

**EPTCS 335**

Proceedings of the  
**Eighteenth Conference on  
Theoretical Aspects of Rationality and  
Knowledge**

**Beijing, China, June 25-27, 2021**

Edited by: Joseph Halpern and Andrés Perea

Published: 22nd June 2021  
DOI: 10.4204/EPTCS.335  
ISSN: 2075-2180  
Open Publishing Association

# Table of Contents

Preface .....	1
<i>Andrés Perea</i>	
A Recursive Measure of Voting Power that Satisfies Reasonable Postulates .....	3
<i>Arash Abizadeh and Adrian Vetta</i>	
Well-Founded Extensive Games with Perfect Information .....	7
<i>Krzysztof R. Apt and Sunil Simon</i>	
Uncertainty-Based Semantics for Multi-Agent Knowing How Logics .....	23
<i>Carlos Areces, Raul Fervari, Andrés R. Saravia and Fernando R. Velázquez-Quesada</i>	
Revisiting Epistemic Logic with Names .....	39
<i>Marta Bílková, Zoé Christoff and Olivier Roy</i>	
Language-based Decisions .....	55
<i>Adam Bjorndahl and Joseph Y. Halpern</i>	
An Awareness Epistemic Framework for Belief, Argumentation and Their Dynamics .....	69
<i>Alfredo Burrieza and Antonio Yuste-Ginel</i>	
Local Dominance .....	85
<i>Emiliano Catonini and Jingyi Xue</i>	
Collective Argumentation: The Case of Aggregating Support-Relations of Bipolar Argumentation Frameworks .....	87
<i>Weiwei Chen</i>	
De Re Updates .....	103
<i>Michael Cohen, Wen Tang and Yanjing Wang</i>	
Dynamically Rational Judgment Aggregation: A Summary .....	119
<i>Franz Dietrich and Christian List</i>	
Deliberation and Epistemic Democracy .....	127
<i>Huihui Ding and Marcus Pivato</i>	
No Finite Model Property for Logics of Quantified Announcements .....	129
<i>Hans van Ditmarsch, Tim French and Rustam Galimullin</i>	
Fire! .....	139
<i>Krisztina Fruzsá, Roman Kuznets and Ulrich Schmid</i>	

Are the Players in an Interactive Belief Model Meta-certain of the Model Itself? .....	155
<i>Satoshi Fukuda</i>	
Knowledge from Probability .....	171
<i>Jeremy Goodman and Bernhard Salow</i>	
Belief Inducibility and Informativeness .....	187
<i>P. Jean-Jacques Herings, Dominik Karos and Toygar Kerman</i>	
Measuring Violations of Positive Involvement in Voting .....	189
<i>Wesley H. Holliday and Eric Pacuit</i>	
Algorithmic Randomness, Bayesian Convergence and Merging .....	211
<i>Simon Huttegger, Sean Walsh and Francesca Zaffora Blando</i>	
Game-Theoretic Models of Moral and Other-Regarding Agents (extended abstract) .....	213
<i>Gabriel Istrate</i>	
Understanding Transfinite Elimination of Non-Best Replies .....	229
<i>Stephan Jagau</i>	
Persuading Communicating Voters .....	231
<i>Toygar Kerman and Anastas P. Tenev</i>	
Knowing How to Plan .....	233
<i>Yanjun Li and Yanjing Wang</i>	
Probabilistic Stability and Statistical Learning .....	249
<i>Krzysztof Mierzewski</i>	
Attainable Knowledge and Omniscience .....	251
<i>Pavel Naumov and Jia Tao</i>	
Failures of Contingent Thinking .....	267
<i>Evan Piermont and Peio Zuazo-Garin</i>	
Reasoning about Emergence of Collective Memory .....	269
<i>R. Ramanujam</i>	
A Deontic Stit Logic Based on Beliefs and Expected Utility .....	281
<i>Aldo Iván Ramírez Abarca and Jan Broersen</i>	
Epistemic Modality and Coordination under Uncertainty .....	295
<i>Giorgio Sbardolini</i>	
Communication Pattern Models: An Extension of Action Models for Dynamic-Network Distributed Systems .....	307
<i>Diego A. Velázquez, Armando Castañeda and David A. Rosenblueth</i>	

# Preface

Andrés Perea

Maastricht University

These proceedings contain the papers that have been accepted for presentation at the Eighteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK XVIII). The conference took place from June 25 to June 27, 2021, at Tsinghua University, Beijing, China. However, due to the COVID-19 pandemic, the conference was offered completely online.

As is to be expected from TARK, these proceedings offer a highly interdisciplinary collection of papers, including areas such as logic, computer science, philosophy, economics, game theory, decision theory and social welfare. The topics covered by the papers include semantic models for knowledge and belief, epistemic logic, computational social choice, rationality in games and decision problems, and foundations of multi-agent systems.

I wish to thank the team of local organizers, chaired by Fenrong Liu, to make this conference possible under these extraordinary circumstances. Another word of gratitude goes to the members of the program committee, not only for reviewing the submissions, but also for their valuable input concerning other aspects of the conference, such as the invited speakers and the precise format of the conference. The members of the program committee are: Christian Bach, Adam Bjorndahl, Giacomo Bonanno, Emiliano Catonini, Franz Dietrich, Davide Grossi, Joseph Halpern (conference chair), Jérôme Lang, Fenrong Liu (local organizing chair), Silvia Milano, Yoram Moses, Eric Pacuit, Andrés Perea (program committee chair), Olivier Roy, Elias Tsakas, Paolo Turrini, Rineke Verbrugge and Kevin Zollman.

I also wish to thank the invited speakers at this conference: Ariel Procaccia, Burkhard Schipper, Sonja Smets and Katie Steele.

On the practical side, the conference and the proceedings have benefitted a lot from the EasyChair platform, and the EPTCS - system. I thank Rob van Glabbeek, editor of EPTCS, for his help during the process of setting up these proceedings.

Last but not least, I am very grateful to Joseph Halpern (conference chair) and Fenrong Liu (local organizing chair) who have done so much for the organization of TARK XVIII. It was an absolute pleasure to work with you, and I am sorry for the many E-mails you had to digest from me.

I sincerely hope that these proceedings will be a source of inspiration for your research, and that you will enjoy reading the papers.

Andrés Perea

Program Committee Chair TARK XVIII

Maastricht, June 2021



# A Recursive Measure of Voting Power that Satisfies Reasonable Postulates

Arash Abizadeh

Department of Political Science  
McGill University  
Montreal, Canada  
arash.abizadeh@mcgill.ca

Adrian Vetta

Department of Mathematics and Statistics  
School of Computer Science  
McGill University  
Montreal, Canada  
adrian.vetta@mcgill.ca

## Extended Abstract

We design a recursive measure of voting power based upon partial voting efficacy as well as full voting efficacy. In contrast, classical indices and measures of voting power incorporate only full voting efficacy. We motivate our design by representing voting games using a division lattice and via the notion of random walks in stochastic processes, and show the viability of our recursive measure by proving it satisfies a plethora of postulates that any reasonable voting measure should satisfy.

There have been two approaches to justifying measures of voting power. The first is the *axiomatic* approach, which seeks to identify a set of reasonable axioms that uniquely pick out a single measure of voting power. To date this justificatory approach has proved a failure: while many have succeeded in providing axiomatic characterizations of various measures, no one has succeeded in doing so for a set of axioms all of which are independently justified, i.e., in showing why it would be reasonable to expect a measure of voting power to satisfy the entire set of axioms that uniquely pick out a proposed measure. For example, Dubey (1975) and Dubey and Shapley (1979) have characterized the classic Shapely-Shubik index (*SS*) and Penrose-Banzhaf measure (*PB*) as uniquely satisfying a distinct set of axioms, respectively, but several of the axioms lack proper justification (Straffin 1982: 292-296; Felsenthal and Machover 1998: 194-195; Laruelle and Valenciano 2001). The second, *two-pronged* approach is more modest and involves combining two prongs of justification. The first prong is to motivate a proposed measure on conceptual grounds, showing the sense in which it captures the intuitive meaning of what voting power is. With this conceptual justification in place, the second prong of justification then requires showing that the measure satisfies a set of reasonable postulates. For the more modest approach, both prongs of justification are necessary, and the satisfaction of reasonable postulates serves, not to pick out a uniquely reasonable measure, but to rule out unreasonable measures.

The first prong of justification has been typically carried out in *probabilistic* terms. For example, the *a priori* Penrose-Banzhaf measure equates a player's voting power, in a given voting structure, with the proportion of logically possible *divisions* or complete vote configurations in which the player is (fully) *decisive* for the division outcome, i.e., in which the player has an alternative voting strategy such that, if it were to choose that alternative instead, the outcome would be different (holding all other players' votes constant). The standard interpretation is that the *a priori* *PB* measure represents the probability a player will be decisive under the assumptions of *equiprobable voting* (the probability a player votes for an alternative is equal to the probability it votes for any other) and *voting independence* (votes are not correlated), which together imply *equiprobable divisions* (the probability of each division is equal) (Felsenthal and Machover 1998: 37-38).

However, measures of voting power based exclusively on the ex ante probability of decisiveness suffer from a crucial conceptual flaw. The motivation for basing a measure of voting power on this notion is that decisiveness is supposed to formalize the idea of a player *making a difference* to the outcome. To equate a player's voting power with the player's ex ante probability of being decisive is to assume that if any particular division were hypothetically to occur, then the player would have efficaciously exercised power to help produce the outcome ex post if and only if that player would have been decisive or necessary for the outcome. Yet this assumption is false: sometimes, as in causally overdetermined outcomes, an actor has efficaciously exercised its power to effect an outcome ex post, and, through the exercise of that power, made a causal contribution to the outcome, even though the actor's contribution was not decisive to it.

More specifically, reducing voting power to the ex ante probability of being decisive fails to take into account players' *partial* causal efficacy in producing outcomes ex post. In this paper, we design a *Recursive Measure (RM)* of voting power that remedies this shortcoming, by taking into account partial efficacy or degrees of causal efficacy. A full conceptual justification for *RM* – i.e., the first prong of justification on the more modest approach – is given in Abizadeh (working paper). *RM* represents, not the *probability* a player will be decisive for the division outcome (the probability the player will be *fully causally efficacious* in bringing it about) but, rather, the player's *expected efficacy*, that is, the probability the player will make a causal contribution to the outcome weighted by the degree of causal efficacy. Whereas decisiveness measures such as *PB* solely track full efficacy, *RM* tracks partial efficacy as well.

Our task in this paper is to furnish the second prong of justification. In particular, we take it that any reasonable measure of a priori voting power  $\pi$  should satisfy, for simple voting games  $\mathcal{G}$  with equiprobable divisions, where  $[n]$  is the set of all voters and a dummy is a voter not decisive in any division, the following postulates:

*Iso-invariance postulate:* For iso-invariant voting games  $\mathcal{G}$  and  $\hat{\mathcal{G}}$ :  $\pi_i = \hat{\pi}_i$  for any player  $i$ .

*Dummy postulates:* For any game  $\hat{\mathcal{G}}$  formed by the addition of a dummy voter to  $\mathcal{G}$ : if  $i$  is a dummy voter, then  $\pi_i = 0$ ;  $\hat{\pi}_i = 0$  only if  $i$  is a dummy voter; and if  $i$  is a non-dummy voter, then  $\pi_i = \hat{\pi}_i$ .

*Dominance postulate:* For any subset  $S \subseteq [n]$  with  $i, j \notin S$ :  $\pi_j \geq \pi_i$  whenever  $j$  weakly dominates  $i$ , and  $\pi_j > \pi_i$  whenever  $j$  strictly dominates  $i$  (where  $j$  *weakly dominates*  $i$  if whenever  $S \cup i$  vote YES and the outcome is YES, then if  $S \cup j$  vote YES the outcome is YES; and  $i$  *strictly dominates*  $j$  if the former weakly dominates the latter but not vice versa).

*Donation postulate:* For any game  $\hat{\mathcal{G}}$  formed from  $\mathcal{G}$  by player  $j$  transferring its vote to player  $i$ :  $\hat{\pi}_i \geq \max(\pi_i, \pi_j)$ .

*Bloc postulate:* For any game  $\hat{\mathcal{G}}$  formed from  $\mathcal{G}$  by player  $i$  annexing  $i$ 's vote to form a bloc  $I = \{i, j\}$ :  $\hat{\pi}_I \geq \max(\pi_i, \pi_j)$ .

*Quarrel postulate:* For any game  $\hat{\mathcal{G}}$  formed from  $\mathcal{G}$  by inducing a symmetric, weak, monotonic quarrel between  $i$  and  $j$ :  $\hat{\pi}_i \leq \pi_i$  and  $\hat{\pi}_j \leq \pi_j$ .

*Added blocker postulate:* For any game  $\mathcal{G}^Y$  resulting from  $\mathcal{G}$  by adding an *added YES-blocker*, and  $\mathcal{G}^N$  resulting from adding an *added NO-blocker*:  $\frac{\pi_i^+(\mathcal{G})}{\pi_j^+(\mathcal{G})} = \frac{\pi_i^+(\mathcal{G}^Y)}{\pi_j^+(\mathcal{G}^Y)}$ , and  $\frac{\pi_i^-(\mathcal{G})}{\pi_j^-(\mathcal{G})} = \frac{\pi_i^-(\mathcal{G}^N)}{\pi_j^-(\mathcal{G}^N)}$  (where  $\pi^+$  is a player's YES-voting power, based solely on divisions in which it votes YES, and  $\pi^-$  is a player's NO-voting power, based solely on divisions in which it votes NO).

In the full paper, we explain the intuitive justification for and fully specify each of these voting-power postulates, and then prove that *RM* satisfies them for a priori power in simple voting games. We prove these postulates by introducing a new way of representing voting games using a division lattice, and show that previous formulations of some of these postulates require revision.

A full version of the paper can be found at: <http://arxiv.org/abs/2105.03006>

## References

- [1] Abizadeh, A. (Working paper). A Recursive Measure of Voting Power with Partial Decisiveness or Efficacy.
- [2] Dubey, P. (1975). On the Uniqueness of the Shapley Value. *International Journal of Game Theory*, 4(3), 131-139. doi:10.1007/BF01780630
- [3] Dubey, P., & Shapley, L.S. (1979). Mathematical Properties of the Banzhaf Power Index. *Mathematics of Operations Research*, 4(2), 99-131. doi:10.1287/moor.4.2.99
- [4] Felsenthal, D.S., & Machover, M. (1998). *The Measurement of Voting Power: Theory and Practice, Problems and Paradoxes*. Cheltenham, UK: Edward Elgar. doi:10.4337/9781840647761
- [5] Laruelle, A., & Valenciano, F. (2001). Shapley-Shubik and Banzhaf Indices Revisited. *Mathematics of Operations Research*, 26(1), 89-104. doi:10.1287/moor.26.1.89.10589
- [6] Straffin, P.D. (1982). Power Indices in Politics. In S. J. Brams, W. F. Lucas, & P. D. Straffin (Eds.), *Political and Related Models* (pp. 256-321). New York: Springer. doi:10.1007/978-1-4612-5430-0\_11



# Well-Founded Extensive Games with Perfect Information

Krzysztof R. Apt

Centrum Wiskunde & Informatica  
Amsterdam, The Netherlands  
University of Warsaw  
Warsaw, Poland  
k.r.ap@cw.i.nl

Sunil Simon

Department of CSE,  
IIT Kanpur, Kanpur, India  
simon@cse.iitk.ac.in

We consider extensive games with perfect information with well-founded game trees and study the problems of existence and of characterization of the sets of subgame perfect equilibria in these games. We also provide such characterizations for two classes of these games in which subgame perfect equilibria exist: two-player zero-sum games with, respectively, two and three outcomes.

## 1 Introduction

Research on strategic games assumes that players have to their disposal infinitely many strategies. This allows one to view strategic games with mixed strategies as customary strategic games with infinitely many strategies. This assumption is also used in a study of various standard examples, such as Cournot or Bertrand competition, in which the players have to set the production level or the price of a product.

In contrast, the exposition of the standard results for the extensive games with perfect information (from now, just ‘extensive games’) is usually limited to finite games. This restriction rules out a study of various natural examples, for example infinite variants of the ultimatum game or some bargaining games, see, e.g., [16]. In the first case the game has just two stages but the first player may have infinitely many actions to choose from, while in the second case the game has an arbitrary, though finite, number of stages. Such games are then analyzed separately, without taking into account general results.

The aim of this paper is to provide a systematic account of extensive games with perfect information in a setting that only requires that the underlying game tree is well-founded (i.e., has no infinite paths). We call such games *well-founded*<sup>1</sup>.

The standard tool to analyze finite extensive games is the concept of a subgame perfect equilibrium. Their existence is established by means of the backward induction algorithm. In infinite well-founded extensive games subgame perfect equilibria may fail to exist. Also one cannot resort to any version of this algorithm since it will not terminate. In principle this could be taken care of by defining a joint strategy as an eventual outcome of an infinite computation. However, for arbitrary well-founded games one would have then to proceed by means of a transfinite induction, which raises a legitimate question whether such a process can be called a computation.

Therefore, instead of trying to define such generalized computations we dispense with the backward induction altogether and simply proceed by transfinite induction. This results in mathematical proofs of existence that are not supported by any algorithm, but still, as illustrated by examples, the obtained results can be used to compute the sets of subgame perfect equilibria in specific well-founded games and to deduce their existence in special cases. Informally, transfinite induction analyzes the game tree ‘top

---

<sup>1</sup>Such games are sometimes called games with *finite horizon* (see, e.g., [15]). We decided to use instead the qualification ‘well-founded’ because ‘finite horizon’ is sometimes used to indicate that the game tree is of bounded depth, i.e., has a finite rank (a concept introduced in the next section).

down', while the backward induction proceeds 'bottom up' and consequently cannot be naturally applied to infinite game trees.

Most results, though not all, are natural generalizations of the corresponding results for the case of finite extensive games. Some of these results fail to hold for infinite games and some of the traditional proofs for finite games, notably the ones involving the backward induction, have to be suitably modified. In what follows we focus both on arbitrary well-founded games and on two-player zero-sum games with, respectively, two and three outcomes.

In the literature we found only one paper in which well-founded games appear, namely [5]. The authors provide using higher-order computability theory a formula that defines the set of subgame perfect equilibria under an assumption that implies their existence, and apply it to determine a subgame perfect equilibrium in an infinite three stage game. In several books various examples of infinite extensive games are studied and various extensions of finite extensive games, for example games with chance moves, see, e.g., [16], simultaneous moves, see, e.g., [15], or repeated extensive games, see [13], are introduced (not to mention games with imperfect information). Also subgame perfect equilibria in games allowing infinite plays have been studied, see, e.g., [8] and a more recent [10]. In [1] a maximally general definition of an extensive form game is proposed that among others covers repeated games, differential games, and stochastic games. In the proposed framework even immediate predecessors of an action may not exist (like in continuous time interactive decisions examples). In our opinion the class of games considered here merits attention as a first natural generalization to study.

In the next section we introduce the relevant concepts and provide natural examples of well-founded extensive games. In Section 3 we establish existence of subgame perfect equilibria for some natural classes of well-founded games and show how to apply a characterization result to compute the set of subgame perfect equilibria for specific example games. Then, in Section 4 we consider two classes of two-player well-founded games: win or lose games and chess-like games. As a stepping stone towards characterizations of the sets of subgame perfect equilibria in these games we show that the well-known result attributed to Zermelo [23] (see also [19]) about existence of winning strategies continues to hold for well-founded games.

## 2 Preliminaries on extensive games

A *tree* is an acyclic directed connected graph, written as  $(V, E)$ , where  $V$  is a non-empty set of nodes and  $E$  is a possibly empty set of edges. In drawings the edges will be directed downwards.

An *extensive game with perfect information* (in short, just an *extensive game*) for  $n \geq 1$  players consists of:

- a set of players  $\{1, \dots, n\}$ ,
- a *game tree*, which is a tree  $T := (V, E)$  with a *turn function*  $turn : V \setminus Z \rightarrow \{1, \dots, n\}$ , where  $Z$  is the set of leaves of  $T$ ,
- the *payoff functions*  $p_i : Z \rightarrow \mathbb{R}$ , for each player  $i$ .

We denote it by  $(T, turn, p_1, \dots, p_n)$ .

The function  $turn$  determines at each non-leaf node which player should move. The edges of  $T$  represent possible *moves* in the considered game, while for a node  $v \in V \setminus Z$  the set of its children  $C(v) := \{w \mid (v, w) \in E\}$  represents possible *actions* of player  $turn(v)$  at  $v$ . For a node  $u$  in  $T$  let  $T^u$  denote the subtree of  $T$  rooted at  $u$ .

We say that an extensive game is *finite*, *finite depth*, *infinite*, or *well-founded* if, respectively, its game tree is finite, finite depth, infinite, or well-founded. Recall that a tree is called *well-founded* if it has no infinite paths (see, e.g., [21, page 224]).

Further, following [2], we say that an extensive game is *without relevant ties* if for all non-leaf nodes  $u$  in  $T$  the function  $p_i$ , where  $turn(u) = i$ , is injective on the leaves of  $T^u$ . This is more general than saying that a game is *generic*, which means that each  $p_i$  is an injective function.

We shall often rely on the concept of a *rank* of a well-founded tree  $T$ . Recall that it is defined inductively as follows, where  $v$  is the root of  $T$ :

$$rank(T) := \begin{cases} 0 & \text{if } T \text{ has one node} \\ \sup\{rank(T^u) + 1 \mid u \in C(v)\} & \text{otherwise,} \end{cases}$$

where  $\sup(X)$  denotes the least ordinal larger than all ordinals in the set  $X$ . Transfinite induction will be needed only to deal with games on the trees with  $rank > \omega$ .

In the figures below we identify the actions with the labels we put on the edges and thus identify each action with the corresponding move. For convenience we do not assume the labels to be unique, but it will not lead to confusion. Further, we annotate the non-leaf nodes with the identity of the player whose turn it is to move and the name of the node. Finally, we annotate each leaf node with the corresponding sequence of the values of the  $p_i$  functions.

**Example 1** The following two-player game is called the *Ultimatum game*. Player 1 moves first and claims a real number  $x \in [0, 100]$ , to be interpreted as a fraction of some good to be shared, leaving the fraction  $100 - x$  for the other player. Player 2 either *accepts* this decision, the outcome is then  $(x, 100 - x)$ , or *rejects* it, the outcome is then  $(0, 0)$ . The game tree is depicted in Figure 1, where the action of player 1 is a number from the set  $[0, 100]$ , and the actions of player 2 are denoted by  $A$  and  $R$ . The resulting game is infinite but is of finite depth. The rank of the game tree is 2.

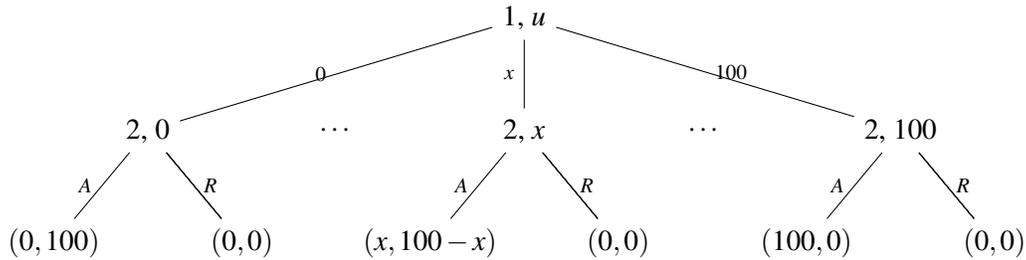


Figure 1: The Ultimatum game

□

**Example 2** Consider the following *Bargaining game* (without depreciation). Player 1 moves first by selecting a natural number  $k \geq 2$ . Such a choice is to be interpreted that he claims the fraction  $1 - \frac{1}{k}$  of some good to be shared, leaving the fraction  $\frac{1}{k}$  for the other player. As long as player 1 selects  $k > 2$ , player 2 asks for a better offer or rejects it. In the first case player 1 selects  $k - 1$ . The game continues until player 1 selects 2, i.e., claims 50 % of the good. At that moment player 2 either accepts this offer, the outcome is then  $(50, 50)$ , or rejects it. All rejections result in the outcome  $(0, 0)$ . The game tree of this game, depicted in Figure 2, has arbitrary long, though finite, branches, so this game is infinite but it is well-founded. The rank of the game tree is  $\omega$ . □

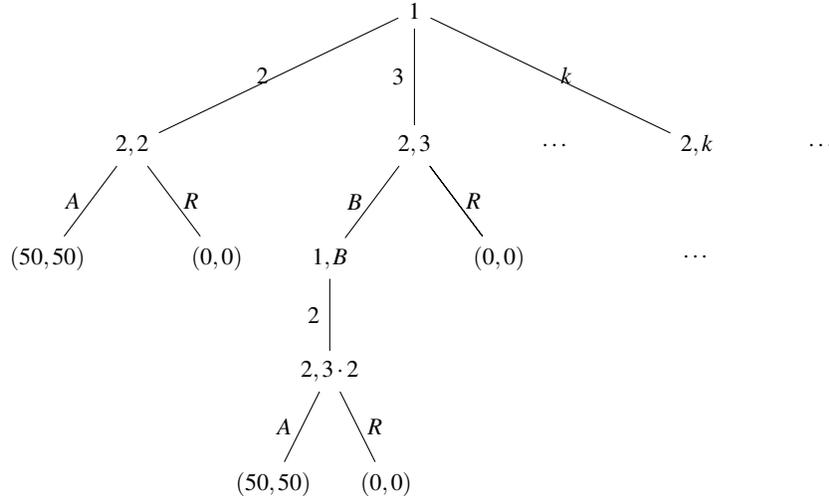


Figure 2: The Bargaining game

Below, given a two-player extensive game we denote the opponent of player  $i$  by  $-i$  instead of  $3-i$ .

**Example 3** We now construct a sequence of games  $G(i, \alpha)$ , where  $i \in \{1, 2\}$  and  $\alpha$  is an ordinal  $> 1$ , by induction as follows:

- $G(1, 2)$  is the the ultimatum game from Example 1 and  $G(2, 2)$  is its version with the roles of players 1 and 2 reversed;
- $G(i, \alpha)$ , where  $\alpha > 2$ , is obtained as follows:
  - its game tree is constructed by selecting a root  $v$ , taking the game trees of the games  $G(-i, \beta)$ , where  $1 < \beta < \alpha$  and selecting their roots as the children of  $v$ ,
  - setting  $turn(v) = i$ .

So for example the root of the game tree of  $G(i, 3)$  has one child, namely the root of the game tree of  $G(-i, 2)$ , the root of the game tree of  $G(i, 4)$  has two children, namely the roots of the game trees of  $G(-i, 3)$  and  $G(-i, 2)$ , etc. Note that the rank of the game tree of  $G(i, \alpha)$  is  $\alpha$ .  $\square$

The class of endogenous games studied in [9] form another example of well-founded extensive games. These games are played in two stages. In the first stage the players are involved in pre-play negotiations that essentially fix the payoff functions and in the second stage they choose their strategies. The resulting game is infinite due to the pre-play negotiations, while the rank of the game tree is 2.

Note that by König's lemma [11] every finitely branching well-founded extensive games is finite. Consequently, interesting well-founded extensive games necessarily have infinite branching.

For an extensive game  $G := (T, turn, p_1, \dots, p_n)$  let  $V_i := \{v \in V \setminus Z \mid turn(v) = i\}$ . So  $V_i$  is the set of nodes at which player  $i$  moves. A **strategy** for player  $i$  is a function  $s_i : V_i \rightarrow V$ , such that  $(v, s_i(v)) \in E$  for all  $v \in V_i$ . We denote the set of strategies of player  $i$  by  $S_i$ .

Let  $S = S_1 \times \dots \times S_n$ . We call each element  $s \in S$  a **joint strategy**, denote the  $i$ th element of  $s$  by  $s_i$ , and abbreviate the sequence  $(s_j)_{j \neq i}$  to  $s_{-i}$ . We write  $(s'_i, s_{-i})$  to denote the joint strategy in which player  $i$ 's strategy is  $s'_i$  and for all  $j \neq i$ , player  $j$ 's strategy is  $s_j$ . Occasionally we write  $(s_i, s_{-i})$  instead of  $s$ . Finally, we abbreviate the Cartesian product  $\times_{j \neq i} S_j$  to  $S_{-i}$ . So in the degenerate situation when the game

tree consists of just one node, each strategy is the empty function, denoted by  $\emptyset$ , and there is only one joint strategy, namely the  $n$ -tuple of these functions. Each joint strategy assigns a unique descendant to every node in  $V \setminus Z$ . In fact, we can identify joint strategies with such assignments.

From now on the above notation will be used in the context of any considered extensive game  $G$ . In particular  $S_i$  will always denote the set of strategies of player  $i$ .

Each joint strategy  $s = (s_1, \dots, s_n)$  determines a rooted path  $play(s) := (v_1, \dots, v_m)$  in  $T$  defined inductively as follows:

- $v_1$  is the root of  $T$ ,
- if  $v_k \notin Z$ , then  $v_{k+1} := s_i(v_k)$ , where  $turn(v_k) = i$ .

So when the game tree consists of just one node,  $v$ , we have  $play(s) = v$ . Informally, given a joint strategy  $s$ , we can view  $play(s)$  as the resulting *play* of the game.

Suppose now that the extensive game is well-founded. Then for each joint strategy  $s$  the rooted path  $play(s)$  is finite. Denote by  $leaf(s)$  the last element of  $play(s)$ . We call  $(p_1(leaf(s)), \dots, p_n(leaf(s)))$  the **outcome** of the game  $G$  when each player  $i$  pursues his strategy  $s_i$  and abbreviate it as  $p(leaf(s))$ . We call two joint strategies  $s$  and  $t$  **payoff equivalent** if  $p(leaf(s)) = p(leaf(t))$ .

We say that a strategy  $s_i$  of player  $i$  is a **best response** to a joint strategy  $s_{-i}$  of his opponents if for all  $s'_i \in S_i$ ,  $p_i(leaf(s)) \geq p_i(leaf(s'_i, s_{-i}))$ . Next, we call a joint strategy  $s$  a **Nash equilibrium** if each  $s_i$  is a best response to  $s_{-i}$ , that is, if

$$\forall i \in \{1, \dots, n\}, \forall s'_i \in S_i, p_i(leaf(s_i, s_{-i})) \geq p_i(leaf(s'_i, s_{-i})).$$

**Example 4** Let us return to the Ultimatum game from Example 1. Each strategy for player 1 is a number, respectively from  $[0, 100]$ , while each strategy for player 2 assigns to every such number  $x$  either  $A$  or  $R$ .

It is easy to check that each Nash equilibrium is of the form  $(100, \text{always } R)$ , with the outcome  $(100, 0)$ , or  $(x, s_2)$  with  $s_2(x) = A$  and  $s_2(y) = R$  for  $y > x$ , where  $x, y \in [0, 100]$ , with the outcome  $(x, 100 - x)$ .  $\square$

Finally, we recall the notion of a subgame perfect equilibrium due to Selten [20] (see also section 6.2 in [15]), though now defined for the larger class of well-founded games.

The **subgame of  $G$  rooted at the node  $w$** , denoted by  $G^w$ , is defined as follows:

- its set of players is  $\{1, \dots, n\}$ ,
- its tree is  $T^w$ ,
- its turn and payoff functions are the restrictions of the corresponding functions of  $G$  to the nodes of  $T^w$ .

Note that some players may ‘drop out’ in  $G^w$ , in the sense that at no node of  $T^w$  it is their turn to move. Still, to keep the notation simple, it is convenient to admit in  $G^w$  all original players in  $G$ .

Each strategy  $s_i$  of player  $i$  in  $G$  uniquely determines his strategy  $s_i^w$  in  $G^w$ . Given a joint strategy  $s = (s_1, \dots, s_n)$  of  $G$  we denote by  $s^w$  the joint strategy  $(s_1^w, \dots, s_n^w)$  in  $G^w$ . Further, we denote by  $S_i^w$  the set of strategies of player  $i$  in the subgame  $G^w$  and by  $S^w$  the set of joint strategies in this subgame.

Suppose now the extensive game  $G$  is well-founded. Then the notion of a Nash equilibrium is well-defined. A joint strategy  $s$  of  $G$  is called a **subgame perfect equilibrium** in  $G$  if for each node  $w$  of  $T$ , the joint strategy  $s^w$  of  $G^w$  is a Nash equilibrium in  $G^w$ . Informally  $s$  is subgame perfect equilibrium in  $G$  if it induces a Nash equilibrium in every subgame of  $G$ .

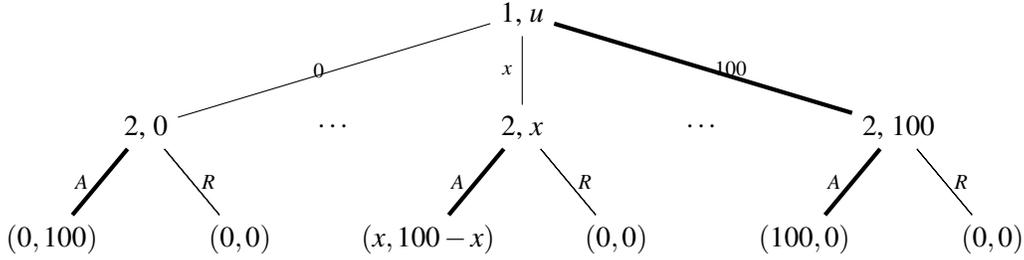


Figure 3: The subgame perfect equilibrium in the Ultimatum game

**Example 5** Return now to the Ultimatum game from Example 1. It is easy to check that it has exactly one subgame equilibrium, depicted in Figure 3 by thick lines.

Note that even though the rank of the game tree is 2, the customary backward induction cannot be applied here to compute this subgame equilibrium. Indeed, to deal with the root node the algorithm has to deal first with infinitely many nodes at which player 2 moves, which leads to divergence.  $\square$

An analysis of the subgame perfect equilibria in the games from Examples 2 and 3 is more involved and will be provided in the next section using a characterization of the set of subgame perfect equilibria of a well-founded extensive game.

### 3 Subgame perfect equilibria in well-founded games

In general subgame perfect equilibria may not exist in well-founded games. As an example take the modification of the Ultimatum game from Example 1 in which instead of  $[0, 100]$  one considers the open interval  $(0, 100)$ . In this section we establish existence of subgame perfect equilibria in some natural classes of well-founded games. This will directly follow from a characterization of the sets of subgame perfect equilibria in such games.

We begin by stating a preparatory lemma, called the ‘one deviation property’ in [15]. To keep the paper self-contained we include in the Appendix the proof. It is more detailed than the one given in [15].

**Lemma 6** *Let  $G$  be a well-founded extensive game over the game tree  $T$ . A joint strategy  $s$  is a subgame perfect equilibrium in  $G$  iff for all non-leaf nodes  $u$  in  $T$  and all  $y \in C(u)$*

- $p_i(\text{leaf}(s^x)) \geq p_i(\text{leaf}(s^y))$ , where  $i = \text{turn}(u)$  and  $s_i(u) = x$ .

**Corollary 7** *Let  $G$  be a well-founded extensive game over the game tree  $T$  with the root  $v$ . A joint strategy  $s$  is a subgame perfect equilibrium in  $G$  iff for all  $u \in C(v)$*

- $p_i(\text{leaf}(s^w)) \geq p_i(\text{leaf}(s^u))$ , where  $i = \text{turn}(v)$  and  $s_i(v) = w$ ,
- $s^u$  is a subgame perfect equilibrium in the subgame  $G^u$ .

Intuitively, the first condition states that among the subgames rooted at the children of the root  $v$ , the one determined by the first move in the game  $G$  yields the maximal outcome for the player who moved. Recall that for a function  $f : X \rightarrow Y$  (with  $X$  non-empty),  $\text{argmax}_{x \in X} f(x) := \{y \in X \mid f(y) = \max_{x \in X} f(x)\}$ . Using this notation this condition can be reformulated as:  $s_i(v) \in \text{argmax}_{u \in C(v)} p_i(\text{leaf}(s^u))$ , where  $i = \text{turn}(v)$ .

**Proof.** If  $C(v) = \emptyset$ , the claim is vacuously true. Otherwise consider any  $u \in C(v)$ . By Lemma 6  $s^u$  is a subgame perfect equilibrium in  $G^u$  iff for all non-leaf nodes  $y$  in  $T^u$  and  $z \in C(y)$ ,  $p_i(\text{leaf}((s^u)^x)) \geq p_i(\text{leaf}((s^u)^z))$ , where  $i = \text{turn}(y)$  and  $s_i^u(y) = x$ .

Since  $(s^u)^x = s^x$  and  $(s^u)^z = s^z$ , the last statement is equivalent to the statement that the inequality in Lemma 6 holds for all non-leaf nodes  $y$  in  $T^u$  and  $z \in C(y)$ . The conclusion now follows by Lemma 6.  $\square$

The above corollary allows us to characterize inductively the set of subgame perfect equilibria in each well-founded extensive game.

Consider a well-founded extensive game  $G$  with the root  $v$  and suppose  $C(v) \neq \emptyset$ . Consider the subgames  $G^w$ , where  $w \in C(v)$ , and a function  $f$  that assigns to each sequence  $t$  of joint strategies in these subgames a child of  $v$ . Then each pair of  $t$  and  $f$  determines a joint strategy in  $G$  that we denote by  $(f, t)$ .

Recall that by  $S^w$  we denote the set of joint strategies in the subgame  $G^w$ . Given subsets  $U^w$  of  $S^w$  for  $w \in C(v)$  and a set of functions  $F$  from  $\times_{w \in C(v)} U^w$  to  $C(v)$ , we denote by  $[F, \times_{w \in C(v)} U^w]$  the set of joint strategies in  $G$  defined by

$$[F, \times_{w \in C(v)} U^w] := \begin{cases} \{(\emptyset, \dots, \emptyset)\} & \text{if } C(v) = \emptyset \\ \{(f, t) \mid f \in F, t \in \times_{w \in C(v)} U^w\} & \text{otherwise} \end{cases}$$

In the first case  $(\emptyset, \dots, \emptyset)$  stands for the joint strategy that consists of the  $n$ -tuple of the empty strategies. Note that when  $C(v) \neq \emptyset$  if any of the sets  $U^w$  or  $F$  is empty, then so is  $[F, \times_{w \in C(v)} U^w]$ . Further, we denote the set of subgame perfect equilibria in  $G$  by  $\text{SPE}(G)$ .

**Theorem 8** Consider a well-founded extensive game  $G$  with the root  $v$  and let  $i = \text{turn}(v)$ . Then

$$\text{SPE}(G) = [F, \times_{w \in C(v)} \text{SPE}(G^w)],$$

where if  $C(v) \neq \emptyset$  then  $F = \{f \mid \forall t \in \times_{w \in C(v)} \text{SPE}(G^w) f(t) \in \text{argmax}_{w \in C(v)} p_i(\text{leaf}(t^w))\}$ .

In particular, if the set  $\text{argmax}_{w \in C(v)} p_i(\text{leaf}(t^w))$  is empty, then  $F = \emptyset$  and hence  $\text{SPE}(G) = \emptyset$ . Intuitively, each function  $f \in F$ , given a sequence of subgame perfect equilibria in the subgames rooted at the children of the root  $v$ , selects a root of the subgame in which the outcome in the equilibrium is maximal for the player who moves at  $v$ .

**Proof of Theorem 8.** If  $C(v) = \emptyset$ , then  $(\emptyset, \dots, \emptyset)$  is a unique subgame perfect equilibrium, so the claim holds.

If  $C(v) \neq \emptyset$ , then by Corollary 7 every subgame perfect equilibrium in the game  $G$  is of the form  $(f, t)$ , where for all  $w \in C(v)$ ,  $t^w$  is a subgame perfect equilibrium in  $G^w$  and for some  $w \in C(v)$  we have  $f(t) = w$  and  $p_i(\text{leaf}(s^w)) \geq p_i(\text{leaf}(s^u))$  for all  $u \in C(v)$ .  $\square$

**Corollary 9** Every well-founded extensive game with finitely many outcomes has a subgame perfect equilibrium.

**Proof.** The claim follows from Theorem 8 by induction on the rank of the game tree and the observation that for every function  $g : X \rightarrow Y$  with a finite range the set  $\text{argmax}_{x \in X} g(x)$  is non-empty.  $\square$

The above result can be generalized to some games with infinitely many outcomes. An example is a game in which for each player the set of outcomes is either finite or equals the set of negative integers. More generally, consider a well-founded extensive game in which for each player the set of outcomes is a reverse well-ordered set, i.e., every subset of this set has a greatest element. Then Theorem 8 implies that the game has a subgame perfect equilibrium.

**Corollary 10** Every well-founded extensive game without relevant ties has at most one subgame perfect equilibrium.

**Proof.** If a game is without relevant ties, then so is every subgame of it. This allows us to proceed by induction on the rank of the game tree. For game trees of rank 0 the claim clearly holds. Suppose that it holds for all well-founded extensive games without relevant ties with the game trees of rank smaller than some ordinal  $\alpha > 0$ . Consider such a game with game tree of rank  $\alpha$  and rooted at  $v$ . Let  $i = \text{turn}(v)$ .

By the induction hypothesis for each  $w \in C(v)$  the set  $\text{SPE}(G^w)$  has at most one element. If one of these sets is empty, then so is  $\text{SPE}(G)$ .

So suppose that each  $\text{SPE}(G^w)$  is a singleton set. Then so is  $\times_{w \in C(v)} \text{SPE}(G^w)$ . Let  $\times_{w \in C(v)} \text{SPE}(G^w) = \{t\}$ . Then for different  $w, w' \in C(v)$ ,  $\text{leaf}(t^w)$  and  $\text{leaf}(t^{w'})$  are different leaves of the game tree of  $G$ , so by the assumption about the game  $p_i(\text{leaf}(t^w)) \neq p_i(\text{leaf}(t^{w'}))$ , since  $i = \text{turn}(v)$ .

This means that the function  $g : C(v) \rightarrow \mathbb{R}$  defined by  $g(w) := p_i(\text{leaf}(t^w))$  is injective. Consequently the set  $\text{argmax}_{w \in C(v)} p_i(\text{leaf}(t^w))$  has at most one element and hence the same successively holds for the sets  $F$  and  $\text{SPE}(G)$ .  $\square$

In particular, every generic well-founded extensive game with finitely many outcomes has a unique subgame perfect equilibrium. We now show how Theorem 8 can be used to reason about subgame perfect equilibria in specific extensive games.

**Example 11** Consider the Bargaining game  $G$  from Example 2. Denote by  $G(k)$  the game in which player 1 first selects the number  $k$ . The inductive structure of these games is depicted in Figure 4, where the actions of player 2 are  $B$  ('make a better offer') or  $A$  and  $R$ , as in Example 1.

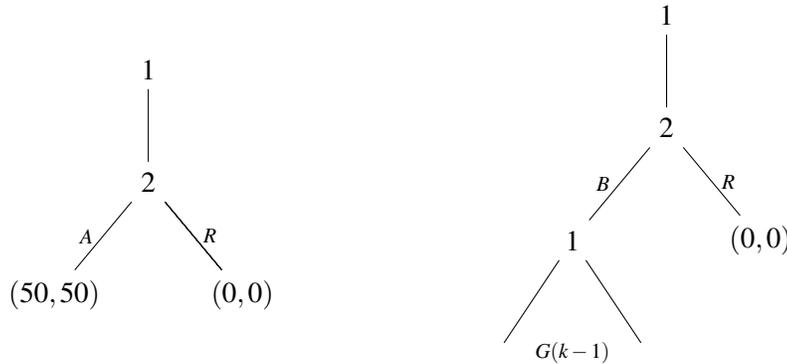


Figure 4: The games  $G(2)$  and  $G(k)$  for  $k > 2$

It is easy to prove by induction using Theorem 8 or simply by the backward induction (the presentation of which we omit) that each game  $G(k)$ , where  $k \geq 2$ , has a unique subgame perfect equilibrium with the outcome  $(50, 50)$ .

Children of the root of the game tree of  $G$  are the roots of the game trees of  $G(k)$ , where  $k \geq 2$ . So for the game  $G$  the set  $\text{argmax}_{w \in C(v)} p_i(\text{leaf}(t^w))$  referred to in Theorem 8 has exactly one element, 50. Hence Theorem 8 implies that  $G$  has a subgame perfect equilibrium, that the outcome in each of them is  $(50, 50)$ , and that for each  $k \geq 2$  there is a unique subgame perfect equilibrium in which player 1 first selects  $k$ .  $\square$

**Example 12** Consider now the games  $G(i, \alpha)$ , where  $i \in \{1, 2\}$  and  $\alpha$  is an ordinal  $> 1$  from Example 3. We noticed in Example 5 that the game  $G(1, 2)$  has a unique subgame perfect equilibrium with the outcome  $(100, 0)$ . By symmetry the game  $G(2, 2)$  has a unique subgame perfect equilibrium with the outcome  $(0, 100)$ . The root of the game tree  $G(i, 3)$  has one child, which is the root of the game tree of  $G(-i, 2)$ . Consequently  $G(i, 3)$  has a unique subgame perfect equilibrium with the outcome  $(0, 100)$  for  $i = 1$  and  $(100, 0)$  for  $i = 2$ .

Using these observations we now show that for  $i \in \{1, 2\}$  and ordinals  $\alpha > 3$  the game  $G(i, \alpha)$  has a subgame perfect equilibrium and the outcomes in all these equilibria are all  $(100, 0)$  for  $i = 1$  and  $(0, 100)$  for  $i = 2$ . We proceed by induction. Consider a game  $G(1, \alpha)$  with  $\alpha > 3$  and assume the claim holds for all  $\beta$  with  $3 \leq \beta < \alpha$ . The root of the game tree has as children the roots of the game trees of  $G(2, \beta)$ , where  $1 < \beta < \alpha$ .

By the induction hypothesis all these games except  $G(2, 3)$  have subgame perfect equilibria with the outcomes  $(0, 100)$ . For the game  $G(2, 3)$ , as just noted, the outcome in the unique subgame perfect equilibrium is  $(100, 0)$ . So for the game  $G(1, \alpha)$  the set  $\operatorname{argmax}_{w \in C(v)} p_i(\operatorname{leaf}(t^w))$  referred to in Theorem 8 has exactly one element, 100. Using this theorem we conclude that the game  $G(1, \alpha)$  has a subgame perfect equilibrium and that the outcome in each equilibrium is  $(100, 0)$ . A symmetric claim, referring to  $(0, 100)$  instead, holds for each game  $G(2, \alpha)$  with  $\alpha > 3$ .

Using Theorem 8 we conclude that each  $G(i, \alpha)$  for  $\alpha > 4$  has multiple subgame perfect equilibria.  $\square$

Next, we establish a result showing that for a class of well-founded extensive games all subgame perfect equilibria are payoff equivalent. The following condition was introduced in [17]:

$$\forall i \in \{1, \dots, n\} \forall s, t \in S [p_i(\operatorname{leaf}(s)) = p_i(\operatorname{leaf}(t)) \rightarrow p(\operatorname{leaf}(s)) = p(\operatorname{leaf}(t))]. \quad (1)$$

This condition is in particular satisfied by the two-player well-founded extensive games that are *strictly competitive*, which means that

$$\forall i \in \{1, 2\} \forall s, s' \in S p_i(\operatorname{leaf}(s)) \geq p_i(\operatorname{leaf}(s')) \text{ iff } p_{-i}(\operatorname{leaf}(s)) \leq p_{-i}(\operatorname{leaf}(s')).$$

(To see it transpose  $i$  and  $-i$  and conjoin both equivalences.)

**Theorem 13** *In every well-founded extensive game that satisfies condition (1) all subgame perfect equilibria are payoff equivalent.*

**Proof.** First we prove the following claim.

**Claim.** If a well-founded extensive game satisfies condition (1), then so does every subgame of it.

*Proof.* Let  $G$  be a well-founded extensive game that satisfies condition (1). Consider any subgame  $G^w$  of  $G$ . Suppose that for some player  $i$  and joint strategies  $s'$  and  $t'$  in  $G^w$  we have  $p_i(\operatorname{leaf}(s')) = p_i(\operatorname{leaf}(t'))$ . Take some joint strategies  $s$  and  $t$  in  $G$  such that  $\operatorname{leaf}(s) = \operatorname{leaf}(s')$ ,  $\operatorname{leaf}(t) = \operatorname{leaf}(t')$ ,  $s^w = s'$  and  $t^w = t'$ . Then  $p_i(\operatorname{leaf}(s)) = p_i(\operatorname{leaf}(t))$ , so by condition (1)  $p(\operatorname{leaf}(s)) = p(\operatorname{leaf}(t))$  and consequently  $p(\operatorname{leaf}(s')) = p(\operatorname{leaf}(t'))$ .  $\square$

We now proceed by induction on the rank of the game tree. For game trees of rank 0 the claim obviously holds. Suppose the claim holds for all well-founded extensive games whose game tree is of rank smaller than some ordinal  $\alpha > 0$ . Consider a well-founded game  $G = (T, \operatorname{turn}, p_1, \dots, p_n)$  over a game tree of rank  $\alpha$  with the root  $v$ . Take two subgame perfect equilibria  $s$  and  $t$  in  $G$ .

If  $\operatorname{path}(s) = \operatorname{path}(t)$ , then  $p(\operatorname{leaf}(s)) = p(\operatorname{leaf}(t))$ . Otherwise take the first non-leaf node  $u$  lying on  $\operatorname{path}(s)$  such that  $s_i(u) \neq t_i(u)$ , where  $i = \operatorname{turn}(u)$ . Let  $s_i(u) = x$  and  $t_i(u) = y$ .

Both  $s^y$  and  $t^y$  are subgame perfect equilibria in the subgame  $G^y$ . By the Claim the game  $G^y$  satisfies condition (1), so by the induction hypothesis  $s^y$  and  $t^y$  are payoff equivalent in  $G^y$ . We thus have

$$p_i(\operatorname{leaf}(s)) = p_i(\operatorname{leaf}(s^x)) \geq p_i(\operatorname{leaf}(s^y)) = p_i(\operatorname{leaf}(t^y)) = p_i(\operatorname{leaf}(t)),$$

where the inequality holds by Lemma 6. Analogously  $p_i(\operatorname{leaf}(t)) \geq p_i(\operatorname{leaf}(s))$ , so  $p_i(\operatorname{leaf}(s)) = p_i(\operatorname{leaf}(t))$  and hence by condition (1)  $p(\operatorname{leaf}(s)) = p(\operatorname{leaf}(t))$ .  $\square$

For finite extensive games this result was stated in [15, page 100] as Exercise 100.2. The most natural proof makes use of the backward induction. For infinite games a different proof is needed.

We say that a well-founded extensive game  $(T, \text{turn}, p_1, \dots, p_n)$  satisfies the **transference of decision-maker indifference (TDI)** condition if:  $\forall i \in \{1, \dots, n\} \forall r_i, t_i \in S_i \forall s_{-i} \in S_{-i}$ ,

$$p_i(\text{leaf}(r_i, s_{-i})) = p_i(\text{leaf}(t_i, s_{-i})) \rightarrow p(\text{leaf}(r_i, s_{-i})) = p(\text{leaf}(t_i, s_{-i})).$$

Informally, this condition states that whenever for some player  $i$  two of his strategies  $r_i$  and  $t_i$  are indifferent w.r.t. some joint strategy  $s_{-i}$  of the other players then this indifference extends to all players.

Clearly, condition (1) implies the TDI condition. The TDI condition was introduced in [14], the results of which imply that in every finite extensive game with perfect information that satisfies the TDI condition all subgame perfect equilibria are payoff equivalent. We conjecture that this result extends to well-founded extensive games.

## 4 Win or lose and chess-like games

In this section we characterize subgame perfect equilibria of two-player zero-sum well-founded extensive games with, respectively, two and three outcomes. By Corollary 9 each of these games has a subgame perfect equilibrium. Below we consider the outcomes  $(1, -1)$ ,  $(0, 0)$ , and  $(-1, 1)$ , but the obtained results hold with the same proofs for arbitrary outcomes as long as the game remains zero-sum.

A two-player extensive game is called a **win or lose game** if the only possible outcomes are  $(1, -1)$  and  $(-1, 1)$ , with 1 associated with winning and 0 with losing. Given a well-founded win or lose game  $G$  we call a strategy  $s_i$  of player  $i$  a **winning strategy** if  $\forall s_{-i} \in S_{-i} p_i(\text{leaf}(s_i, s_{-i})) = 1$ . Below we denote the (possibly empty) set of winning strategies of player  $i$  in  $G$  by  $\text{win}_i(G)$ .

A classic result, attributed to Zermelo [23], implies that in finite win or lose games one of the players has a winning strategy. This result also holds for arbitrary well-founded games.

**Theorem 14** *Let  $G$  be a well-founded win or lose game. For all players  $i$  we have  $\text{win}_i(G) \neq \emptyset$  iff  $\text{win}_{-i}(G) = \emptyset$ .*

**Proof.** We have the following sequences of equivalences, where  $i = \text{turn}(v)$ :

$$\begin{aligned} & s_i \in \text{win}_i(G) \\ \text{iff} & \quad \{ \text{the definition of } \text{win}_i(G) \} \\ & \text{for all } s_{-i} \in S_{-i}, p_i(\text{leaf}(s_i, s_{-i})) = 1 \\ \text{iff} & \quad \{ i = \text{turn}(v) \} \\ & \text{for all } s_{-i}^w \in S_{-i}^w, p_i(\text{leaf}(s_i^w, s_{-i}^w)) = 1, \text{ where } w = s_i(v) \\ \text{iff} & \quad \{ \text{definition of a winning strategy} \} \\ & s_i^w \in \text{win}_i(G^w), \text{ where } w = s_i(v). \end{aligned}$$

and

$$\begin{aligned} & s_{-i} \in \text{win}_{-i}(G) \\ \text{iff} & \quad \{ \text{the definition of } \text{win}_{-i}(G) \} \\ & \text{for all } s_i \in S_i, p_{-i}(\text{leaf}(s_i, s_{-i})) = 1 \\ \text{iff} & \quad \{ i = \text{turn}(v) \} \\ & \text{for all } w \in C(v) \text{ and } s_i^w \in S_i^w, p_{-i}(\text{leaf}(s_i^w, s_{-i}^w)) = 1 \\ \text{iff} & \quad \{ \text{definition of a winning strategy} \} \\ & \text{for all } w \in C(v), s_{-i}^w \in \text{win}_{-i}(G^w). \end{aligned}$$

We now prove the claim by induction on the rank of the game tree. For game trees of rank 0 the claim clearly holds. Suppose that it holds for all well-founded win or lose games with game trees of rank smaller than some ordinal  $\alpha > 0$  and consider a win or lose game  $G$  with the well-founded game tree of rank  $\alpha$  and rooted at  $v$ . Let  $i = \text{turn}(v)$ .

By the induction hypothesis for all  $w \in C(v)$ ,  $\text{win}_i(G^w) \neq \emptyset$  iff  $\text{win}_{-i}(G^w) = \emptyset$ , so the above equivalences imply the following string of equivalences:

$\text{win}_i(G) \neq \emptyset$  iff for some  $w \in C(v)$ ,  $\text{win}_i(G^w) \neq \emptyset$  iff for some  $w \in C(v)$ ,  $\text{win}_{-i}^w(G) = \emptyset$  iff  $\text{win}_{-i}(G) = \emptyset$   
and hence also  $\text{win}_{-i}(G) \neq \emptyset$  iff  $\text{win}_i(G) = \emptyset$ . □

From Corollary 9 we know that every well-founded win or lose game has a subgame perfect equilibrium, thus in particular a Nash equilibrium. The following result clarifies the relation between Nash equilibria and winning strategies. We denote the set of Nash equilibria in an extensive game  $G$  by  $NE(G)$ .

**Corollary 15** *Consider a well-founded win or lose game  $G$ . For some player  $i$ ,  $NE(G) = \text{win}_i(G) \times S_{-i}$ .*

**Proof.** By Theorem 14  $\text{win}_1(G) \neq \emptyset$  or  $\text{win}_2(G) \neq \emptyset$ . Suppose without loss of generality that  $\text{win}_1(G) \neq \emptyset$ .

( $\Rightarrow$ ) Let  $(s_1, s_2)$  be a Nash equilibrium and  $t_1$  be a winning strategy for player 1. Then we have  $p_1(\text{leaf}(s_1, s_2)) \geq p_1(\text{leaf}(t_1, s_2)) = 1$  and hence  $p_2(\text{leaf}(s_1, s_2)) = -1$ . If  $s_1$  is not a winning strategy for player 1, then for some player 2 strategy  $t_2$  we have  $p_1(\text{leaf}(s_1, t_2)) = -1$ , i.e.,  $p_2(\text{leaf}(s_1, t_2)) = 1 > p_2(\text{leaf}(s_1, s_2))$ , which contradicts the fact that  $(s_1, s_2)$  is a Nash equilibrium. So  $NE(G) \subseteq \text{win}_1(G) \times S_2$ .

( $\Leftarrow$ ) Take a winning strategy  $s_1$  for player 1. Then for all strategies  $t_1$  of player 1 and  $s_2$  and  $t_2$  of player 2,

$$p_1(\text{leaf}(t_1, s_2)) \leq p_1(\text{leaf}(s_1, s_2)) = p_1(\text{leaf}(s_1, t_2)).$$

So  $(s_1, s_2)$  is a Nash equilibrium. Hence  $\text{win}_1(G) \times S_2 \subseteq NE(G)$ . □

In general the sets of subgame perfect equilibria and Nash equilibria differ, so we cannot replace in the above result  $NE(G)$  by  $SPE(G)$ . However, the above corollary directly implies the following characterization of subgame perfect equilibria.

**Corollary 16** *Let  $G$  be a well-founded win or lose game on a game tree  $(V, E)$  with the set of leaves  $Z$ . Then  $SPE(G) = \{s \in S \mid \forall w \in V \setminus Z \exists i [s^w \in \text{win}_i(G^w) \times S_{-i}^w]\}$ .*

It is easy to see that one cannot reverse here the order of the quantifiers.

We now consider a related class of games often called **chess-like games**. These are two-player well-founded extensive games in which the only possible outcomes are  $(1, -1)$ ,  $(0, 0)$ , and  $(-1, 1)$ , with 0 interpreted as a *draw*. We say that a strategy  $s_i$  of player  $i$  in such a game **guarantees him at least a draw** if

$$\forall s_{-i} \in S_{-i} p_i(\text{leaf}(s_i, s_{-i})) \geq 0,$$

and denote the (possibly empty) set of such strategies by  $\text{draw}_i(G)$ .

We now prove the following result for well-founded chess-like games. The set  $\text{win}_i(G)$  is defined as above.

**Theorem 17** *In every well-founded chess-like game  $G$*

$$\text{win}_1(G) \neq \emptyset \text{ or } \text{win}_2(G) \neq \emptyset \text{ or } (\text{draw}_1(G) \neq \emptyset \text{ and } \text{draw}_2(G) \neq \emptyset).$$

It states that in every chess-like game either one of the players has a winning strategy or each player has a strategy that guarantees him at least a draw. These three alternatives are mutually exclusive, since for all  $i \in \{1, 2\}$ ,  $win_i(G) \neq \emptyset$  implies both  $win_{-i}(G) = \emptyset$  and  $draw_{-i}(G) = \emptyset$ .

**Proof.** We introduce the following abbreviations:

- $A$  for  $win_1(G) \neq \emptyset$ ,
- $B$  for  $draw_2(G) \neq \emptyset$ ,
- $C$  for  $win_2(G) \neq \emptyset$ ,
- $D$  for  $draw_1(G) \neq \emptyset$ .

Let  $G_1$  and  $G_2$  be the modifications of  $G$  in which each outcome  $(0, 0)$  is replaced for  $G_1$  by  $(-1, 1)$  and for  $G_2$  by  $(1, -1)$ . Then  $win_1(G_1) = win_1(G)$ ,  $win_2(G_1) = draw_2(G)$ ,  $win_1(G_2) = draw_1(G)$ , and  $win_2(G_2) = win_2(G)$ .

Hence by Theorem 14 applied to the games  $G_1$  and  $G_2$  we have  $A \vee B$  and  $C \vee D$ , so  $(A \wedge C) \vee (A \wedge D) \vee (B \wedge C) \vee (B \wedge D)$ , which implies  $A \vee C \vee (B \wedge D)$ , since  $\neg(A \wedge C)$ ,  $(A \wedge D) \equiv A$ , and  $(B \wedge C) \equiv C$ .  $\square$

For finite games, the above result is formulated in [22, page 125]. The proof first uses backward induction (apparently the first use of it in the literature on game theory) to establish the existence of a Nash equilibrium. Subsequently, (what is now called) the Minimax theorem is invoked to conclude that the payoff to the first player (and hence the second, as well) in any Nash equilibrium is unique. Finally, it is observed that each possible payoff value corresponds to one of the three disjuncts in the above theorem. The above theorem clarifies that this result holds for well-founded games as well, and that it can be proved in a simple way, without the use of backward induction.

In [6] a proof of this result is provided for chess-like games in which infinite plays, interpreted as draw, are allowed. The proof does not rely on backward induction and is also valid for well-founded chess-like games.

**Corollary 18** Consider a well-founded chess-like game  $G$ . For some player  $i$

$$NE(G) = win_i(G) \times S_{-i} \text{ or } NE(G) = draw_i(G) \times draw_{-i}(G).$$

**Proof.** Consider the games  $G_1$  and  $G_2$  from the proof of Theorem 17. We noticed there that  $win_1(G_1) = win_1(G)$ ,  $win_2(G_1) = draw_2(G)$ ,  $win_1(G_2) = draw_1(G)$ , and  $win_2(G_2) = win_2(G)$ .

So if  $win_1(G) \neq \emptyset$ , then by Corollary 15 applied to the game  $G_1$  we get  $NE(G_1) = win_1(G) \times S_2$ , and if  $win_2(G) \neq \emptyset$ , then by Corollary 15 applied to the game  $G_2$  we get  $NE(G_2) = S_1 \times win_2(G)$ .

Suppose now that for both players  $i$ ,  $win_i(G) = \emptyset$ . Then both  $win_1(G_1) = \emptyset$  and  $win_2(G_2) = \emptyset$ , so by Corollary 15 applied to the games  $G_2$  and  $G_1$  we get both  $NE(G_2) = draw_1(G) \times S_2$  and  $NE(G_1) = S_1 \times draw_2(G)$ . This implies  $NE(G_1) \cap NE(G_2) = draw_1(G) \times draw_2(G)$ .

Further, it is easy to see that  $NE(G) \subseteq NE(G_1)$  and  $NE(G) \subseteq NE(G_2)$ . Thus we have established that for some player  $i$

$$NE(G) \subseteq win_i(G) \times S_{-i} \text{ or } NE(G) \subseteq draw_i(G) \times draw_{-i}(G).$$

To complete the proof, let  $p_i$  denote the payoff function of player  $i$  in the game  $G$ . Suppose there exists a player  $i$  such that  $win_i(G) \neq \emptyset$  and let  $s \in win_i(G) \times S_{-i}$ . By the definition of  $win_i(G)$  for all  $s'_{-i}$  we have  $p_i(leaf(s_i, s'_{-i})) = 1$ . Hence, since  $G$  is a zero-sum game, for all  $s'_{-i}$  we have  $p_{-i}(leaf(s_i, s'_{-i})) = -1$ .

Since 1 is a maximum payoff for all  $s'_i$  we also have  $p_i(\text{leaf}(s)) \geq p_i(\text{leaf}(s'_i, s_{-i}))$ . This shows that  $s$  is a Nash equilibrium of  $G$ .

Suppose now that for all players  $i$ ,  $\text{win}_i(G) = \emptyset$ . Fix some  $i \in \{1, 2\}$  and let  $s \in \text{draw}_i(G) \times \text{draw}_{-i}(G)$ . By the definition of the sets  $\text{draw}_i(G)$

- for all  $s'_{-i}$ ,  $p_i(\text{leaf}(s_i, s'_{-i})) \geq 0$  and
- for all  $s'_i$ ,  $p_{-i}(\text{leaf}(s'_i, s_{-i})) \geq 0$ .

Hence, since  $G$  is a zero-sum game,  $p(\text{leaf}(s)) = (0, 0)$  and

- for all  $s'_{-i}$ ,  $p_{-i}(\text{leaf}(s_i, s'_{-i})) \leq 0$  and
- for all  $s'_i$ ,  $p_i(\text{leaf}(s'_i, s_{-i})) \leq 0$ .

This means that  $s$  is a Nash equilibrium of  $G$ . □

**Corollary 19** *Consider a well-founded chess-like game  $G$  on a game tree  $(V, E)$  with the set of leaves  $Z$ . Then*

$$\text{SPE}(G) = \{s \in S \mid \forall w \in V \setminus Z \exists i [s^w \in (\text{win}_i(G^w) \times S_{-i}^w) \cup (\text{draw}_i(G^w) \times \text{draw}_{-i}(G^w))]\}.$$

## 5 Conclusions

In this paper we studied well-founded extensive games with perfect information. We focused on the existence and structural characterization of the sets of subgame perfect equilibria. We also provided such characterizations for two classes of two-player zero-sum games: win or lose games and chess-like games. It will be interesting to consider in this setting other notions and solution concepts that have been well-studied in finite games.

One of them is weak dominance. For finite games, its relation to backward induction was studied in [14]. The authors showed that for finite game that satisfy the TDI condition from Section 3 the elimination of weakly dominated strategies is order independent and is guaranteed to solve the game. The author of [6, 7] studied zero-sum extensive games and showed that every such game with finitely many outcomes can be solved by iterated elimination of weakly dominated strategies.

The definition of weak dominance applies to well-founded extensive games, as well, but the resulting dynamics may be different. For instance, it is possible that the iterated elimination of weakly dominated strategies can then result in empty strategy sets for all or for some players. Also, it may happen that the elimination process has to be iterated over ordinals larger than  $\omega$ . It would be interesting to identify subclasses of well-founded extensive games which can be solved by the iterated elimination of weakly dominated strategies and for which it is order independent.

Another direction is a study of the dynamics of strategy improvement in terms of best (or better) response updates. For finite extensive games, the relation between the improvement dynamics and Nash equilibria was analyzed in [12, 4]. For restricted classes of infinite games of perfect information, improvement dynamics were studied in [3, 18]. It is an interesting question how the improvement dynamics and Nash and subgame perfect equilibria relate in well-founded extensive games.

## Acknowledgements

We would like to thank the reviewers and Marcin Dziubiński for helpful comments. The second author was partially supported by the grant MTR/2018/001244.

## References

- [1] C. Alós-Ferrer & K. Ritzberger (2016): *The Theory of Extensive Form Games*. Springer, doi:10.1007/978-3-662-49944-3.
- [2] P. Battigalli (1997): *On rationalizability in extensive games*. *Journal of Economic Theory* 74, pp. 40–61, doi:10.1006/jeth.1996.2252.
- [3] E. Boros, K. Elbassioni, V. Gurvich & K. Makino (2012): *On Nash equilibria and improvement cycles in pure positional strategies for Chess-like and Backgammon-like n-person games*. *Discrete Mathematics* 312(4), pp. 772–788, doi:10.1016/j.disc.2011.11.011.
- [4] T. Brihaye, G. Geeraerts, M. Hallet & S. Le Roux (2017): *Dynamics and Coalitions in Sequential Games*. In: *Proc. 8th International Symposium on Games, Automata, Logics and Formal Verification*, p. 136–150, doi:10.4204/EPTCS.256.10.
- [5] M. Escardo & P. Oliva (2012): *Computing Nash Equilibria of Unbounded Games*. In: *Turing-100. The Alan Turing Centenary, EPiC Series in Computing* 10, pp. 53–65, doi:10.29007/1wpl.
- [6] C. Ewerhart (2002): *Backward Induction and the Game-Theoretic Analysis of Chess*. *Games and Economic Behaviour* 39, pp. 206–214, doi:10.1006/game.2001.0900.
- [7] C. Ewerhart (2002): *Iterated Weak Dominance in Strictly Competitive Games of Perfect Information*. *Journal of Economic Theory* 107(2), pp. 474–482, doi:10.1006/jeth.2001.2958.
- [8] D. Fudenberg & D. Levine (1983): *Subgame-perfect equilibria of finite and infinite-horizon games*. *Journal of Economic Theory* 31(2), pp. 251–268, doi:10.1016/0022-0531(83)90076-5.
- [9] M.O. Jackson & S. Wilkie (2005): *Endogenous Games and Mechanisms: Side Payments Among Players*. *Review of Economic Studies* 72, p. 543–566, doi:10.1111/j.1467-937X.2005.00342.x.
- [10] M.M. Kaminski (2019): *Generalized Backward Induction: Justification for a Folk Algorithm*. *Games* 34(3), doi:10.3390/g10030034.
- [11] D. König (1927): *Über eine Schlußweise aus dem Endlichen ins Unendliche*. *Acta Litt. Ac. Sci.* 3, pp. 121–130.
- [12] N.S. Kukushkin (2002): *Perfect information and potential games*. *Games and Economic Behavior* 38(2), pp. 306–317, doi:10.1006/game.2001.0859.
- [13] G.J. Mailath & L. Samuelson (2006): *Repeated Games and Reputation: Long-Run Relationships*. Oxford University Press, doi:10.1093/acprof:oso/9780195300796.001.0001.
- [14] L.M. Marx & J.M. Swinkels (1997): *Order Independence for Iterated Weak Dominance*. *Games and Economic Behaviour* 18, pp. 219–245, doi:10.1006/game.1997.0525.
- [15] M.J. Osborne & A. Rubinstein (1994): *A Course in Game Theory*. The MIT Press.
- [16] K. Ritzberger (2001): *Foundations of Non-cooperative Game Theory*. Oxford University Press, Oxford, UK.
- [17] J.C. Rochet (1980): *Selection on an Unique Equilibrium Value for Extensive Games with Perfect Information*. *Cahiers de mathématiques de la décision*, Université Paris IX-Dauphine.
- [18] S. Le Roux & A. Pauly (2020): *A Semi-Potential for Finite and Infinite Games in Extensive Form*. *Dynamic Games and Applications* 10, pp. 120–144, doi:10.1007/s13235-019-00301-7.
- [19] U. Schwalbe & P. Walker (2001): *Zermelo and the Early History of Game Theory*. *Games and Economic Behavior* 34(1), pp. 123–137, doi:10.1006/game.2000.0794.
- [20] R. Selten (1965): *Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit*. *Zeitschrift für die gesamte Staatswissenschaft* 121, pp. 301–324 and 667–689.
- [21] A.S. Troelstra & D. van Dalen (1988): *Constructivism in Mathematics an Introduction (Volume 1)*. *Studies in Logic and the Foundations of Mathematics* 121, Elsevier, doi:10.1016/S0049-237X(09)70523-3.
- [22] J. von Neumann & O. Morgenstern (2004): *Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition)*. Princeton Classic Editions, Princeton University Press.

[23] E. Zermelo (1913): *Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels*. In: *Proc. of The Fifth International Congress of Mathematicians*, Cambridge University Press, pp. 501–504.

## Appendix

**Lemma 6** *Let  $G$  be a well-founded extensive game over the game tree  $T$ . A joint strategy  $s$  is a subgame perfect equilibrium in  $G$  iff for all non-leaf nodes  $u$  in  $T$  and all  $y \in C(u)$*

- $p_i(\text{leaf}(s^x)) \geq p_i(\text{leaf}(s^y))$ , where  $i = \text{turn}(u)$  and  $s_i(u) = x$ .

### Proof.

( $\Rightarrow$ ) Suppose  $s$  is a subgame perfect equilibrium in  $G$ . Consider a non-leaf node  $u$  in  $T$ . Let  $i = \text{turn}(u)$ ,  $x = s_i(u)$  and take some  $y \in C(u)$ . Let  $t_i^u$  be the strategy obtained from  $s_i^u$  by assigning the node  $y$  to  $u$ .

We now have  $p_i(\text{leaf}(s^x)) = p_i(\text{leaf}(s^u)) \geq p_i(\text{leaf}(t_i^u, s_{-i}^u)) = p_i(\text{leaf}(s^y))$ , where the inequality holds by since  $s^u$  a Nash equilibrium in  $G^u$ .

( $\Leftarrow$ ) We proceed by induction on the rank of the game tree of  $G$ . For game trees of rank 0 the induction hypothesis is vacuously true. Suppose the claim holds for all well-founded extensive games whose game tree is of rank smaller than some ordinal  $\alpha > 0$ . Consider a well-founded game  $G$  over a game tree  $T$  of rank  $\alpha$  with the root  $v$ .

Consider any node  $u$  in  $T$  such that  $u \neq v$ . (Since  $\alpha > 0$ , such a node  $u$  exists.) Then  $\text{rank}(T^u)$  is smaller than  $\alpha$  and for all nodes  $w$  in  $T^u$  we have  $(s^u)^w = s^w$ . By the induction hypothesis  $s^u$  is a subgame perfect equilibrium in  $G^u$ , so a fortiori it is a Nash equilibrium in  $G^u$ . It remains to prove that  $s$  is a Nash equilibrium in  $G$ .

Suppose not. Then there exists player  $i$  and  $t_i \in S_i$  such that for  $t = (t_i, s_{-i})$  we have  $p_i(\text{leaf}(s)) < p_i(\text{leaf}(t))$ . Recall that every joint strategy  $s'$  in  $G$  defines a rooted path  $\text{play}(s')$  in  $T$ . By the definition of  $t$  these paths differ for  $s$  and  $t$  at a node at which player  $i$  moves. So for some non-leaf node  $u$  in  $G$  with  $\text{turn}(u) = i$  we have  $\text{play}(s) = \sigma u x \pi_1$  and  $\text{play}(t) = \sigma u y \pi_2$ , where  $\sigma, \pi_1$  and  $\pi_2$  are possibly empty sequences of nodes and  $x \neq y$ . So  $s_i(u) = x$  and  $t_i(u) = y$ .

*Case 1.  $s^y \neq t^y$ .*

Take the first, starting from the root, non-leaf node  $w$  in  $T^y$  such that  $s_i(w) \neq t_i(w)$ . We have  $p_i(\text{leaf}(s)) = p_i(\text{leaf}(s^x)) \geq p_i(\text{leaf}(s^y)) = p_i(\text{leaf}(s^w)) \geq p_i(\text{leaf}(t^w)) = p_i(\text{leaf}(t))$ , where the first inequality holds by the assumptions for the considered implication for the node  $v$  and the second by the fact that  $s^w$  is a Nash equilibrium. So we get a contradiction.

*Case 2.  $s^y = t^y$ .*

We have  $p_i(\text{leaf}(s^x)) = p_i(\text{leaf}(s)) < p_i(\text{leaf}(t)) = p_i(\text{leaf}(t^y)) = p_i(\text{leaf}(s^y))$ . But given that  $s_i(u) = x$  this contradicts the assumption for the node  $u$ .

This concludes the proof. □



# Uncertainty-Based Semantics for Multi-Agent Knowing How Logics

Carlos Areces

Raul Fervari

Andrés R. Saravia

FAMAF, Universidad Nacional de Córdoba, & CONICET, Argentina

{carlos.areces, rfervari}@unc.edu.ar, andresrsaravia@mi.unc.edu.ar

Fernando R. Velázquez-Quesada

ILLC, Universiteit van Amsterdam, The Netherlands

F.R.VelazquezQuesada@uva.nl

We introduce a new semantics for a multi-agent epistemic operator of *knowing how*, based on an indistinguishability relation between plans. Our proposal is, arguably, closer to the standard presentation of *knowing that* modalities in classical epistemic logic. We study the relationship between this semantics and previous approaches, showing that our setting is general enough to capture them. We also define a sound and complete axiomatization, and investigate the computational complexity of its model checking and satisfiability problems.

## 1 Introduction

Epistemic logic (EL; [19, 9]) is a logical formalism tailored for reasoning about the knowledge of abstract autonomous entities, commonly called agents (e.g., a human being, a robot, a vehicle). Most standard epistemic logics deal with an agent’s knowledge about the truth-value of propositions (the notion of *knowing that*). Thus, they focus on the study of sentences like “*the agent knows that it is sunny in Paris*” or “*the robot knows that it is standing next to a wall*”.

At the semantic level, EL formulas are typically interpreted over relational models [6, 5]: essentially, labeled directed graphs. The elements of the domain (called *states* or *worlds*) represent different possible situations. Each agent has associated a relation (interpreted as an *epistemic indistinguishability* relation), used to represent its uncertainty: related states are considered indistinguishable for the agent. An agent is said to know that a proposition  $\varphi$  is true at a given state  $s$  if and only if  $\varphi$  holds in all states she cannot distinguish from  $s$ . It is typically assumed that the indistinguishability relation is an equivalence relation. In spite of its simplicity, this indistinguishability-based representation of knowledge has several advantages. First, it also represents the agent’s *high-order* knowledge (knowledge about her own knowledge and that of other agents). Second, it allows a very natural representation of actions through which knowledge changes (epistemic updates, see, e.g., [7, 4]).

In recent years, other patterns of knowledge besides knowing that have been investigated (see the discussion in [31]). Some examples are *knowing whether* [16, 10], *knowing why* [1, 33] and *knowing the value* [14, 2, 8]. Motivated by different scenarios in philosophy and AI, languages for reasoning about *knowing how* assertions [11] are particularly interesting. Intuitively, an agent

knows how to achieve  $\varphi$  given  $\psi$  if she has the *ability* to guarantee that  $\varphi$  will be the case whenever she is in a situation in which  $\psi$  holds.

There is a large literature connecting knowing how with logics of knowledge and action (see, e.g., [27, 28, 24, 20, 18]). However, these proposals for representing knowing how have been the target of criticisms. The main issue is that a simple combination of standard operators expressing *knowing that* and *ability* (see, e.g., [21]) does not seem to lead to a natural notion of knowing how (see [22, 17] for a discussion).

Taking these considerations into account, [30, 31, 32] introduced a framework based on a knowing how binary modality  $\text{Kh}(\psi, \varphi)$ . At the semantic level, this language is also interpreted over relational models — called in this context labeled transition systems (LTSs). But relations do not represent indistinguishability anymore; they rather describe the actions an agent has at her disposal (in some sense, her *abilities*). An edge labeled  $a$  going from state  $w$  to state  $u$  indicates that the agent can execute action  $a$  to transform state  $w$  into  $u$ . In the proposed semantics,  $\text{Kh}(\psi, \varphi)$  holds if and only if there is a “plan” (a sequence of actions satisfying a constraint called strong executability) in the LTS that unerringly leads from every  $\psi$ -state only to  $\varphi$ -states. Other variants of this knowing how operator follow a similar approach (see [25, 26, 12, 29]).

In these proposals, relations are interpreted as the agent’s available actions (as it is done in, e.g., propositional dynamic logic [15]); and the knowing how of an agent is directly defined by what these actions can achieve. This is in sharp contrast with EL, where relational models have two kinds of information: ontic facts about a given situation (represented by the current state in the model), and the particular perspective that agents have (represented by the possible states available in the model, and their respective indistinguishability relation between them).<sup>1</sup> If one would like to mirror the situation in EL, it seems natural that knowing how should be defined in terms of some kind of indistinguishability over the information provided by an LTS. Such an extended model would be able to capture both the abilities of an agent as given by her available actions, together with the (in)abilities that arise when considering two different actions/plans/executions indistinguishable.

This paper introduces a new semantics for  $\text{Kh}_i(\psi, \varphi)$ , a multi-agent version of the knowing how modality. The crucial idea is the inclusion of a notion of epistemic indistinguishability over plans, in the spirit of the *strategy indistinguishability* of, e.g., [23, 3]. We interpret formulas over an *uncertainty-based LTS* ( $\text{LTS}^U$ ) which is an LTS equipped with an indistinguishability relation over plans. An agent may have different alternatives at her disposal to try to achieve a goal, all “as good as any other” (and in that sense indistinguishable) as far as she can tell. In this way,  $\text{LTS}^U$ s aims to reintroduce the notion of epistemic indistinguishability, now at the level of plans. Moreover, the use of  $\text{LTS}^U$ s leads to a natural definition of operators that represent dynamic aspects of knowing how (e.g., the concept of *learning how* can be modeled by eliminating uncertainty between plans).

**Our contributions.** They can be summarized as follows: (1) We introduce a new semantics for  $\text{Kh}_i(\psi, \varphi)$  (for  $i$  an agent) that reintroduces the notion of epistemic indistinguishability from classical EL. (2) We show that the logic obtained is weaker (and this is an advantage, as we will discuss) than the logic from [30, 31, 32]. Still, the new semantics is general enough to capture previous proposals by imposing adequate conditions on the class of models. (3) We present a sound and complete axiomatization for the logic over the class of all  $\text{LTS}^U$ s. (4) We prove that

<sup>1</sup>Notice that in a multi-agent scenario, all agents share the same ontic information, and differ on their epistemic interpretation of it. We will come back to this later.

the satisfiability problem for the new logic is NP-complete, whereas model checking is in P.

**Outline.** Sec. 2 recalls the framework of [30, 31, 32], including its axiom system. Sec. 3 introduces *uncertainty-based LTS*, indicating how it can be used for interpreting a multi-agent version of the knowing how language, and providing an axiom system in Sec. 3.1. Sec. 3.2 studies the correspondence between our semantics and the ones in the previous proposals. Sec. 3.3 studies the computational complexity of model checking and the satisfiability problem for our logic. Sec. 4 provides conclusions and future lines of research.

## 2 A logic of knowing how

This section recalls the basic *knowing how* framework from [30, 31, 32].

**Syntax and semantics.** Throughout the text, let  $\text{Prop}$  be a countable non-empty set of propositional symbols.

**Definition 2.1** Formulas of the language  $L_{\text{Kh}}$  are given by the following grammar:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid \text{Kh}(\varphi, \varphi),$$

with  $p \in \text{Prop}$ . Other Boolean connectives are defined as usual. Formulas of the form  $\text{Kh}(\psi, \varphi)$  are read as “when  $\psi$  holds, the agent knows how to make  $\varphi$  true”.  $\dashv$

In [30, 31, 32] (and variations like [26, 25]), formulas of  $L_{\text{Kh}}$  are interpreted over *labeled transition systems*: relational models in which the relations describe the state-transitions available to the agent. Throughout the text, let  $\text{Act}$  be a denumerable set of (basic) action names.

**Definition 2.2 (Actions and plans)** Let  $\text{Act}^*$  be the set of finite sequences over  $\text{Act}$ . Elements of  $\text{Act}^*$  are called *plans*, with  $\epsilon$  being the *empty plan*. Given  $\sigma \in \text{Act}^*$ , let  $|\sigma|$  be the length of  $\sigma$  ( $|\epsilon| := 0$ ). For  $0 \leq k \leq |\sigma|$ , the plan  $\sigma_k$  is  $\sigma$ 's initial segment up to (and including) the  $k$ th position (with  $\sigma_0 := \epsilon$ ). For  $0 < k \leq |\sigma|$ , the action  $\sigma[k]$  is the one in  $\sigma$ 's  $k$ th position.  $\dashv$

**Definition 2.3 (Labeled transition systems)** A *labeled transition system* (LTS) for  $\text{Prop}$  and  $\text{Act}$  is a tuple  $\mathcal{S} = \langle W, R, V \rangle$  where  $W$  is a non-empty set of states (also denoted by  $D_{\mathcal{S}}$ ),  $R = \{R_a \subseteq W \times W \mid a \in \text{Act}\}$  is a collection of binary relations on  $W$ , and  $V : W \rightarrow 2^{\text{Prop}}$  is a labelling function. Given an LTS  $\mathcal{S}$  and  $w \in D_{\mathcal{S}}$ , the pair  $(\mathcal{S}, w)$  is a *pointed LTS* (parentheses are usually dropped).  $\dashv$

An LTS describes the *abilities* of the agent; thus, sometimes (e.g., [30, 31, 32]) it is also called an *ability map*. Here are some useful definitions.

**Definition 2.4** Let  $\{R_a \subseteq W \times W \mid a \in \text{Act}\}$  be a collection of binary relations. Define  $R_\epsilon := \{(w, w) \mid w \in W\}$  and, for  $\epsilon \neq \sigma \in \text{Act}^*$  and  $a \in \text{Act}$ ,  $R_{\sigma a} := \{(w, u) \in W \times W \mid \exists v \in W \text{ s.t. } (w, v) \in R_\sigma \text{ and } (v, u) \in R_a\}$ . Take a plan  $\sigma \in \text{Act}^*$ : for  $u \in W$  define  $R_\sigma(u) := \{v \in W \mid (u, v) \in R_\sigma\}$ , and for  $U \subseteq W$  define  $R_\sigma(U) := \bigcup_{u \in U} R_\sigma(u)$ .  $\dashv$

Intuitively, [30, 31, 32] defines that an agent knows how to achieve  $\varphi$  given  $\psi$  when she has an appropriate plan that allows her to go from any situation in which  $\psi$  holds to only states in which  $\varphi$  holds. A crucial part is, then, what “appropriate” is taken to be.

**Definition 2.5 (Strong executability)** Let  $\{R_a \subseteq W \times W \mid a \in \text{Act}\}$  be a collection of binary relations. A plan  $\sigma \in \text{Act}^*$  is *strongly executable* (SE) at  $u \in W$  if and only if  $v \in R_{\sigma_k}(u)$  implies  $R_{\sigma[k+1]}(v) \neq \emptyset$  for every  $k \in [0..|\sigma| - 1]$ . We define the set  $\text{SE}(\sigma) := \{w \in W \mid \sigma \text{ is SE at } w\}$ .  $\dashv$

Thus, strong executability asks for *every* partial execution of the plan (which might be  $\epsilon$ ) to be completed. With this notion, formulas in  $L_{\text{Kh}}$  are interpreted over an LTS as follows (the semantic clause for the Kh modality is equivalent to the one found in the original papers).

**Definition 2.6 ( $L_{\text{Kh}}$  over LTSs)** The relation  $\models$  between a pointed LTS  $\mathcal{S}, w$  (with  $\mathcal{S} = \langle W, R, V \rangle$  an LTS over **Act** and **Prop**) and formulas in  $L_{\text{Kh}}$  (over **Prop**) is defined inductively as follows:

$$\begin{aligned} \mathcal{S}, w \models p & \quad \text{iff}_{\text{def}} \quad w \in V(p), \\ \mathcal{S}, w \models \neg\varphi & \quad \text{iff}_{\text{def}} \quad \mathcal{S}, w \not\models \varphi, \\ \mathcal{S}, w \models \varphi \vee \psi & \quad \text{iff}_{\text{def}} \quad \mathcal{S}, w \models \varphi \text{ or } \mathcal{S}, w \models \psi, \\ \mathcal{S}, w \models \text{Kh}(\psi, \varphi) & \quad \text{iff}_{\text{def}} \quad \exists \sigma \in \text{Act}^* \text{ such that } \mathbf{(Kh-1)} \llbracket \psi \rrbracket^{\mathcal{S}} \subseteq \text{SE}(\sigma) \text{ and } \mathbf{(Kh-2)} R_{\sigma}(\llbracket \psi \rrbracket^{\mathcal{S}}) \subseteq \llbracket \varphi \rrbracket^{\mathcal{S}}, \end{aligned}$$

with  $\llbracket \varphi \rrbracket^{\mathcal{S}} := \{w \in W \mid \mathcal{S}, w \models \varphi\}$  (the elements of  $\llbracket \varphi \rrbracket^{\mathcal{S}}$  are sometimes called  $\varphi$ -states).  $\dashv$

$\text{Kh}(\psi, \varphi)$  holds when there is a plan  $\sigma$  such that, when it is executed at any  $\psi$ -state, it will always complete every partial execution (condition **(Kh-1)**), ending unerringly in states satisfying  $\varphi$  (condition **(Kh-2)**). Notice that Kh acts *globally*, i.e.,  $\llbracket \text{Kh}(\psi, \varphi) \rrbracket^{\mathcal{S}}$  is either  $D_{\mathcal{S}}$  or  $\emptyset$ .

**Axiomatization.** The universal modality [13], interpreted as truth in every state of the model, is definable in  $L_{\text{Kh}}$  as  $\mathbf{A}\varphi := \text{Kh}(\neg\varphi, \perp)$ . This is justified by the following proposition, whose proof relies on the fact that  $\text{Act}^*$  is never empty (it always contains  $\epsilon$ ).

**Proposition 2.1 ([30])** *Let  $\mathcal{S}, w$  be a pointed LTS. Then,  $\mathcal{S}, w \models \text{Kh}(\neg\varphi, \perp)$  iff  $\llbracket \varphi \rrbracket^{\mathcal{S}} = D_{\mathcal{S}}$ .*  $\blacktriangleleft$

<u>Block <math>\mathcal{L}</math>:</u>	TAUT $\vdash \varphi$ for $\varphi$ a propositional tautology	MP	From $\vdash \varphi$ and $\vdash \varphi \rightarrow \psi$ infer $\vdash \psi$
	DISTA $\vdash \mathbf{A}(\varphi \rightarrow \psi) \rightarrow (\mathbf{A}\varphi \rightarrow \mathbf{A}\psi)$	NECA	From $\vdash \varphi$ infer $\vdash \mathbf{A}\varphi$
	TA $\vdash \mathbf{A}\varphi \rightarrow \varphi$		
	4KhA $\vdash \text{Kh}(\psi, \varphi) \rightarrow \mathbf{A}\text{Kh}(\psi, \varphi)$	5KhA	$\vdash \neg\text{Kh}(\psi, \varphi) \rightarrow \mathbf{A}\neg\text{Kh}(\psi, \varphi)$
<u>Block <math>\mathcal{L}_{\text{LTS}}</math>:</u>	EMP $\vdash \mathbf{A}(\psi \rightarrow \varphi) \rightarrow \text{Kh}(\psi, \varphi)$	COMPKh	$\vdash (\text{Kh}(\psi, \varphi) \wedge \text{Kh}(\varphi, \chi)) \rightarrow \text{Kh}(\psi, \chi)$

Table 1: Axiom system  $\mathcal{L}_{\text{Kh}}^{\text{LTS}}$ , for  $L_{\text{Kh}}$  w.r.t. LTSs.

The axiom system  $\mathcal{L}_{\text{Kh}}^{\text{LTS}}$  in Tab. 1 shows that **A** and **Kh** are strongly interconnected.

**Theorem 1 ([30])**  $\mathcal{L}_{\text{Kh}}^{\text{LTS}}$  is sound and strongly complete for  $L_{\text{Kh}}$  w.r.t. the class of all LTSs.  $\blacktriangleleft$

Some axioms deserve comment. If **A** is taken as primitive, and  $\mathbf{A}\varphi$  is interpreted as  $\varphi$  is true at every state in an LTS, then EMP states that if  $\psi \rightarrow \varphi$  is a globally true implication, then given  $\psi$  the agent has the ability to make  $\varphi$  true. In simpler words, global ontic information turns into knowledge. One could argue that, more realistically, there are global truths in a model that are still beyond the abilities of the agent. The case of COMPKh is similar (as it also implies a certain level of omniscience) but perhaps less controversial. It might well be that an agent knows how to make  $\varphi$  true given  $\psi$ , and how to make  $\chi$  true given  $\varphi$ , but still have not worked out how to put the two together to ensure that  $\chi$  given  $\psi$ . As we will see in the next section, both these axioms can be correlated with strong assumptions on the uncertainty relation between plans that an agent might have.

### 3 Uncertainty-based semantics

The LTS-based semantics provides a possible representation of an agent's abilities: the agent knows how to achieve  $\varphi$  given  $\psi$  if and only if there is a plan that, when run at any  $\psi$ -state, will always complete every partial execution, ending unerringly in states satisfying  $\varphi$ . One could argue that this representation involves a certain level of idealization.

Consider an agent *lacking* a certain ability. In the LTS-based semantics, this can only happen when the environment does not provide the required (sequence of) action(s). But one can think of scenarios where an adequate plan exists, and yet the agent lacks the ability for a different reason. Indeed, she might *fail to distinguish* the adequate plan from a non-adequate one, in the sense of not being able to tell that, in general, those plans produce a different outcome. Consider, for example, an agent baking a cake. She might have the ability to do the [nine different mixing methods](#) (beating, blending, creaming, cutting, folding, kneading, sifting, stirring, whipping), and she might even recognize them as different actions. However, she might not be able to perfectly distinguish one from the others in the sense of not recognizing that sometimes they produce different results. In such cases, one would say that the agent does not know how to make certain cake: sometimes she gets good outcomes (when she uses the adequate mixing method) and sometimes she does not.

Indistinguishability among *basic* actions can account for the example above (with each mixing method a basic action). Still, one can also think of situations in which a more general indistinguishability *among plans* is involved. Consider the baking agent again. It is reasonable to assume that she can tell the difference between "adding milk" and "adding flour", but perhaps she does not realize the effect that *the order* for mixing ingredients might have in the final result. Here, the issue is not that she cannot distinguish between basic actions; rather, two plans are indistinguishable because the order of their actions is being considered irrelevant. For a last possibility, the agent might not know that, while opening the oven once to check whether the baking goods are done is reasonable, this must not be done in excess. In this case, the problem is not being able to tell the difference between the effect of executing an action once and executing it multiple times. Thus, even plans of *different length* might be considered indistinguishable.

The previous examples suggest that one can devise a more general representation of an agent's abilities. This representation involves taking into account not only the plans she has available (the LTS structure), but also her skills for telling two different plans apart (a form of *indistinguishability among plans*). As we will see, the use of an indistinguishability relation among plans will also let us define a natural model for a multi-agent scenario. In this setting, agents share the same set of *affordances* (provided by the actual environment), but still have different *abilities* depending on which of these affordances they have available, and how well they can tell these affordances apart.

**Definition 3.1 (Uncertainty-based LTS)** Let  $\text{Agt}$  be a finite non-empty set of agents. A *multi-agent uncertainty-based LTS* ( $\text{LTS}^U$ ) for  $\text{Prop}$ ,  $\text{Act}$  and  $\text{Agt}$  is a tuple  $\mathcal{M} = \langle W, R, \sim, V \rangle$  where  $\langle W, R, V \rangle$  is an LTS and  $\sim$  assigns, to each agent  $i \in \text{Agt}$ , an equivalence *indistinguishability* relation over a non-empty set of plans  $P_i \subseteq \text{Act}^*$ . Given an  $\text{LTS}^U$   $\mathcal{M}$  and  $w \in D_{\mathcal{M}}$ , the pair  $(\mathcal{M}, w)$  (parenthesis usually dropped) is called a *pointed LTS*<sup>U</sup>. ⊣

Intuitively,  $P_i$  is the set of plans that agent  $i$  has at her disposal. Similarly as in classical epistemic logic,  $\sim_i \subseteq P_i \times P_i$  describes agent  $i$ 's indistinguishability. But this time, this relation is not defined over possible states of affairs, but rather over her available plans.

**Remark 3.1** *The following change in notation will simplify some definitions later on, and will make the comparison with the LTS-based semantics clearer. Let  $\langle W, R, \sim, V \rangle$  be an  $LTS^U$  and take  $i \in \text{Agt}$ ; for a plan  $\sigma \in P_i$ , let  $[\sigma]_i$  be its equivalence class in  $\sim_i$  (i.e.,  $[\sigma]_i := \{\sigma' \in P_i \mid \sigma \sim_i \sigma'\}$ ).*

*There is a one-to-one correspondence between  $\sim_i$  and its induced set of equivalence classes  $S_i := \{[\sigma]_i \mid \sigma \in P_i\}$ . Hence, from now on an  $LTS^U$  will be presented as a tuple  $\langle W, R, S, V \rangle$ , with  $S = \{S_i \mid i \in \text{Agt}\}$ . Notice the following properties: (1)  $\pi_1 \neq \pi_2 \in S_i$  implies  $\pi_1 \cap \pi_2 = \emptyset$ , (2)  $P_i = \bigcup_{\pi \in S_i} \pi$  and (3)  $\emptyset \notin S_i$ . ◀*

Given her uncertainty over  $\text{Act}^*$ , the abilities of an agent  $i$  depend not on what a single plan can achieve, but rather on what a set of them can guarantee.

**Definition 3.2** For  $\pi \subseteq \text{Act}^*$  and  $U \cup \{u\} \subseteq W$  define  $R_\pi := \bigcup_{\sigma \in \pi} R_\sigma$ ,  $R_\pi(u) := \bigcup_{\sigma \in \pi} R_\sigma(u)$ , and  $R_\pi(U) := \bigcup_{u \in U} R_\pi(u)$ . ◀

We can now define strong executability for sets of plans.

**Definition 3.3 (Strong executability)** A set of plans  $\pi \subseteq \text{Act}^*$  is *strongly executable* at  $u \in W$  if and only if *every* plan  $\sigma \in \pi$  is *strongly executable* at  $u$ . Hence,  $\text{SE}(\pi) = \bigcap_{\sigma \in \pi} \text{SE}(\sigma)$  is the set of the states in  $W$  where  $\pi$  is strongly executable. ◀

**Definition 3.4 (Kh<sub>i</sub> over LTS<sup>U</sup>s)** The satisfiability relation  $\models$  between a pointed  $LTS^U$   $\mathcal{M}, w$  (with  $\mathcal{M} = \langle W, R, S, V \rangle$  an  $LTS^U$  over  $\text{Act}$ ,  $\text{Prop}$  and  $\text{Agt}$ ) and formulas in the multi-agent version of  $L_{\text{Kh}}$  (denoted by  $L_{\text{Kh}_i}$ , and obtained by replacing  $\text{Kh}$  with  $\text{Kh}_i$ ,  $i \in \text{Agt}$ ) is defined inductively. The atomic and Boolean cases are as before. For *knowing how* formulas,

$\mathcal{M}, w \models \text{Kh}_i(\psi, \varphi)$  *iff<sub>def</sub>*  $\exists \pi \in S_i$  such that **(Kh-1)**  $\llbracket \psi \rrbracket^{\mathcal{M}} \subseteq \text{SE}(\pi)$  and **(Kh-2)**  $R_\pi(\llbracket \psi \rrbracket^{\mathcal{M}}) \subseteq \llbracket \varphi \rrbracket^{\mathcal{M}}$ , with  $\llbracket \varphi \rrbracket^{\mathcal{M}} := \{w \in W \mid \mathcal{M}, w \models \varphi\}$ . ◀

It is worth comparing Def. 2.6 and Def. 3.4. As before,  $\text{Kh}_i(\psi, \varphi)$  acts *globally*. Moreover, we now require *for agent i* to have a *set of plans* satisfying strong executability in every  $\psi$ -state (condition **(Kh-1)**). Still, the set of plans should work as the single plan did before: when executed at  $\psi$ -states, it should end unerringly in states satisfying  $\varphi$  (condition **(Kh-2)**).

The rest of the section is devoted to explore the properties of the logic with our new semantics. Moreover, we compare it to the well-known framework from [30, 31, 32].

### 3.1 Axiomatization

We start by establishing that the universal modality is again definable within  $L_{\text{Kh}_i}$  over  $LTS^U$  (it is crucial that  $S_i \neq \emptyset$  and  $\emptyset \notin S_i$ , see Remark 3.1).

**Proposition 3.1** *Given  $\mathcal{M}, w$  a pointed  $LTS^U$ , then  $(\exists i \in \text{Agt}, \mathcal{M}, w \models \text{Kh}_i(\neg\varphi, \perp))$  iff  $\llbracket \varphi \rrbracket^{\mathcal{M}} = D_{\mathcal{M}}$ . ◀*

Hence, by taking  $\text{A}\varphi := \bigvee_{i \in \text{Agt}} \text{Kh}_i(\neg\varphi, \perp)$  (recall that  $\text{Agt}$  is non-empty and finite) and  $\text{E}\varphi := \neg\text{A}\neg\varphi$ , it turns out that formulas in  $\mathcal{L}$  (first part of Tab. 1) are still valid, generalizing  $\text{Kh}$  to  $\text{Kh}_i$ . As discussed in the next section, some valid formulas in  $\mathcal{L}_{\text{LTS}}$  can be falsified over  $LTS^U$ s. But the weaker theorems of  $\mathcal{L}_{\text{Kh}}^{\text{LTS}}$  shown in Tab. 2 (see Prop. 3.8) are still valid, and can be used to define a complete axiomatic system.

$\text{KhA}$  can be subjected to some of the criticism that apply to  $\text{EMP}$  and  $\text{COMPKh}$  but, in our opinion, to a lesser extent. It implies certain level of idealization, as it entails that the knowing how of an agent is, in a sense, closed under global entailment.  $\text{KhE}$  on the other hand, seems plausible: if  $\text{Kh}_i(\psi, \varphi)$  is not trivial (given that  $\text{E}\psi$  holds), then  $\text{E}\varphi$  should be assured.

---

<u>Block <math>\mathcal{L}_{LTS^U}</math>:</u>	$\text{KhE} \quad \vdash (\text{E}\psi \wedge \text{Kh}_i(\psi, \varphi)) \rightarrow \text{E}\varphi$	$\text{KhA} \quad \vdash (\text{A}(\chi \rightarrow \psi) \wedge \text{Kh}_i(\psi, \varphi) \wedge \text{A}(\varphi \rightarrow \theta)) \rightarrow \text{Kh}_i(\chi, \theta)$
--	--	---

---

Table 2: Axioms  $\mathcal{L}_{LTS^U}$ , for  $L_{\text{Kh}_i}$  w.r.t.  $LTS^U$ s.

Let us define the system  $\mathcal{L}_{\text{Kh}_i}^{LTS^U} := \mathcal{L} + \mathcal{L}_{LTS^U}$  (Tab. 2). We will show that the system is sound and strongly complete over  $LTS^U$ s. The proof of soundness is rather straightforward, thus we will focus on completeness. Following [30, 32], the strategy is to build, for any  $\mathcal{L}_{\text{Kh}_i}^{LTS^U}$ -consistent set of formulas, an  $LTS^U$  satisfying them. Note:

**Proposition 3.2** *The following are theorems of  $\mathcal{L}_{\text{Kh}_i}^{LTS^U}$ :*

$$\text{SCOND:} \quad \vdash \text{A}\neg\psi \rightarrow \text{Kh}_i(\psi, \varphi);$$

$$\text{COND:} \quad \vdash \text{Kh}_i(\perp, \varphi). \quad \blacktriangleleft$$

We proceed with the definition of the *canonical model*.

**Definition 3.5 (Canonical model)** Let  $\Phi$  be the set of all maximally  $\mathcal{L}_{\text{Kh}_i}^{LTS^U}$ -consistent sets (MCS) of formulas in  $L_{\text{Kh}_i}$ . For any  $\Delta \in \Phi$ , define

$$\Delta|_{\text{Kh}_i} := \{\xi \in \Delta \mid \xi \text{ is of the form } \text{Kh}_i(\psi, \varphi)\}, \quad \Delta|_{\text{Kh}} := \bigcup_{i \in \text{Agt}} \Delta|_{\text{Kh}_i}.$$

Let  $\Gamma$  be a set in  $\Phi$ . Define, for each agent  $i \in \text{Agt}$ , the set of basic actions  $\text{Act}_i^\Gamma := \{\langle \psi, \varphi \rangle \mid \text{Kh}_i(\psi, \varphi) \in \Gamma\}$ , and  $\text{Act}^\Gamma := \bigcup_{i \in \text{Agt}} \text{Act}_i^\Gamma$ . Notice that COND implies that  $\text{Kh}_i(\perp, \perp) \in \Gamma$  for every  $i \in \text{Agt}$ ; since there is at least one agent, this implies that  $\text{Act}^\Gamma$  is non-empty, and thus it is an adequate set of actions. It is worth noticing that the set  $\text{Act}^\Gamma$  fixes a new signature. However, since the operators of the language cannot see the names of the actions, we can define a mapping from  $\text{Act}^\Gamma$  to any particular  $\text{Act}$ , to preserve the original signature.

Then, the structure  $\mathcal{M}^\Gamma$ , defined over  $\text{Act}^\Gamma$ ,  $\text{Agt}$  and  $\text{Prop}$ , is the tuple  $\langle W^\Gamma, R^\Gamma, \{S_i^\Gamma\}_{i \in \text{Agt}}, V^\Gamma \rangle$  where

- $W^\Gamma := \{\Delta \in \Phi \mid \Delta|_{\text{Kh}} = \Gamma|_{\text{Kh}}\}$ ,
- $R_{\langle \psi, \varphi \rangle}^\Gamma := \bigcup_{i \in \text{Agt}} R_{\langle \psi, \varphi \rangle^i}^\Gamma$ , with  $R_{\langle \psi, \varphi \rangle^i}^\Gamma := \{(\Delta_1, \Delta_2) \in W^\Gamma \times W^\Gamma \mid \text{Kh}_i(\psi, \varphi) \in \Gamma, \psi \in \Delta_1, \varphi \in \Delta_2\}$ ,
- $S_i^\Gamma := \{\{\langle \psi, \varphi \rangle\} \mid \langle \psi, \varphi \rangle \in \text{Act}_i^\Gamma\}$ ,
- $V^\Gamma(\Delta) := \{p \in \text{Prop} \mid p \in \Delta\}$ . \(\blacktriangleleft\)

If  $\Gamma \in \Phi$ , then  $\mathcal{M}^\Gamma$  is a structure of the required type.

**Proposition 3.3** *The structure  $\mathcal{M}^\Gamma = \langle W^\Gamma, R^\Gamma, \{S_i^\Gamma\}_{i \in \text{Agt}}, V^\Gamma \rangle$  is an  $LTS^U$ .*

*Proof.* It is enough to show that each  $S_i^\Gamma$  defines a partition over a non-empty subset of  $\varphi(\text{Act}^*)$ . First, COND implies  $\text{Kh}_i(\perp, \perp) \in \Gamma$ , so  $\langle \perp, \perp \rangle \in \text{Act}_i^\Gamma$  and hence  $\{\langle \perp, \perp \rangle\} \in S_i^\Gamma$ ; thus,  $\bigcup_{\pi \in S_i} \pi \neq \emptyset$ . Then,  $S_i$  indeed defines a partition over  $\bigcup_{\pi \in S_i} \pi$ : its elements are mutually disjoint (they are singletons with different elements), collective exhaustiveness is immediate and, finally,  $\emptyset \notin S_i^\Gamma$ . ■

Let  $\Gamma \in \Phi$ , the following properties of  $\mathcal{M}^\Gamma$  are useful (proofs are similar to the ones in [32]).

**Proposition 3.4** *For any  $\Delta_1, \Delta_2 \in W^\Gamma$  we have  $\Delta_1|_{\text{Kh}} = \Delta_2|_{\text{Kh}}$ .* \(\blacktriangleleft\)

**Proposition 3.5** *Take  $\Delta \in W^\Gamma$ . If  $\Delta$  has a  $R_{\langle \psi, \varphi \rangle}^\Gamma$ -successor, then every  $\Delta' \in W^\Gamma$  with  $\varphi \in \Delta'$  can be  $R_{\langle \psi, \varphi \rangle}^\Gamma$ -reached from  $\Delta$ .* \(\blacktriangleleft\)

**Proposition 3.6** *Let  $\varphi$  be an  $L_{\text{Kh}_i}$ -formula. If  $\varphi \in \Delta$  for every  $\Delta \in W^\Gamma$ , then  $\text{A}\varphi \in \Delta$  for every  $\Delta \in W^\Gamma$ .* \(\blacktriangleleft\)

**Proposition 3.7** Take  $\psi, \psi', \varphi'$  in  $L_{\text{Kh}_i}$ . Suppose that every  $\Delta \in W^\Gamma$  with  $\psi \in \Delta$  has a  $R_{\langle \psi', \varphi' \rangle}^\Gamma$ -successor. Then,  $A(\psi \rightarrow \psi') \in \Delta$  for all  $\Delta \in W^\Gamma$ . ◀

With these properties at hand, we can prove the truth lemma for  $\mathcal{M}^\Gamma$ .

**Lemma 3.1 (Truth lemma for  $\mathcal{M}^\Gamma$ )** Given  $\Gamma \in \Phi$ , take  $\mathcal{M}^\Gamma = \langle W^\Gamma, R^\Gamma, \{S_i^\Gamma\}_{i \in \text{Agt}}, V^\Gamma \rangle$ . Then, for every  $\Theta \in W^\Gamma$  and every  $\varphi \in L_{\text{Kh}_i}$ ,  $\mathcal{M}^\Gamma, \Theta \models \varphi$  if and only if  $\varphi \in \Theta$ .

*Proof.* The proof is by induction on  $\varphi$ , with the atom and Boolean cases as usual. For the rest:

**Case  $\text{Kh}_i(\psi, \varphi)$ . ( $\Rightarrow$ )** Suppose  $\mathcal{M}^\Gamma, \Theta \models \text{Kh}_i(\psi, \varphi)$ , then consider two cases.

- $\llbracket \psi \rrbracket^{\mathcal{M}^\Gamma} = \emptyset$ . Then, for each  $\Delta \in W^\Gamma$  we have  $\Delta \notin \llbracket \psi \rrbracket^{\mathcal{M}^\Gamma}$ , so  $\psi \notin \Delta$  (by IH) and thus  $\neg\psi \in \Delta$  (by maximal consistency). Hence, by Prop. 3.6,  $A\neg\psi \in \Delta$  for every  $\Delta \in W^\Gamma$ . In particular,  $A\neg\psi \in \Theta$  and thus, by SCOND and MP,  $\text{Kh}_i(\psi, \varphi) \in \Theta$ .
- $\llbracket \psi \rrbracket^{\mathcal{M}^\Gamma} \neq \emptyset$ . By hypothesis, there is  $\langle \psi', \varphi' \rangle \in S_i^\Gamma$  with **(Kh-1)**  $\llbracket \psi \rrbracket^{\mathcal{M}^\Gamma} \subseteq \text{SE}(\langle \psi', \varphi' \rangle)$  and **(Kh-2)**  $R_{\langle \psi', \varphi' \rangle}^\Gamma(\llbracket \psi \rrbracket^{\mathcal{M}^\Gamma}) \subseteq \llbracket \varphi \rrbracket^{\mathcal{M}^\Gamma}$ . In other words, there is  $\langle \psi', \varphi' \rangle \in \text{Act}_i^\Gamma$  such that
  - (Kh-1)** for all  $\Delta \in W^\Gamma$ , if  $\Delta \in \llbracket \psi \rrbracket^{\mathcal{M}^\Gamma}$  then  $\Delta \in \text{SE}(\langle \psi', \varphi' \rangle)$ , so  $\Delta \in \text{SE}(\langle \psi', \varphi' \rangle)$  and therefore  $\Delta$  has a  $R_{\langle \psi', \varphi' \rangle}^\Gamma$ -successor.
  - (Kh-2)** for all  $\Delta' \in W^\Gamma$ , if there is  $\Delta \in \llbracket \psi \rrbracket^{\mathcal{M}^\Gamma}$  such that  $(\Delta, \Delta') \in R_{\langle \psi', \varphi' \rangle}^\Gamma$ , then  $\Delta' \in \llbracket \varphi \rrbracket^{\mathcal{M}^\Gamma}$ .

This case requires three pieces.

- (1) Take any  $\Delta \in W^\Gamma$  with  $\psi \in \Delta$ . Then, by IH,  $\Delta \in \llbracket \psi \rrbracket^{\mathcal{M}^\Gamma}$  and thus, by **(Kh-1)**,  $\Delta$  has a  $R_{\langle \psi', \varphi' \rangle}^\Gamma$ -successor. Thus, every  $\Delta \in W^\Gamma$  with  $\psi \in \Delta$  has such successor; then (Prop. 3.7), it follows that  $A(\psi \rightarrow \psi') \in \Delta$  for every  $\Delta \in W^\Gamma$ . In particular,  $A(\psi \rightarrow \psi') \in \Theta$ .
- (2) From  $\langle \psi', \varphi' \rangle \in \text{Act}_i^\Gamma$  it follows that  $\text{Kh}_i(\psi', \varphi') \in \Gamma$ . But  $\Theta \in W^\Gamma$ , so  $\Theta|_{\text{Kh}} = \Gamma|_{\text{Kh}}$  (by definition of  $W^\Gamma$ ). Hence,  $\text{Kh}_i(\psi', \varphi') \in \Theta$ .
- (3) Since  $\llbracket \psi \rrbracket^{\mathcal{M}^\Gamma} \neq \emptyset$ , there is  $\Delta \in \llbracket \psi \rrbracket^{\mathcal{M}^\Gamma}$ . By **(Kh-1)**,  $\Delta$  should have at least one  $R_{\langle \psi', \varphi' \rangle}^\Gamma$ -successor. Then, by Prop. 3.5, every  $\Delta' \in W^\Gamma$  satisfying  $\varphi' \in \Delta'$  can be  $R_{\langle \psi', \varphi' \rangle}^\Gamma$ -reached from  $\Delta$ ; in other words, every  $\Delta' \in W^\Gamma$  satisfying  $\varphi' \in \Delta'$  is in  $R_{\langle \psi', \varphi' \rangle}^\Gamma(\Delta)$ . But  $\Delta \in \llbracket \psi \rrbracket^{\mathcal{M}^\Gamma}$ , so every  $\Delta' \in W^\Gamma$  satisfying  $\varphi' \in \Delta'$  is in  $R_{\langle \psi', \varphi' \rangle}^\Gamma(\llbracket \psi \rrbracket^{\mathcal{M}^\Gamma})$ . Then, by **(Kh-2)**, every  $\Delta' \in W^\Gamma$  satisfying  $\varphi' \in \Delta'$  is in  $\llbracket \varphi \rrbracket^{\mathcal{M}^\Gamma}$ . By IH on the latter part, every  $\Delta' \in W^\Gamma$  satisfying  $\varphi' \in \Delta'$  is such that  $\varphi \in \Delta'$ . Thus,  $\varphi' \rightarrow \varphi \in \Delta'$  for every  $\Delta' \in W^\Gamma$ , and hence (Prop. 3.6)  $A(\varphi' \rightarrow \varphi) \in \Delta'$  for every  $\Delta' \in W^\Gamma$ . In particular,  $A(\varphi' \rightarrow \varphi) \in \Theta$ .

Thus,  $\{A(\psi \rightarrow \psi'), \text{Kh}_i(\psi', \varphi'), A(\varphi' \rightarrow \varphi)\} \subset \Theta$ . Therefore, by KhA and MP,  $\text{Kh}_i(\psi, \varphi) \in \Theta$ .

**( $\Leftarrow$ )** Suppose  $\text{Kh}_i(\psi, \varphi) \in \Theta$ . Thus (Prop. 3.4),  $\text{Kh}_i(\psi, \varphi) \in \Gamma$ , so  $\langle \psi, \varphi \rangle \in \text{Act}_i^\Gamma$  and therefore  $\langle \psi, \varphi \rangle \in S_i^\Gamma$ . The rest of the proof is split in two cases.

- Suppose there is no  $\Delta_\psi \in W^\Gamma$  with  $\psi \in \Delta$ . Then, by IH, there is no  $\Delta_\psi \in W^\Gamma$  with  $\Delta_\psi \in \llbracket \psi \rrbracket^{\mathcal{M}^\Gamma}$ , that is,  $\llbracket \neg\psi \rrbracket^{\mathcal{M}^\Gamma} = D_{W^\Gamma}$ . Since  $\mathcal{M}^\Gamma$  is in  $\text{LTS}^U$  (Prop. 3.3), the latter yields  $(\mathcal{M}^\Gamma, \Delta) \models \text{Kh}_i(\psi, \chi)$  for any  $i \in \text{Agt}$ ,  $\chi \in L_{\text{Kh}_i}$  and  $\Delta \in W^\Gamma$  (cf. Prop. 3.1); hence,  $(\mathcal{M}^\Gamma, \Theta) \models \text{Kh}_i(\psi, \varphi)$ .
- Suppose there is  $\Delta_\psi \in W^\Gamma$  with  $\psi \in \Delta_\psi$ . It will be shown that the strategy  $\{\langle \psi, \varphi \rangle\} \in S_i^\Gamma$  satisfies the requirements.
  - (Kh-1)** Take any  $\Delta \in \llbracket \psi \rrbracket^{\mathcal{M}^\Gamma}$ . By IH,  $\psi \in \Delta$ . Moreover, from  $\text{Kh}_i(\psi, \varphi) \in \Theta$  and Prop. 3.4 it follows that  $\text{Kh}_i(\psi, \varphi) \in \Delta$ . Then, from  $R_{\langle \psi, \varphi \rangle}^\Gamma$ 's definition, every  $\Delta' \in W^\Gamma$  with  $\varphi \in \Delta'$  is such that  $(\Delta, \Delta') \in R_{\langle \psi, \varphi \rangle}^\Gamma$ , and therefore such that  $(\Delta, \Delta') \in R_{\langle \psi, \varphi \rangle}^\Gamma$ . Now note how,

since there is  $\Delta_\psi \in W^\Gamma$  with  $\psi \in \Delta_\psi$ , there should be  $\Delta_\varphi \in W^\Gamma$  with  $\varphi \in \Delta_\varphi$  (the proof uses KhE and TA). This implies that  $(\Delta, \Delta_\varphi) \in R_{\langle\psi, \varphi\rangle}^\Gamma$  and thus, since  $\langle\psi, \varphi\rangle$  is a basic action,  $\Delta \in SE(\langle\psi, \varphi\rangle)$  so  $\Delta \in SE(\{\langle\psi, \varphi\rangle\})$ . Since  $\Delta$  is an arbitrary state in  $\llbracket\psi\rrbracket^{\mathcal{M}^\Gamma}$ , the required  $\llbracket\psi\rrbracket^{\mathcal{M}^\Gamma} \subseteq SE(\{\langle\psi, \varphi\rangle\})$  follows.

**(Kh-2)** Take any  $\Delta' \in R_{\langle\psi, \varphi\rangle}^\Gamma(\llbracket\psi\rrbracket^{\mathcal{M}^\Gamma})$ . Then, there is  $\Delta \in \llbracket\psi\rrbracket^{\mathcal{M}^\Gamma}$  such that  $(\Delta, \Delta') \in R_{\langle\psi, \varphi\rangle}^\Gamma$ . By definition of  $R^\Gamma$ , it follows that  $\varphi \in \Delta'$  so, by IH,  $\Delta' \in \llbracket\varphi\rrbracket^{\mathcal{M}^\Gamma}$ . Since  $\Delta'$  is an arbitrary state in  $R_{\langle\psi, \varphi\rangle}^\Gamma(\llbracket\psi\rrbracket^{\mathcal{M}^\Gamma})$ , the required  $R_{\langle\psi, \varphi\rangle}^\Gamma(\llbracket\psi\rrbracket^{\mathcal{M}^\Gamma}) \subseteq \llbracket\varphi\rrbracket^{\mathcal{M}^\Gamma}$  follows. ■

**Theorem 2** The axiom system  $\mathcal{L}_{\text{Kh}_i}^{\text{LTS}^\text{U}} := \mathcal{L} + \mathcal{L}_{\text{LTS}^\text{U}}$  (Tab. 2) is sound and strongly complete for  $\text{L}_{\text{Kh}_i}$  w.r.t. the class of all  $\text{LTS}^\text{U}$ s.

*Proof.* Take any  $\mathcal{L}_{\text{Kh}_i}^{\text{LTS}^\text{U}}$ -consistent set of formulas  $\Gamma' \subseteq \text{L}_{\text{Kh}_i}$ . Since  $\text{L}_{\text{Kh}_i}$  is enumerable,  $\Gamma'$  can be extended into a maximally  $\mathcal{L}_{\text{Kh}_i}^{\text{LTS}^\text{U}}$ -consistent set  $\Gamma \supseteq \Gamma'$  by a standard Lindenbaum's construction (see, e.g., [6, Chapter 4]). By Lemma 3.1,  $\Gamma'$  is satisfiable in  $\mathcal{M}^\Gamma$  at  $\Gamma$ . The fact that  $\mathcal{M}^\Gamma$  is in  $\text{LTS}^\text{U}$  (Prop. 3.3) completes the proof. ■

### 3.2 Comparing LTS semantics and $\text{LTS}^\text{U}$ semantics

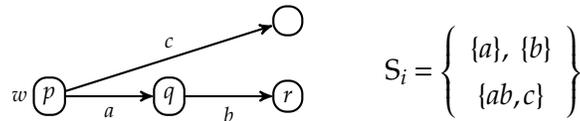
The provided axiom system can be used to compare the notion of *knowing how* under LTSs with that under  $\text{LTS}^\text{U}$ s. Here is a first observation.

**Proposition 3.8** Axioms KhE and KhA are  $\mathcal{L}_{\text{Kh}_i}^{\text{LTS}^\text{U}}$ -derivable (thus,  $\mathcal{L}_{\text{Kh}_i}^{\text{LTS}^\text{U}}$  is a subsystem of  $\mathcal{L}_{\text{Kh}_i}^{\text{LTS}}$ ). ◀

Hence, the *knowing how* operator under LTS is at least as strong as its  $\text{LTS}^\text{U}$ -based counterpart: every formula valid under  $\text{LTS}^\text{U}$ s is also valid under LTSs. The following fact shows that the converse is not the case.

**Proposition 3.9** Within  $\text{LTS}^\text{U}$ , axioms EMP and COMPKh are not valid.

*Proof.* Consider the  $\text{LTS}^\text{U}$   $\mathcal{M}$  shown below, with the collection of sets of available plans for agent  $i$  (i.e., the set  $S_i$ ) depicted on the right. Recall that  $\text{Kh}_i$  acts globally.



With respect to EMP, notice that  $\mathbf{A}(p \rightarrow p)$  holds; yet,  $\text{Kh}_i(p, p)$  fails since there is no  $\pi \in S_i$  leading from  $p$ -states to  $p$ -states. More generally, EMP is valid over LTSs because the empty plan  $\epsilon$ , strongly executable everywhere, is always available. However, in a  $\text{LTS}^\text{U}$ , the plan  $\epsilon$  might not be available to the agent (i.e.,  $\epsilon \notin P_i$ ); and even if  $\epsilon$  is available, it might be indistinguishable from other plans with different behaviour.

With respect to COMPKh, notice that  $\text{Kh}_i(p, q)$  and  $\text{Kh}_i(q, r)$  hold, witness  $\{a\}$  and  $\{b\}$ , resp. However, there is no  $\pi \in S_i$  containing only plans that, when starting on  $p$ -states, lead only to  $r$ -states. This is due to the fact that, although  $ab$  acts as needed, it cannot be distinguished from  $c$ , which behaves differently. Thus,  $\text{Kh}_i(p, r)$  fails. More generally, COMPKh is valid over LTS because the sequential composition of the plans that make true the conjuncts in the antecedent is a witness that makes true the consequent. However, in an  $\text{LTS}^\text{U}$ , this composition might be unavailable or else indistinguishable from other plans. ■

From these two observations it follows that Kh under  $LTS^U$ s is strictly weaker than Kh under LTSs: adding uncertainty about the effect of actions does change the logic. However, the  $LTS^U$  framework is general enough to capture the LTS semantics. To establish the connection, let us work in a single-agent setting (i.e., with a single modality Kh and no subindexes for  $P_i$  and  $S_i$ ).

Given the discussion in Prop. 3.9, it should be clear that there is an obvious class of  $LTS^U$ s in which EMP and COMPKh are valid. This is the class of  $LTS^U$ s in which the agent has every plan available and can distinguish between any two of them (i.e.,  $S = \{\{\sigma\} \mid \sigma \in Act^*\}$ ). This is because, in such models,  $\epsilon$  is available and distinguishable from other plans (for EMP) and from  $\{\sigma_1\} \in S$  and  $\{\sigma_2\} \in S$  it follows that  $\{\sigma_1\sigma_2\} \in S$  (for COMPKh). Clearly, other, more general, classes can be defined, but the one introduced here serves as an example.

**Proposition 3.10** *Let  $\mathcal{S} = \langle W, R, V \rangle$  be an LTS over Act, define  $\mathcal{M}_{\mathcal{S}} = \langle W, R, S, V \rangle$ , where  $S = \{\{\sigma\} \mid \sigma \in Act^*\}$ . Let  $C := \{\mathcal{M}_{\mathcal{S}} \mid \mathcal{S} \text{ is an LTS}\}$ . Given  $\mathcal{M} = \langle W, R, S, V \rangle$  an  $LTS^U$  in  $C$ , define  $\mathcal{S}_{\mathcal{M}} = \langle W, R, V \rangle$ . Then, for every  $\varphi \in L_{Kh}$ ,  $\llbracket \varphi \rrbracket^{\mathcal{S}} = \llbracket \varphi \rrbracket^{\mathcal{M}_{\mathcal{S}}}$  and  $\llbracket \varphi \rrbracket^{\mathcal{M}} = \llbracket \varphi \rrbracket^{\mathcal{S}_{\mathcal{M}}}$ .  $\blacktriangleleft$*

Since we have a class of  $LTS^U$ s in correspondence with the class of all LTSs, we get a direct completeness result:

**Theorem 3** *The axiom system  $\mathcal{L}_{Kh}^{LTS}$  (Tab. 1) is sound and strongly complete for  $L_{Kh}$  w.r.t. the class  $C$ .  $\blacktriangleleft$*

### 3.3 Complexity

Here we investigate the computational complexity of the satisfiability problem of  $L_{Kh_i}$  under the  $LTS^U$ -based semantics. We will establish membership in NP by showing a polynomial size model property.

Given a formula, we will show that it is possible to select just a piece of the canonical model which is relevant for its evaluation. The selected model will preserve satisfiability, and moreover, its size will be polynomial w.r.t. the size of the input formula.

**Definition 3.6 (Selection function)** Let  $\mathcal{M}^{\Gamma} = \langle W^{\Gamma}, R^{\Gamma}, \{S_i^{\Gamma}\}_{i \in \text{Agt}}, V^{\Gamma} \rangle$  be a canonical model for an MCS  $\Gamma$  (see Def. 3.5); take an MCS  $w \in W^{\Gamma}$  and a formula  $\varphi \in L_{Kh_i}$ . Define  $Act_{\varphi} := \{(\theta_1, \theta_2) \in Act^{\Gamma} \mid Kh_i(\theta_1, \theta_2) \text{ is a subformula of } \varphi\}$ . A *canonical selection function*  $sel_w^{\varphi}$  is a function that takes  $\mathcal{M}^{\Gamma}$ ,  $w$  and  $\varphi$  as input, returns a set  $W' \subseteq W^{\Gamma}$ , and is s.t.:

- (1)  $sel_w^{\varphi}(p) = \{w\}$ ;  $sel_w^{\varphi}(\neg\varphi_1) = sel_w^{\varphi}(\varphi_1)$ ;  $sel_w^{\varphi}(\varphi_1 \vee \varphi_2) = sel_w^{\varphi}(\varphi_1) \cup sel_w^{\varphi}(\varphi_2)$ ;
- (2) If  $\llbracket Kh_i(\varphi_1, \varphi_2) \rrbracket^{\mathcal{M}^{\Gamma}} \neq \emptyset$  and  $\llbracket \varphi_1 \rrbracket^{\mathcal{M}^{\Gamma}} = \emptyset$ :  $sel_w^{\varphi}(Kh_i(\varphi_1, \varphi_2)) = \{w\}$ ;
- (3) If  $\llbracket Kh_i(\varphi_1, \varphi_2) \rrbracket^{\mathcal{M}^{\Gamma}} \neq \emptyset$  and  $\llbracket \varphi_1 \rrbracket^{\mathcal{M}^{\Gamma}} \neq \emptyset$ :  
 $sel_w^{\varphi}(Kh_i(\varphi_1, \varphi_2)) = \{w_1, w_2\} \cup sel_{w_1}^{\varphi}(\varphi_1) \cup sel_{w_2}^{\varphi}(\varphi_2)$ , where  $w_1, w_2$  are s.t.  $(w_1, w_2) \in R_{\langle \varphi_1, \varphi_2 \rangle}^{\Gamma}$ ;
- (4) If  $\llbracket Kh_i(\varphi_1, \varphi_2) \rrbracket^{\mathcal{M}^{\Gamma}} = \emptyset$  (note that  $\llbracket \varphi_1 \rrbracket^{\mathcal{M}^{\Gamma}} \neq \emptyset$ ):

For all set of plans  $\pi$ , either  $\llbracket \varphi_1 \rrbracket^{\mathcal{M}^{\Gamma}} \not\subseteq SE(\pi)$  or  $R_{\pi}^{\Gamma}(\llbracket \varphi_1 \rrbracket^{\mathcal{M}^{\Gamma}}) \not\subseteq \llbracket \varphi_2 \rrbracket^{\mathcal{M}^{\Gamma}}$ . For each  $a \in Act_{\varphi}$ :

- (a) if  $\llbracket \varphi_1 \rrbracket^{\mathcal{M}^{\Gamma}} \not\subseteq SE(\{a\})$ : we add  $\{w_1\} \cup sel_{w_1}^{\varphi}(\varphi_1)$  to  $sel_w^{\varphi}(Kh_i(\varphi_1, \varphi_2))$ , where  $w_1 \in \llbracket \varphi_1 \rrbracket^{\mathcal{M}^{\Gamma}}$  and  $w_1 \notin SE(\{a\})$ ;
- (b) if  $R_{\pi}^{\Gamma}(\llbracket \varphi_1 \rrbracket^{\mathcal{M}^{\Gamma}}) \not\subseteq \llbracket \varphi_2 \rrbracket^{\mathcal{M}^{\Gamma}}$  we add  $\{w_1, w_2\} \cup sel_{w_1}^{\varphi}(\varphi_1) \cup sel_{w_2}^{\varphi}(\varphi_2)$  to  $sel_w^{\varphi}(Kh_i(\varphi_1, \varphi_2))$ , where  $w_1 \in \llbracket \varphi_1 \rrbracket^{\mathcal{M}^{\Gamma}}$ ,  $w_2 \in R_a^{\Gamma}(w_1)$  and  $w_2 \notin \llbracket \varphi_2 \rrbracket^{\mathcal{M}^{\Gamma}}$ .  $\dashv$

We can now select a small model which preserves the satisfiability of a given formula.

**Definition 3.7 (Selected model)** Let  $\mathcal{M}^\Gamma$  be the canonical model for an MCS  $\Gamma$ ,  $w$  a state in  $\mathcal{M}^\Gamma$ , and  $\varphi$  an  $\mathsf{L}_{\text{Kh}_i}$ -formula. Let  $\text{sel}_w^\varphi$  be a selection function, we define the *model selected by*  $\text{sel}_w^\varphi$  as  $\mathcal{M}_w^\varphi = \langle W_w^\varphi, R_w^\varphi, \{(S_w^\varphi)_i\}_{i \in \text{Agt}}, V_w^\varphi \rangle$ , where

- $W_w^\varphi := \text{sel}_w^\varphi(\varphi)$ ;
- $(R_w^\varphi)_{\langle \theta_1, \theta_2 \rangle} := R_{\langle \theta_1, \theta_2 \rangle}^\Gamma \cap (W_w^\varphi)^2$  for each  $\langle \theta_1, \theta_2 \rangle \in \text{Act}_\varphi$ ;
- $(S_w^\varphi)_i := \{\{a\} \mid a \in \text{Act}_\varphi\} \cup \{\{\perp, \top\}\}$ , for  $i \in \text{Agt}$  (and  $(R_w^\varphi)_{\langle \perp, \top \rangle} := \emptyset$ );
- $V_w^\varphi$  is the restriction of  $V^\Gamma$  to  $W_w^\varphi$ . ◄

Note that, although  $\text{Act}_\varphi$  can be an empty set, each collection of sets of plans  $(S_w^\varphi)_i$  is not. Therefore,  $\mathcal{M}_w^\varphi$  is an  $\text{LTS}^U$ .

**Proposition 3.11** Let  $\mathcal{M}^\Gamma$  be a canonical model,  $w$  a state in  $\mathcal{M}^\Gamma$  and  $\varphi$  an  $\mathsf{L}_{\text{Kh}_i}$ -formula. Let  $\mathcal{M}_w^\varphi$  be the selected model by a selection function  $\text{sel}_w^\varphi$ .  $\mathcal{M}^\Gamma, w \models \varphi$  implies that for all  $\psi$  subformula of  $\varphi$ , and for all  $v \in W_w^\varphi$  we have that  $\mathcal{M}^\Gamma, v \models \psi$  iff  $\mathcal{M}_w^\varphi, v \models \psi$ . Moreover,  $\mathcal{M}_w^\varphi$  is polynomial on the size of  $\varphi$ . ◀

*Proof.* The proof proceeds by induction in the size of the formula. Boolean cases are simple, so we will proceed with the case in which  $\psi = \text{Kh}_i(\psi_1, \psi_2)$ .

Suppose that  $\mathcal{M}^\Gamma, v \models \text{Kh}_i(\psi_1, \psi_2)$ . Then, we have two cases:

- $\llbracket \psi_1 \rrbracket^{\mathcal{M}^\Gamma} \neq \emptyset$ : by  $\mathcal{M}^\Gamma, v \models \text{Kh}_i(\psi_1, \psi_2)$ , there exists a  $\pi \in S_i^\Gamma$  s.t.  $\llbracket \psi_1 \rrbracket^{\mathcal{M}^\Gamma} \subseteq \text{SE}^{\mathcal{M}^\Gamma}(\pi)$  and  $R_\pi^\Gamma(\llbracket \psi_1 \rrbracket^{\mathcal{M}^\Gamma}) \subseteq \llbracket \psi_2 \rrbracket^{\mathcal{M}^\Gamma}$ . By Truth Lemma,  $\text{Kh}_i(\psi_1, \psi_2) \in v$ , then  $\text{Kh}_i(\psi_1, \psi_2) \in \Gamma$  and  $\langle \psi_1, \psi_2 \rangle \in \text{Act}_\Gamma$ . By the definition of  $R_{\langle \psi_1, \psi_2 \rangle}^\Gamma$ , we have that for all  $w \in \llbracket \psi_1 \rrbracket^{\mathcal{M}^\Gamma}$ , it holds that  $R_{\langle \psi_1, \psi_2 \rangle}^\Gamma(w) \neq \emptyset$  and  $R_{\langle \psi_1, \psi_2 \rangle}^\Gamma(w) \subseteq \llbracket \psi_2 \rrbracket^{\mathcal{M}^\Gamma}$ . Thus,  $\llbracket \psi_1 \rrbracket^{\mathcal{M}^\Gamma} \subseteq \text{SE}^{\mathcal{M}^\Gamma}(\{\langle \psi_1, \psi_2 \rangle\})$  and  $R_{\langle \psi_1, \psi_2 \rangle}^\Gamma(\llbracket \psi_1 \rrbracket^{\mathcal{M}^\Gamma}) \subseteq \llbracket \psi_2 \rrbracket^{\mathcal{M}^\Gamma}$ . Since  $\llbracket \psi_1 \rrbracket^{\mathcal{M}^\Gamma} \neq \emptyset$ , there exist  $w_1, w_2 \in W^\Gamma$  s.t.  $(w_1, w_2) \in R_{\langle \psi_1, \psi_2 \rangle}^\Gamma$ .

Notice that by definition of  $\mathcal{M}_w^\varphi$ , we have that  $\{\langle \psi_1, \psi_2 \rangle\} \in (S_w^\varphi)_i$  and that  $(R_w^\varphi)_{\langle \psi_1, \psi_2 \rangle}$  is defined. Also, by the definition of  $\text{sel}_w^\varphi$ , Item (3), there exist  $w'_1, w'_2 \in W_w^\varphi$  s.t.  $(w'_1, w'_2) \in (R_w^\varphi)_{\langle \psi_1, \psi_2 \rangle}$ . Let  $v_1 \in \llbracket \psi_1 \rrbracket^{\mathcal{M}_w^\varphi} \subseteq \llbracket \psi_1 \rrbracket^{\mathcal{M}^\Gamma}$  (the inclusion holds by IH). Then, we have  $v_1 \in \text{SE}^{\mathcal{M}^\Gamma}(\{\langle \psi_1, \psi_2 \rangle\})$  and  $R_{\langle \psi_1, \psi_2 \rangle}^\Gamma(v_1) \subseteq \llbracket \psi_2 \rrbracket^{\mathcal{M}^\Gamma}$ . Since for all  $v_2 \in R_{\langle \psi_1, \psi_2 \rangle}^\Gamma(v_1)$ , we have  $v_2 \in \llbracket \psi_2 \rrbracket^{\mathcal{M}^\Gamma}$ , (in particular  $v_2 = w'_2$ ), then  $w'_2 \in (R_w^\varphi)_{\langle \psi_1, \psi_2 \rangle}(v_1)$ . Thus,  $v_1 \in \text{SE}^{\mathcal{M}_w^\varphi}(\{\langle \psi_1, \psi_2 \rangle\})$ .

Aiming for a contradiction, suppose now that  $(R_w^\varphi)_{\langle \psi_1, \psi_2 \rangle}(v_1) = R_{\langle \psi_1, \psi_2 \rangle}^\Gamma(v_1) \cap W_w^\varphi \not\subseteq \llbracket \psi_2 \rrbracket^{\mathcal{M}_w^\varphi}$ ; and let  $v_2 \in (R_w^\varphi)_{\langle \psi_1, \psi_2 \rangle}(v_1)$  s.t.  $v_2 \notin \llbracket \psi_2 \rrbracket^{\mathcal{M}_w^\varphi}$ . Then we have that  $(R_w^\varphi)_{\langle \psi_1, \psi_2 \rangle}(v_1) \subseteq R_{\langle \psi_1, \psi_2 \rangle}^\Gamma(v_1)$ , but also by IH  $v_2 \notin \llbracket \psi_2 \rrbracket^{\mathcal{M}^\Gamma}$ . Thus,  $\mathcal{M}^\Gamma, v \not\models \text{Kh}_i(\psi_1, \psi_2)$ , which is a contradiction. Then, it must be the case that  $(R_w^\varphi)_{\langle \psi_1, \psi_2 \rangle}(v_1) \subseteq \llbracket \psi_2 \rrbracket^{\mathcal{M}_w^\varphi}$ . Since we showed that  $\llbracket \psi_1 \rrbracket^{\mathcal{M}_w^\varphi} \subseteq \text{SE}^{\mathcal{M}_w^\varphi}(\{\langle \psi_1, \psi_2 \rangle\})$  and  $(R_w^\varphi)_{\langle \psi_1, \psi_2 \rangle}(\llbracket \psi_1 \rrbracket^{\mathcal{M}_w^\varphi}) \subseteq \llbracket \psi_2 \rrbracket^{\mathcal{M}_w^\varphi}$ , we can conclude  $\mathcal{M}_w^\varphi, v \models \text{Kh}_i(\psi_1, \psi_2)$ .

- $\llbracket \psi_1 \rrbracket^{\mathcal{M}^\Gamma} = \emptyset$ : this case is direct.

Suppose now that  $\mathcal{M}_w^\varphi, v \models \text{Kh}_i(\psi_1, \psi_2)$ :

- $\llbracket \psi_1 \rrbracket^{\mathcal{M}_w^\varphi} \neq \emptyset$ : first, notice that by IH,  $\llbracket \psi_1 \rrbracket^{\mathcal{M}^\Gamma} \neq \emptyset$ . Also, by  $\mathcal{M}_w^\varphi, v \models \text{Kh}_i(\psi_1, \psi_2)$ , we get  $\llbracket \psi_1 \rrbracket^{\mathcal{M}_w^\varphi} \subseteq \text{SE}^{\mathcal{M}_w^\varphi}(\pi')$  and  $(R_w^\varphi)_{\pi'}(\llbracket \psi_1 \rrbracket^{\mathcal{M}_w^\varphi}) \subseteq \llbracket \psi_2 \rrbracket^{\mathcal{M}_w^\varphi}$ , for some  $\pi' \in (S_w^\varphi)_i$ . Aiming for a contradiction, suppose  $\mathcal{M}^\Gamma, v \not\models \text{Kh}_i(\psi_1, \psi_2)$ . This implies that for all  $\pi \in S_i^\Gamma$ ,  $\llbracket \psi_1 \rrbracket^{\mathcal{M}^\Gamma} \not\subseteq \text{SE}^{\mathcal{M}^\Gamma}(\pi)$  or  $R_\pi^\Gamma(\llbracket \psi_1 \rrbracket^{\mathcal{M}^\Gamma}) \not\subseteq \llbracket \psi_2 \rrbracket^{\mathcal{M}^\Gamma}$ . Also, by definition of  $\text{Act}_\varphi$  we have that for all  $\pi =$

$\{a\} \in (S_w^\varphi)_i$ , with  $a \in \text{Act}_\varphi$ ,  $\llbracket \psi_1 \rrbracket^{\mathcal{M}^\Gamma} \not\subseteq \text{SE}^{\mathcal{M}^\Gamma}(\pi)$  or  $R_\pi^\Gamma(\llbracket \psi_1 \rrbracket^{\mathcal{M}^\Gamma}) \not\subseteq \llbracket \psi_2 \rrbracket^{\mathcal{M}^\Gamma}$ ; i.e., for all  $a \in \text{Act}_\varphi$ ,  $\llbracket \psi_1 \rrbracket^{\mathcal{M}^\Gamma} \not\subseteq \text{SE}^{\mathcal{M}^\Gamma}(\{a\})$  or  $R_{\{a\}}^\Gamma(\llbracket \psi_1 \rrbracket^{\mathcal{M}^\Gamma}) \not\subseteq \llbracket \psi_2 \rrbracket^{\mathcal{M}^\Gamma}$ . Thus, there exists  $w_1 \in \llbracket \psi_1 \rrbracket^{\mathcal{M}^\Gamma}$  s.t.  $w_1 \notin \text{SE}^{\mathcal{M}^\Gamma}(a)$  or there exists  $w_2 \in R_a^\Gamma(w_1)$  s.t.  $w_2 \notin \llbracket \psi_2 \rrbracket^{\mathcal{M}^\Gamma}$ . By definition of  $\text{sel}_w^\varphi$ , Item (4), we add witnesses for each  $a \in \text{Act}_\varphi$ . So, let  $\pi' \in (S_w^\varphi)_i$ . If  $\pi' = \{\perp, \top\}$ , trivially we obtain  $\emptyset \neq \llbracket \psi_1 \rrbracket^{\mathcal{M}_w^\varphi} \not\subseteq \text{SE}^{\mathcal{M}_w^\varphi}(\pi') = \emptyset$ . Then, take another  $\pi' = \{a\}$  s.t.  $a \in \text{Act}_\varphi$ , and  $w'_1 \in \llbracket \psi_1 \rrbracket^{\mathcal{M}_w^\varphi} \subseteq \llbracket \psi_1 \rrbracket^{\mathcal{M}^\Gamma}$ . If  $w'_1 \notin \text{SE}^{\mathcal{M}^\Gamma}(\{a\})$ ,  $R_a^\Gamma(w'_1) = \emptyset$  and thus  $(R_w^\varphi)_a(w'_1) = \emptyset$  and therefore  $w'_1 \notin \text{SE}^{\mathcal{M}_w^\varphi}(\{a\})$ . On the other hand, if there exists  $w_2 \in R_a^\Gamma(w'_1)$  s.t.  $w_2 \notin \llbracket \psi_2 \rrbracket^{\mathcal{M}^\Gamma}$ , then by  $\text{sel}_w^\varphi$  and IH, there exists  $w'_2 \in W_w^\varphi$  s.t.  $w'_2 \in R_a^\Gamma(w'_1)$  and  $w'_2 \notin \llbracket \psi_2 \rrbracket^{\mathcal{M}_w^\varphi}$ , and consequently, there exists  $w'_2 \in (R_w^\varphi)_a(w'_1)$  s.t.  $w'_2 \notin \llbracket \psi_2 \rrbracket^{\mathcal{M}_w^\varphi}$ . In any case, it leads to  $\mathcal{M}_w^\varphi, v \not\models \text{Kh}_i(\psi_1, \psi_2)$ , a contradiction. Therefore,  $\mathcal{M}^\Gamma, v \models \text{Kh}_i(\psi_1, \psi_2)$ .

- $\llbracket \psi_1 \rrbracket^{\mathcal{M}_w^\varphi} = \emptyset$ : similar to the previous case.

Thus, we proved the case  $\mathcal{M}^\Gamma, v \models \text{Kh}_i(\psi_1, \psi_2)$  iff  $\mathcal{M}_w^\varphi, v \models \text{Kh}_i(\psi_1, \psi_2)$ . Therefore, we get that for all  $\psi$  subformula of  $\varphi$  and  $v \in W_w^\varphi$ ,  $\mathcal{M}^\Gamma, v \models \psi$  iff  $\mathcal{M}_w^\varphi, v \models \psi$ . Notice that the selection function adds worlds from  $\mathcal{M}^\Gamma$ , only for each  $\text{Kh}_i$ -formula that appears as a subformula of  $\varphi$ . Clearly, there is a polynomial number of such subformulas. Moreover, the number of worlds added at each time is also polynomial in the size of  $\varphi$ . Hence,  $W_w^\varphi$  is of polynomial size. Since  $(S_w^\varphi)_i$  is also polynomial, we have that the size of  $\mathcal{M}_w^\varphi$  is polynomial in the size of  $\varphi$ .

In order to prove that the satisfiability problem of  $L_{\text{Kh}_i}$  is in NP, it remains to show that the model checking problem is in P.

**Proposition 3.12** *The model checking problem for  $L_{\text{Kh}_i}$  is in P.* ◀

*Proof.* Given a pointed LTS<sup>U</sup>  $\mathcal{M}, w$  and a formula  $\varphi$ , we define a bottom-up labeling algorithm running in polynomial time which checks whether  $\mathcal{M}, w \models \varphi$ . We follow the same ideas as for the basic modal logic K (see e.g., [5]). Below we introduce the case for formulas of the shape  $\text{Kh}_i(\psi, \varphi)$ , over an LTS<sup>U</sup>  $\mathcal{M} = \langle W, R, S, V \rangle$ :

```

Procedure ModelChecking( $(\mathcal{M}, w), \text{Kh}_i(\psi, \varphi)$ )
   $lab(\text{Kh}_i(\psi, \varphi)) \leftarrow \emptyset$ ;
  for all  $\pi \in S_i$  do
     $kh \leftarrow \text{True}$ ;
    for all  $\sigma \in \pi$  do
      for all  $v \in lab(\psi)$  do
         $kh \leftarrow (kh \ \& \ v \in \text{SE}(\sigma) \ \& \ R_\sigma(v) \subseteq lab(\varphi))$ ;
      end for
    end for
  if  $kh$  then
     $lab(\text{Kh}_i(\psi, \varphi)) \leftarrow W$ ;
  end if
end for

```

As  $S_i$  and each  $\pi \in S_i$  are not empty, the first two **for** loops are necessarily executed. If  $lab(\psi) = \emptyset$ , then the formula  $\text{Kh}_i(\psi, \varphi)$  is trivially true. Otherwise,  $kh$  will remain true only if the appropriate conditions for the satisfiability of  $\text{Kh}_i(\psi, \varphi)$  hold. If no  $\pi$  succeeds, then the initialization of  $lab(\text{Kh}_i(\psi, \varphi))$  as  $\emptyset$  will not be overwritten, as it should be. Both  $v \in \text{SE}(\sigma)$  and  $R_\sigma$  can be verified in polynomial time. Hence, the model checking problem is in P. ■

The intended result for satisfiability now follows.

**Theorem 4** *The satisfiability problem for  $L_{\text{Kh}_i}$  over  $\text{LTS}^{\text{U}}$ s is NP-complete.*

*Proof.* Hardness follows from NP-completeness of propositional logic (a fragment of  $L_{\text{Kh}_i}$ ). By Prop. 3.11, each satisfiable formula  $\varphi$  has a model of polynomial size on  $\varphi$ . Thus, we can guess a polynomial model  $\mathcal{M}, w$ , and verify  $\mathcal{M}, w \models \varphi$  (which can be done in polynomial time, due to Prop. 3.12). Thus, the satisfiability problem is in the class NP. ■

## 4 Final Remarks

In this article, we introduce a new semantics for the *knowing how* modality from [30, 31, 32], over multiple agents. It is defined in terms of *uncertainty-based labeled transition systems* ( $\text{LTS}^{\text{U}}$ ). The novelty in our proposal is that  $\text{LTS}^{\text{U}}$ s are equipped with an indistinguishability relation among plans. In this way, the epistemic notion of uncertainty of an agent –which in turn defines her epistemic state– is reintroduced, bringing the notion of *knowing how* closer to the notion of *knowing that* from classical epistemic logics. We believe that the semantics based on  $\text{LTS}^{\text{U}}$  can represent properly the situation of a shared, objective description of the affordances of a given situation, together with the different, subjective and personal abilities of a group of agents; this seems difficult to achieve using a semantics based on LTSs alone.

We show that the logic of [30, 31, 32] can be obtained by imposing particular conditions over  $\text{LTS}^{\text{U}}$ s; thus, the new semantics is more general. In particular, it provides counter-examples to EMP and COMP, which directly link Kh to properties of the universal modality.<sup>2</sup> Indeed, consider EMP: even though  $A(\psi \rightarrow \varphi)$  objectively holds in the underlying LTS of an  $\text{LTS}^{\text{U}}$ , it could be argued that an agent might not have actions or plans at her disposal to turn those facts into knowledge, resulting in  $\text{Kh}(\psi, \varphi)$  failing on the model. Moreover, we have introduced a sound and strongly complete axiom system for the new semantics over  $\text{LTS}^{\text{U}}$ s. Finally, we showed that the satisfiability problem for our multi-agent knowing how logic over the  $\text{LTS}^{\text{U}}$ -based semantics is NP-complete, via a selection argument (and model checking is polynomial).

**Future work.** There are several interesting lines of research to explore in the future. First, our framework easily accommodates other notions of executability. For instance, one could require only some of the plans in a set  $\pi$  to be strongly executable, weaken the condition of *strong* executability, etc. We can also explore the effects of imposing different restrictions on the construction of the indistinguishability relation between plans. It would be interesting to investigate which logics we obtain in these cases, and their relations with the LTS semantics.

Second, to our knowledge, the exact complexity of the satisfiability problem for knowing how over LTSs is open. It would be interesting to see whether an adaptation of our selection procedure works over LTSs.

Third, the  $\text{LTS}^{\text{U}}$  semantics, in the multi-agent setting, leads to natural definitions of concepts such as *global*, *distributed* and *common knowing how*, which should be investigated in detail.

Finally, dynamic modalities capturing epistemic updates can be defined via operations that modify the indistinguishability relation among plans (as is done with other dynamic epistemic operators, see, e.g., [7]). This would allow to express different forms of communication, such as *public*, *private* and *semi-private* announcements concerning (sets of) plans.

---

<sup>2</sup>The rest of the axioms and rules in  $\mathcal{L}_{\text{Kh}}^{\text{LTS}}$  (those shown in block  $\mathcal{L}$ ) merely state properties of the universal modality and the fact that Kh is global.

**Acknowledgments.** This work is partially supported by projects ANPCyT-PICT-2017-1130, Stic-AmSud 20-STIC-03 ‘DyLo-MPC’, Secyt-UNC, GRFT Mincyt-Cba, and by the Laboratoire International Associé SINFIN.

## References

- [1] S. Artemov (2008): *The logic of justification*. *The Review of Symbolic Logic* 1(04), pp. 477–513, doi:[10.1017/S1755020308090060](https://doi.org/10.1017/S1755020308090060).
- [2] A. Baltag (2016): *To Know is to Know the Value of a Variable*. In: *Advances in Modal Logic 11*, pp. 135–155.
- [3] F. Belardinelli (2014): *Reasoning about Knowledge and Strategies: Epistemic Strategy Logic*. In: *Proceedings of SR 2014*, pp. 27–33, doi:[10.4204/EPTCS.146.4](https://doi.org/10.4204/EPTCS.146.4).
- [4] J. van Benthem (2011): *Logical Dynamics of Information and Interaction*. Cambridge University Press, doi:[10.1017/CBO9780511974533](https://doi.org/10.1017/CBO9780511974533).
- [5] P. Blackburn & J. van Benthem (2006): *Modal Logic: A Semantic Perspective*. In: *Handbook of Modal Logic*, Elsevier, pp. 1–84, doi:[10.1016/s1570-2464\(07\)80004-8](https://doi.org/10.1016/s1570-2464(07)80004-8).
- [6] P. Blackburn, M. de Rijke & Y. Venema (2002): *Modal Logic*. Cambridge University Press, doi:[10.1017/CBO9781107050884](https://doi.org/10.1017/CBO9781107050884).
- [7] H. van Ditmarsch, W. van der Hoek & B. Kooi (2007): *Dynamic Epistemic Logic*. Springer, doi:[10.1007/978-1-4020-5839-4](https://doi.org/10.1007/978-1-4020-5839-4).
- [8] J. van Eijck, M. Gatteringer & Y. Wang (2017): *Knowing Values and Public Inspection*. In: *Logic and Its Applications - 7th Indian Conference, ICLA 2017*, pp. 77–90, doi:[10.1007/978-3-662-54069-5\\_7](https://doi.org/10.1007/978-3-662-54069-5_7).
- [9] R. Fagin, J. Y. Halpern, Y. Moses & M. Y. Vardi (1995): *Reasoning about knowledge*. The MIT Press, Cambridge, Mass., doi:[10.7551/mitpress/5803.001.0001](https://doi.org/10.7551/mitpress/5803.001.0001).
- [10] J. Fan, Y. Wang & H. van Ditmarsch (2015): *Contingency and Knowing Whether*. *The Review of Symbolic Logic* 8, pp. 75–107, doi:[10.1017/S1755020314000343](https://doi.org/10.1017/S1755020314000343).
- [11] J. Fantl (2017): *Knowledge How*. In E. Zalta, editor: *The Stanford Encyclopedia of Philosophy*, fall 2017 edition, Metaphysics Research Lab, Stanford University.
- [12] R. Fervari, A. Herzig, Y. Li & Y. Wang (2017): *Strategically knowing how*. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*, pp. 1031–1038, doi:[10.24963/ijcai.2017/143](https://doi.org/10.24963/ijcai.2017/143).
- [13] V. Goranko & S. Passy (1992): *Using the Universal Modality: Gains and Questions*. *Journal of Logic and Computation* 2(1), pp. 5–30, doi:[10.1093/logcom/2.1.5](https://doi.org/10.1093/logcom/2.1.5).
- [14] T. Gu & Y. Wang (2016): *“Knowing value” logic as a normal modal logic*. In: *Advances in Modal Logic 11*, pp. 362–381.
- [15] D. Harel, D. Kozen & J. Tiuryn (2000): *Dynamic Logic*. The MIT Press, doi:[10.7551/mitpress/2516.001.0001](https://doi.org/10.7551/mitpress/2516.001.0001).
- [16] S. Hart, A. Heifetz & D. Samet (1996): *Knowing Whether, Knowing That, and The Cardinality of State Spaces*. *Journal of Economic Theory* 70(1), pp. 249–256, doi:[10.1006/jeth.1996.0084](https://doi.org/10.1006/jeth.1996.0084).
- [17] A. Herzig (2015): *Logics of knowledge and action: critical analysis and challenges*. *Autonomous Agents and Multi-Agent Systems* 29(5), pp. 719–753, doi:[10.1007/s10458-014-9267-z](https://doi.org/10.1007/s10458-014-9267-z).
- [18] A. Herzig & N. Troquard (2006): *Knowing how to play: uniform choices in logics of agency*. In: *Proceedings of AAMAS 06*, pp. 209–216, doi:[10.1145/1160633.1160666](https://doi.org/10.1145/1160633.1160666).
- [19] J. Hintikka (1962): *Knowledge and Belief*. Cornell University Press, Ithaca N.Y.
- [20] W. van der Hoek, B. van Linder & J.-J. Ch. Meyer (2000): *On Agents That Have the Ability to Choose*. *Studia Logica* 66(1), pp. 79–119, doi:[10.1023/A:1026796912842](https://doi.org/10.1023/A:1026796912842).

- [21] W. van der Hoek & A. Lomuscio (2003): *Ignore at your peril – towards a logic for ignorance*. In: *Proceedings of AAMAS 03*, pp. 1148–1149, doi:[10.1145/860575.860839](https://doi.org/10.1145/860575.860839).
- [22] W. Jamroga & T. Ågotnes (2007): *Constructive knowledge: what agents can achieve under imperfect information*. *Journal of Applied Non-Classical Logics* 17(4), pp. 423–475, doi:[10.3166/jancl.17.423-475](https://doi.org/10.3166/jancl.17.423-475).
- [23] W. Jamroga & W. van der Hoek (2004): *Agents that Know How to Play*. *Fundamenta Informaticae* 63(2-3), pp. 185–219.
- [24] Y. Lespérance, H. J. Levesque, F. Lin & R. B. Scherl (2000): *Ability and Knowing How in the Situation Calculus*. *Studia Logica* 66(1), pp. 165–186, doi:[10.1023/A:1026761331498](https://doi.org/10.1023/A:1026761331498).
- [25] Y. Li (2017): *Stopping Means Achieving: A Weaker Logic of Knowing How*. *Studies in Logic* 9(4), pp. 34–54.
- [26] Y. Li & Y. Wang (2017): *Achieving While Maintaining: A Logic of Knowing How with Intermediate Constraints*. In: *Logic and Its Applications - 7th Indian Conference, ICLA 2017*, pp. 154–167, doi:[10.1007/978-3-662-54069-5\\_12](https://doi.org/10.1007/978-3-662-54069-5_12).
- [27] J. McCarthy & P. J. Hayes (1969): *Some Philosophical Problems from the Standpoint of Artificial Intelligence*. In: *Machine Intelligence*, Edinburgh University Press, pp. 463–502.
- [28] R. Moore (1985): *A formal theory of knowledge and action*. In: *Formal Theories of the Commonsense World*, Ablex Publishing Corporation.
- [29] X. Wang (2019): *A Logic of Knowing How with Skippable Plans*. In: *Logic, Rationality, and Interaction – 7th International Workshop, LORI 2019*, pp. 413–424, doi:[10.1007/978-3-662-60292-8\\_30](https://doi.org/10.1007/978-3-662-60292-8_30).
- [30] Y. Wang (2015): *A Logic of Knowing How*. In: *Logic, Rationality, and Interaction – 5th International Workshop, LORI 2015*, pp. 392–405, doi:[10.1007/978-3-662-48561-3\\_32](https://doi.org/10.1007/978-3-662-48561-3_32).
- [31] Y. Wang (2018): *Beyond knowing that: a new generation of epistemic logics*. In H. van Ditmarsch & G. Sandu, editors: *J. Hintikka on knowledge and game theoretical semantics*, Springer, pp. 499–533, doi:[10.1007/978-3-319-62864-6\\_21](https://doi.org/10.1007/978-3-319-62864-6_21).
- [32] Y. Wang (2018): *A logic of goal-directed knowing how*. *Synthese* 195(10), pp. 4419–4439, doi:[10.1007/s11229-016-1272-0](https://doi.org/10.1007/s11229-016-1272-0).
- [33] C. Xu, Y. Wang & T. Studer (2021): *A Logic of Knowing Why*. *Synthese* 198, pp. 1259–1285, doi:[10.1007/s11229-019-02104-0](https://doi.org/10.1007/s11229-019-02104-0).



# Revisiting Epistemic Logic with Names\*

Marta Bílková

Czech Acad Sci, Inst Comp Sci  
bilkova@cs.cas.cz

Zoé Christoff

University of Groningen  
z.l.christoff@rug.nl

Olivier Roy

University of Bayreuth  
olivier.roy@uni-bayreuth.de

This paper revisits the multi-agent epistemic logic presented in [10], where agents and sets of agents are replaced by abstract, intensional “names”. We make three contributions. First, we study its model theory, providing adequate notions of bisimulation and frame morphisms, and use them to study the logic’s expressive power and definability. Second, we show that the logic has a natural neighborhood semantics, which in turn allows to show that the axiomatization in [10] does not rely on possibly controversial introspective properties of knowledge. Finally, we extend the logic with common and distributed knowledge operators, and provide a sound and complete axiomatization for each of these extensions. These results together put the original epistemic logic with names in a more modern context and opens the door for a logical analysis of epistemic phenomena where group membership is uncertain or variable.

In [10], Grove and Halpern studied a generalized version of multi-agents epistemic logic where the usual labels for agents and sets of agents are replaced by abstract names whose extension might vary from state to state.<sup>1</sup> Despite being interpreted in standard multi-agents epistemic models, the resulting language does away with the familiar  $K_i$  modalities, and instead contains two families of epistemic operators:  $S_n$ , standing for “someone with name  $n$  knows”, and  $E_n$ , standing for “everyone with name  $n$  knows”.

This generalization is conceptually important. The “names” that index the  $S_n$  and the  $E_n$  modalities can refer intensionally to both individuals and groups. Since these extensions are not fixed in a given model, the logic allows to study social-epistemic phenomena that involve uncertainty or variability in the agents’ identities or group membership. These phenomena are pervasive. [19, 10] already provide convincing examples for distributed systems. Massive coordinated actions or social movements, especially online, also provide contemporary cases [1, 4], where for instance we refer to group labels like “Trump supporters” or “trolls” without knowing exactly who the members of these groups are or even failing to know whether we, ourselves, are members of those groups.<sup>2</sup> The study in [10], however, focuses on the two modalities mentioned above, and in particular leaves aside notions of common and distributed knowledge. These notions are, however, central to theories of social conventions [16, 2] and collective action [25]. [19] study a closely related notion of common knowledge, to which we come back briefly in Section 4, but do not provide an axiomatization. Distributed knowledge for intensional or indexical group names has been studied in [20], but in a more expressive language with explicit quantification.

---

\*The research of Marta Bílková was supported by RVO: 67985807. The research of Zoé Christoff and Olivier Roy was partly supported by the DFG-GACR project “SEGA” (RO 4548/6-1). This publication is also part of Zoé Christoff’s project “Democracy on Social Networks” (VI.Veni.201F.032) of the research programme VENI financed by the Dutch Research Council (NWO).

<sup>1</sup>The idea of replacing standard labels with abstract names appeared earlier, for instance in [7] and [19]. The former define a notion of belief as a “society of minds” along the lines of the operator  $S_n$  defined below. The latter define the notion of “everyone in a group  $n$ ” following the same semantic idea as for the operator written  $E_n$  here, and use it to define a notion of implicit common knowledge like the one later studied in [8]. We briefly come back to this notion in Section 4. [7] proposes an axiomatization of belief as society of minds notions, but not together with the  $E_n$  modality, as [10] do. [19] do not axiomatize the notion of common knowledge they put forward.

<sup>2</sup>Some of these phenomena have been studied using tools from multi-agent epistemic logic, c.f. [24, 6].

Epistemic logic with names is also technically interesting. Even though the idea appears earlier [28, 18, 13], Grove and Halpern’s contribution, as well as [9], has been seminal for the development of so-called term modal logic—c.f. [17] and the references therein—which in turn have helped to understand, among others, *de dicto* and *de re* knowledge attributions. Since many term modal logics turn out to be undecidable, one important question in that literature has been to identify decidable fragments—c.f. [26] and the references therein for the case of epistemic logic. The basic system in [10] is one of them. That paper, however, does not address questions of definability, invariance, expressive power, or proof theory. [22] makes headway in that direction for a very closely related, but still non-equivalent language.

Epistemic logic with names thus stands at the crossroad of important conceptual questions regarding group knowledge and group agency, on the one hand, and technical questions in the landscape of extended modal languages, on the other. As we show below, this generalization of epistemic logic can also be studied from the perspective of neighborhood semantics [21], bringing a third tradition to bear on the understanding of this system. Although some of these areas have had contacts with each other, many dots still need to be connected, which is what this paper sets itself to do.

After introducing the basics of epistemic logic with names, we start by formulating adequate notions of bisimulations and frame morphisms for this logic. Based on the observation that these notions are structurally similar to the corresponding notions for neighborhood frames [11, 12], we show that the basic system can indeed be given a natural interpretation in neighborhood semantics. This allows to import a number of results regarding definability and expressive power from non-normal modal logic, as well as to show that this basic system is actually not dependent on assuming that the agent’s epistemic state is represented by partitions/equivalence relations. We finally turn to group notions, showing sound and complete axiomatizations of common and distributed knowledge with names.

## 1 Epistemic Logic with Names

Epistemic logic with names replaces the familiar individual epistemic modalities  $K_i$ , standing for “agent  $i$  knows...”, with modalities  $E_n$  for “everybody with name  $n$  knows” and  $S_n$  for “somebody with name  $n$  knows”.

**Definition 1** (syntax). *The language  $\mathcal{L}_N$  is defined as follows:*

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \varphi \mid E_n\varphi \mid S_n\varphi$$

where  $p \in \Phi$  with  $\Phi$  a countable set of atomic propositions;  $n \in N \subseteq \mathbb{N}$  is a name.

The basic idea is that these “names” can refer both to individuals and to groups, or even not refer to anyone at all, and that these references are intensional. Who is named by  $n$  might change from state to state. In what follows we often refer to the agent(s) named by  $n$  at the particular state as the “group  $n$ ” in that state. Beside this modification, however, in [10] this language is interpreted in standard epistemic, i.e. “partitional” or “S5” models.

**Definition 2** (frames and models). *Let  $N$  be a given set of names. A frame over  $N$  is a tuple  $\mathcal{F} = (W, A, \mathcal{R}, \mu)$  where:*

- $W, A$  are (nonempty) sets of states and agents, respectively.
- $\mu : W \times N \rightarrow \mathcal{P}(A)$  is a naming function that assigns to each world and name the set of agents that have that name in this world.

- $\mathcal{R} : A \rightarrow W \times W$  assigns to each agent a reflexive binary relation on  $W$  such that  $wR_a w'$  only if  $a \in \mu(w, n)$  for some  $n$ . When each  $R_a$  is an equivalence relation we call  $F$  an epistemic frame. We often write  $R_a(w)$  for  $\{v \mid wR_a v\}$ .

An (epistemic) model over  $N$  and  $\Phi$  is an (epistemic) frame over  $N$  together with a valuation function  $\pi$  for a given set of atomic sentences  $\Phi$ .

This definition is slightly different from the one in [10]. First they use an existence function  $\alpha : W \rightarrow \mathcal{P}(A)$ , which tells at each state which agents exist, and require that only existing agents have a name. We instead implicitly define existence using the naming function. To exist is to have a name, or be a group member. It can easily be shown that truth in the language above is invariant under this modification. In [10], each  $R_a$  is furthermore only allowed to connect the current state to ones where  $a$  exists, which we do not assume here.

Truth and validity for this language are defined as expected, revealing the implicit quantification over agents behind the  $S_n$  and the  $E_n$  modalities. The clauses for the atomic formulas and Boolean connectives are standard.

**Definition 3** (satisfaction). Let  $M$  be a model and  $w \in W$ :

$$M, w \models E_n \varphi \text{ iff for all } a \in \mu(w, n) : \forall w' \in R_a(w), M, w' \models \varphi$$

$$M, w \models S_n \varphi \text{ iff for some } a \in \mu(w, n) : \forall w' \in R_a(w), M, w' \models \varphi$$

[10] work with epistemic frames, but the logical behavior of  $E_n$  and  $S_n$  is much weaker than the usual  $S5$  individual knowledge operators. Neither satisfy positive nor negative introspection,  $E_n$  does not satisfy  $T$  and  $S_n$  is not normal. The following example illustrates this:

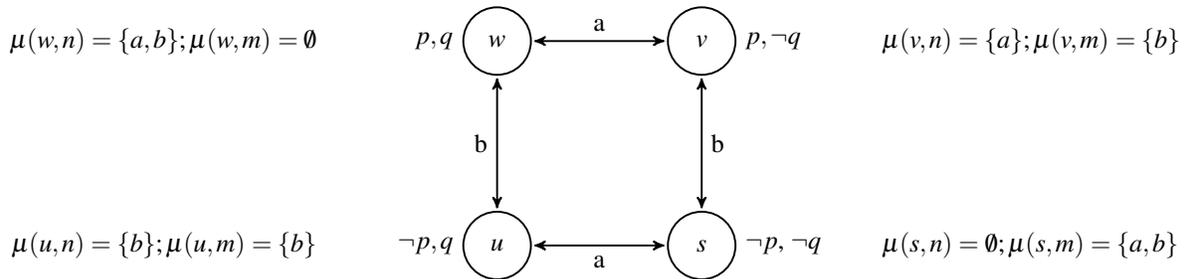


Figure 1: An epistemic model with names. The labeled arrows represent agents  $a$  and  $b$ 's respective indistinguishably relations (we omit reflexive arrows).

In the model depicted in Figure 1, agent  $a$  knows whether  $p$ , say “Trump was impeached”, but does not know whether  $q$ , say “Trump lost the election”, and vice versa for agent  $b$ . Consider state  $w$ , where both  $a$  and  $b$  are labeled by  $n$ , say “Trump supporters”, while neither of them is labeled  $m$ , say “trolls”. Agent  $a$  knows that she herself is a Trump supporter and that  $b$  is either a fellow Trump supporter or a troll. In  $w$ , some but not all Trump supporters know that Trump was impeached,  $w \models S_n p \wedge \neg E_n p$ . And since there is no troll in  $w$ , no troll knows that Trump was impeached, while at the same time, trivially, all trolls know that he was impeached, in as much as they know that he was not,  $w \models \neg S_m p \wedge E_m p \wedge E_m \neg p$ . In state  $u$ , some troll knows that Trump lost the election but it is not the case that some troll knows that some troll knows that Trump lost,  $u \models S_m q \wedge \neg S_m S_m q$ . And in state  $s$ , no Trump supporter knows that Trump was impeached yet no supporter knows that no supporter knows it,  $s \models \neg S_n p \wedge \neg S_n \neg S_n p$ .

The comparatively weak logical behavior of the two modalities is reflected at the axiomatic level. The system below is indeed shown in [10] to be sound and complete with respect to the special class of *epistemic frames*, even though the axioms 4 and 5 are absent for both  $S_n$  and  $E_n$ .

**Definition 4** (axiom system  $AX_{\mathcal{N}}$ ). *The system  $AX_{\mathcal{N}}$  comprises the following axioms and rules:*

$PL$	<i>All instances of propositional tautologies</i>
$MP$	<i>From <math>\varphi</math> and <math>\varphi \rightarrow \psi</math>, infer <math>\psi</math></i>
$T(S_n)$	$S_n \varphi \rightarrow \varphi$
$K(E_n)$	$E_n \varphi \wedge E_n(\varphi \rightarrow \psi) \rightarrow E_n \psi$
$Nec(E_n)$	<i>From <math>\varphi</math>, infer <math>E_n \varphi</math></i>
$Int_1$	$S_n \varphi \wedge E_n(\varphi \rightarrow \psi) \rightarrow S_n \psi$
$Int_2$	$\neg E_n \perp \rightarrow S_n \top$

[10] study further extensions of this system, to cover for instance cases where no two agents have the same name or every agent has its own proper name. They study in greater detail the case where the agents know their own names/which groups they belong to, which, interestingly, shed light on the source of introspection for standard epistemic modalities. Since in this paper we focus on the general system as defined above, we leave the discussion of these special cases for future work.

## 2 Morphisms and bisimulations

To allow for further model-theoretic considerations, we start with the definition of an adequate notion of frame morphisms and bisimulations for epistemic logic with names.

**Definition 5.** *A frame morphism from a frame  $F = (W, A, \mathcal{R}, \mu)$  to a frame  $F' = (W', A', \mathcal{R}', \mu')$  is a map  $f : W \rightarrow W'$  satisfying the following conditions:*

$$(there) \quad \forall a \in \mu(w, n) \exists a' \in \mu'(f(w), n) R'_{a'}(f(w)) = f[R_a(w)]$$

$$(back) \quad \forall a' \in \mu'(f(w), n) \exists a \in \mu(w, n) R'_a(f(w)) = f[R_a(w)]$$

In both items,  $R'_{a'}(f(w)) = f[R_a(w)]$  can be equivalently split into the two usual there-and-back conditions:  $wR_a v$  implies  $f(w)R'_{a'} f(v)$ , and  $f(w)R'_a w'$  implies  $\exists v \in W (wR_a v \wedge f(v) = w')$ . Frame validity is, as expected, preserved under frame morphisms. If each world and its image satisfy the same atomic propositions, we obtain an invariance result for the language of epistemic logic with names.  $F, w \models \varphi$  is defined as  $\forall \pi F, w, \pi \models \varphi$ .

**Lemma 1.** *Assume that  $f : F \rightarrow F'$  is a frame morphism, and valuations  $\pi, \pi'$  are given so that  $\pi(w) = \pi'(f(w))$  for each  $w \in W$ . Then for each formula  $\varphi \in \mathcal{L}_{\mathcal{N}}$ ,*

$$F, \pi, w \models \varphi \text{ if and only if } F', \pi', f(w) \models \varphi.$$

*From this it follows that  $F, w \models \varphi$  implies  $F', f(w) \models \varphi$ .*

The usual model theoretic constructs can be grasped using frame morphisms: that  $F'$  is a generated subframe of  $F$  is defined via an injective frame morphism  $f : F' \hookrightarrow F$ , a morphic image frame via a surjective frame morphism  $f : F \twoheadrightarrow F'$ , and we can see that inclusions in a disjoint union of frames  $f_i : F_i \hookrightarrow \biguplus_{i \in I} F_i$  are frame morphisms. This provides us with the usual validity preservation results. Regarding frame definability, we can already show that the language behaves differently from standard

modal logics: for example, the set of formulas  $\{S_n p \rightarrow p \mid n \in N\}$  defines the class of frames satisfying the condition  $\forall x \forall n (\forall a \in \mu(x, n). xR_a x)$ . The condition requires reflexive loops to exist on each state  $x$  for all relations indexed by the agents named in  $x$ . The language *cannot*, however, define reflexivity of specific individual relations. The proof is short but illustrative, so we state it explicitly.

**Observation 1.** *The class of frames satisfying  $\forall x. xR_a x$  for some fixed  $a \in A$  is not definable in the language  $\mathcal{L}_{\mathcal{N}}$ .*

*Proof.* Consider frame  $F_1$  to consist of a single state  $xR_a x$  with  $\mu(x, n) = \{a\}$ , and frame  $F_2$  to consist of a single state  $x'R_b x'$  with  $\mu'(x', n) = \{b\}$ . Now observe that putting  $f(x) = x'$  we obtain a frame morphism from a  $R_a$ -reflexive frame to a  $R_a$ -non-reflexive frame.  $\square$

Similar definability/undefinability examples can be constructed for many well-known properties and modal axioms, but we do not go into the general definability characterization here, as it turns out we can place the frames in the context of monotone neighborhood structures and address the phenomenon there.

When extended with valuations, frame morphisms become morphisms between models. As a simple application of invariance under morphisms between models, one can see that the modality "someone knows that" with truth condition  $w \models S\varphi \equiv \exists a \in \alpha(w) \forall v \in R_a(w) v \models \varphi$  is not expressible in epistemic logic with names, as morphisms cannot distinguish worlds without agents from worlds without named agents<sup>3</sup>. A straightforward argument shows that extending the language  $\mathcal{L}_{\mathcal{N}}$  with this modality would allow to differentiate frames with and without an existence function mentioned above. The individual knowledge modalities  $K_a$  are not expressible in  $\mathcal{L}_{\mathcal{N}}$  either, as morphisms cannot distinguish between different agents with the same name.

Morphisms between models, via their graph, implicitly encompass the notion of bisimulation between models. This latter notion is worthwhile defining explicitly, as this will open the door to a re-interpretation of epistemic logic with names in neighborhood semantics. This definition turns out to be essentially of the same shape as in [22], but taking into account the naming function.

**Definition 6.** *A bisimulation between a model  $M = (W, A, \mathcal{R}, \mu, \pi)$  and a model  $M' = (W', A', \mathcal{R}', \mu', \pi')$  is a binary relation  $B \subseteq W \times W'$  satisfying the following conditions:  $wBw'$  implies, for each  $n$ ,*

- (0)  $\pi(w) = \pi'(w')$
- (1)  $\forall a \in \mu(w, n) \exists a' \in \mu'(w', n) (\forall u' \in R'_{a'}(w') \exists u \in R_a(w) uBu') \wedge (\forall u \in R_a(w) \exists u' \in R'_{a'}(w') uBu')$
- (2)  $\forall a' \in \mu'(w', n) \exists a \in \mu(w, n) (\forall u \in R_a(w) \exists u' \in R'_{a'}(w') uBu') \wedge (\forall u' \in R'_{a'}(w') \exists u \in R_a(w) uBu')$

*If  $B$  is a bisimulation as above and  $wBw'$ , we call  $(W, A, \mathcal{R}, \pi, w)$  and  $(W', A', \mathcal{R}', \pi', w')$  bisimilar.*

As expected, bisimilarity implies modal equivalence for the language of epistemic logic with names, and the converse holds for image-finite models. These are models where for every state  $w \in W$  and  $n \in N$ , both  $\mu(w, n)$  and  $R_a(w)$  are finite.

**Lemma 2.** *Assume  $B$  is a bisimulation between a model  $M$  and a model  $M'$ , and that  $wBw'$ . Then for each formula  $\varphi \in \mathcal{L}_{\mathcal{N}}$ ,*

$$M, w \models \varphi \text{ if and only if } M', w' \models \varphi.$$

*Furthermore, if both  $M$  and  $M'$  are image-finite, then modal equivalence implies bisimilarity.*

*Proof.* The first part is standard, and proceeds by induction on the complexity of the formula  $\varphi$ . For the second part, the proof follows the argument in [22].  $\square$

<sup>3</sup>The modality  $S$  is close to a modality considered in [22], only the one used in [22] quantifies globally over all agents in  $A$ , not only those that exist in a particular state. It is not expressible in epistemic logic with names either.

Graphs of frame morphisms are prominent examples of bisimulations, as shown in the following Lemma. The proof is omitted for the proceedings version of the paper.

**Lemma 3.** *Assume that  $f : F \rightarrow F'$  is a frame morphism, and valuations  $\pi, \pi'$  are given so that  $\pi(w) = \pi'(f(w))$  for each  $w \in W$ . Then the graph relation  $G(f) = \{(w, f(w)) \mid w \in W\}$  is a bisimulation between a model  $F = (W, A, \mathcal{R}, \mu, \pi)$  and a model  $F' = (W', A', \mathcal{R}', \mu', \pi')$ . Moreover, functional bisimulations are graphs of frame morphisms.*

The semantics of the two modalities, the definitions of frame morphism and bisimulation above, and also the condition corresponding to the  $T(S_n)$  axiom, bear striking resemblance to the model theory of monotone neighborhood models, if we see the sets  $\{R_a(w) \mid a \in \mu(w, n)\}$  as so-called core neighborhood sets [11, 21]. In particular, the notions of core bounded morphism and core monotone bisimulation for so-called core-complete monotone neighborhood models (cf. again [11, 21] for the definition) are relevant here. As we will show in the following section, we can indeed view frames and models in an equivalent way as certain neighborhood structures. This should not come as a surprise, since  $S_n$  are monotone non-normal modalities of the  $\exists\forall$  type. Under closer scrutiny, frame morphisms are core bounded morphisms in disguise, while our definition of bisimulation is stronger than that of core monotone bisimulation, in that it requires that  $B$  is full between  $R_a(w)$  and  $R'_a(w')$  in both (1) and (2). This is because we have an additional  $\forall\forall$  type of modality (namely  $E_n$ ) in the language while the general theory of monotone neighborhood structures only considers  $\exists\forall$  modalities. Interestingly, and in contrast to core monotone bisimulations, functional bisimulations in our sense correspond to graphs of frame morphisms (and thus core bounded morphisms).

### 3 Neighborhood semantics

Beyond the similarity in their underlying notions of model-theoretic invariance, there are also two conceptual reasons to study epistemic logic with names in neighborhood semantics. First, it allows to study the modalities  $E_n$  and  $S_n$  as collective epistemic attitudes in their own right, attitudes that can be instantiated by a number of different assignments of agents to names or groups. Indeed, even though they are present in the semantic structures, the language of this logic makes no direct references to individuals. Different assignments of agents to a name or group  $n$ , i.e. different values of  $\mu(w, n)$ , can yield equivalent sets of statements regarding what some or all agents in  $n$  know at  $w$ . Abstracting from the concrete frames and models defined in the previous sections and moving to neighborhood semantics allows us to study some of the variability in group membership/naming that this allows.

The second conceptual reason to study epistemic logic with names in neighborhood semantics is that it helps assessing the importance of using epistemic, i.e. partitioned/S5 models in the semantics, as done in [10]. We already observed that in the general case neither positive nor negative introspection are valid for  $E_n$  or  $S_n$ , even when the underlying individual relations are assumed to be transitive and Euclidean. These assumptions being controversial anyway [31, 27], this raises the question of whether the logic  $AX_{\mathcal{N}}$  is sound and complete with respect to a larger class of frames. The completeness result provided in [22] provides evidence that this is the case, but a precise argument for this was still missing.

**Definition 7.** *A neighborhood frame  $\mathfrak{F}$  for a given index set  $I$  is a tuple  $(W, \{v_i\}_{i \in I})$  where  $W$  is a set of states and for each  $i \in I$ ,  $v_i : W \rightarrow \mathcal{P}\mathcal{P}(W)$  is a neighborhood function that assigns to each state  $w$  a set of sets of states. We call  $v_i(w)$  the  $i$ -neighborhood of  $w$ . Whenever for all  $w$  and  $X \in v_i(w)$ , we have that  $w \in X$ , we say that  $v_i$  is reflexive.*

*A neighborhood model is a neighborhood frame together with a valuation function  $\pi$ . When  $I$  is a set  $N$  of names we call  $\mathfrak{F}$  neighborhood frame for  $N$ .*

Note that neighborhood frames for  $N$  do not explicitly contain agents. To avoid confusion, in this section we will sometimes refer to the frames defined in Definition 2 as *Kripke frames*, in contrast to the neighborhood frames that we have just defined. The two epistemic modalities  $S_n$  and  $E_n$  are interpreted using the "inexact" semantic clause, which builds in monotonicity.

**Definition 8.** (*satisfaction in neighborhood models*)

$$\begin{aligned} w \models S_n \varphi &\equiv \exists X \in \nu_n(w) X \subseteq \|\varphi\| \\ w \models E_n \varphi &\equiv \forall X \in \nu_n(w) X \subseteq \|\varphi\| \end{aligned}$$

As a first step, we provide a simple representation of reflexive neighborhood frames into frames as defined above. Crucially these frames are not necessarily epistemic ones.

**Observation 2.** *Let  $N$  be a set of names and  $\mathcal{F} = (W, A, \mathcal{R}, \mu)$  be a frame for  $N$ . There is a reflexive neighborhood frame  $\mathfrak{F}^{\mathcal{F}}$  for  $N$  such that for all  $\pi$  and  $\varphi \in \mathcal{L}_N$ ,  $\mathcal{F}, \pi, w \models \varphi$  iff  $\mathfrak{F}^{\mathcal{F}}, \pi, w \models \varphi$ . Conversely, if  $\mathfrak{F} = (W, \{\nu_i\}_{i \in N})$  is a reflexive neighborhood frame for  $N$  then there is a frame  $\mathcal{F}^{\mathfrak{F}}$  such that  $\mathfrak{F}, \pi, w \models \varphi$  iff  $\mathcal{F}^{\mathfrak{F}}, \pi, w \models \varphi$ .*

*Proof.* Going from Kripke frames to neighborhood frames is straightforward, putting for all  $w$ ,  $\nu_n(w) = \{R_a(w) : a \in \mu(w, n)\}$ . For the other direction, we put for all  $w$ ,  $\mu(w, n) = \nu_n(w)$ ,  $A = \bigcup_{n \in N} \bigcup_{w \in W} \mu(w, n)$ , and for all  $a$ ,  $R_a(w) = a$ . The reader might want to compare with the construction in the completeness proof of [10], and also with the construction of ultrafilter frames in subsection 3.3, where the agents that are needed to witness the truth of formulas  $S_n \varphi$  in at a state  $w$  are constructed by identifying them with truth sets of specific formulas, namely those also containing  $\varphi$  together with all formulas  $\psi$  for which  $E_n \psi$  is true at  $w$ .  $\square$

Given a reflexive neighborhood frame  $\mathfrak{F}$ , the Kripke frame  $\mathcal{F}^{\mathfrak{F}}$  as constructed above is reflexive but neither necessarily transitive nor symmetric. Furthermore, the frame  $\mathcal{F}^{\mathfrak{F}}$ , although modally equivalent, is not necessarily isomorphic to  $\mathcal{F}$  in the sense of the first order meta-theory. However, the identity map  $\iota : \mathcal{F} \rightarrow \mathcal{F}^{\mathfrak{F}}$  is a frame morphism, so it is isomorphic in the sense of frame morphisms. We can in fact present the equivalence between the two kinds of semantics as an equivalence of the corresponding categories.

### 3.1 Categorical Equivalence between Frames and Neighborhood frames

We consider the category ( $C$ ) of frames and frame morphisms as defined in Definition 5. On the other hand we consider the category ( $C'$ ) of neighborhood frames of Definition 7, and their morphisms, i.e. maps  $f : W \rightarrow W'$  satisfying:

$$\begin{aligned} \text{(there-n)} \quad X \in \nu_n(w) \text{ implies } f[X] \in \nu'_n(f(w)), \\ \text{(back-n)} \quad Y \in \nu'_n(f(w)) \text{ implies } \exists X (f[X] = Y \ \& \ X \in \nu_n(w)). \end{aligned}$$

The definition is that of core bounded morphisms from [11, Definition 4.6]. The two constructions used in the proof of Observation 2 constitute in fact functors between the two categories, we only need to say what happens to maps  $f : W \rightarrow W'$ : as the underlying sets  $W, W'$  remain unchanged, we can literally take the same map going both ways (from ( $C$ ) to ( $C'$ ) or back). It remains to see that (i) if  $f$  is a frame morphism  $f : \mathcal{F} \rightarrow \mathcal{F}'$ , then it is also a (core bounded) morphism  $f : \mathfrak{F}^{\mathcal{F}} \rightarrow \mathfrak{F}'^{\mathcal{F}'}$ , and (ii) vice versa - if  $f$  is a (core bounded) morphism  $f : \mathfrak{F} \rightarrow \mathfrak{F}'$ , then it is also a frame morphism  $f : \mathcal{F}^{\mathfrak{F}} \rightarrow \mathcal{F}'^{\mathfrak{F}'}$ .

For (i), assume  $f$  is a frame morphism  $f : \mathcal{F} \rightarrow \mathcal{F}'$ . For (there-n) assume that  $X \in \nu_n(w) = \{R_a(w) \mid a \in \mu(w, n)\}$ . So, there is some  $a \in \mu(w, n)$  for which  $f[X] = f[R_a(w)]$ , and therefore there

is some  $a' \in \mu'(f(w), n)$  with  $f[R_a(w)] = R_{a'}(f(w))$  by (there). This shows that  $f[X] = f[R_a(w) = R_{a'}(f(w))] \in \nu'_n(f(w))$ . For (back-n) assume  $Y \in \nu'_n(f(w))$ , so  $Y = R_{a'}(f(w))$  for some  $a' \in \mu'(f(w), n)$ . Then there is  $a \in \mu(w, n)$  with  $f[R_a(w)] = R_{a'}(f(w))$  by (back), so  $R_a(w) \in \nu_n(w)$  is the required set.

For (ii), assume  $f$  is a (core bounded) morphism  $f : \mathfrak{F} \rightarrow \mathfrak{F}'$ . For (there), assume  $a \in \mu(w, n) = \nu_n(w)$  is given. By (there-n),  $f[a] \in \nu'_n(f(w)) = \mu'(f(w), n)$ , so this is our  $a'$ . Now  $R_a(w) = a$  and  $f[R_a(w)] = f(a) = a' = R_{a'}(f(w))$  as required. For (back), assume  $a' \in \mu'(f(w), n) = \nu'_n(f(w))$  is given. By (back-n) there is some  $a \in \nu_n(w) = \mu(w, n)$  with  $f[a] = a'$ . The rest is similar as above.

It remains to observe that the maps  $\iota_{\mathcal{F}} : \mathcal{F} \rightarrow \mathcal{F}^{\mathfrak{F}}$  and  $\iota_{\mathfrak{F}} : \mathfrak{F} \rightarrow \mathfrak{F}^{\mathcal{F}}$ , given as identity on  $W$ , are morphisms (iso-morphisms in fact) in the respective categories.

Observe that nothing in the above (morphisms, their properties, or the translations between the two kinds of frames) depends on reflexivity of frames. In other words, we can establish an equivalence both for frames and neighborhood frames, and for reflexive frames and neighborhood reflexive frames.

### 3.2 Completeness for reflexive neighborhood frames

With this equivalence in hand we can proceed to show that  $AX_{\mathcal{N}}$  is indeed complete with respect to the class of reflexive neighborhood frames. Together with the representation result this gives us that this logic is actually sound and complete with respect to the class of reflexive Kripke frames augmented with a naming function.

**Theorem 1.** *The logic  $AX_{\mathcal{N}}$  is sound and complete w.r.t. the class of reflexive neighborhood frames.*

*Proof.* The proof proceeds by a standard canonical model construction for neighborhood semantics.  $\square$

Besides completeness for the class of reflexive Kripke frames that comes as a corollary of the completeness and representation results for neighborhood frames, this connection allows us to relate to existing results in the model and the proof theory of non-normal modal logics. We are not aware of an algebraic duality, Goldblatt-Thomason definability theorem, or van Benthem theorem existing in the literature which would apply to our case as it is<sup>4</sup>, so we present it in more details in the next section.

On the proof theoretical side, there has been to our knowledge no study of sequent calculi dedicated to epistemic logic with names specifically. [15], however, provides a sound and complete nested sequent system that has cut elimination for Brown's "logic of ability" [3]. This bi-modal logic receives the same interpretation in neighborhood semantics as epistemic logic with names, with the  $S_n$  and the  $E_n$  modalities corresponding to Brown's "can" and "will" operators, respectively. The only difference is that Brown's neighborhood frames are not necessarily reflexive, and so the corresponding axiomatization omits  $T(S_n)$ . We conjecture that the system in [15] can be extended with the standard sequent rule for T, c.f. [23], without breaking the cut elimination result, but we leave the development of the details for future work.

### 3.3 Algebraic duality

Any of the categories of frames described in 3.1 (we leave out reflexivity for this part) can be seen as dual to the category of the following modal algebras with names. We pick the neighborhood frames of

<sup>4</sup>Recall that in contrast to general model theory of monotone neighborhood logics we have additional  $\forall\forall$  type modalities in the language, which in particular affects the definition of standard translation or ultrafilter extensions of frames and models.

Definition 7 to show this is so. A modal algebra with names over  $N$  is  $\mathbb{A} = (A, \wedge, \neg, \{E_n, S_n \mid n \in N\})$ , a Boolean algebra with modalities satisfying the following equations:

$$\begin{aligned} E_n \top &= \top & \neg E_n \perp &= S_n \top \\ E_n(a \wedge b) &= (E_n a \wedge E_n b) & S_n a \wedge E_n b &\leq S_n(a \wedge b). \end{aligned}$$

It is not hard to see that this presentation is equivalent to the one obtained by simply algebraizing  $AX_{\mathcal{N}}$  (without the  $T(S_n)$  axiom). Homomorphisms of modal algebras with names are Boolean homomorphisms preserving the  $S_n, E_n$  modalities.

**Ultrafilter frames:** Given a modal algebras with names  $\mathbb{A} = (A, \wedge, \neg, \{E_n, S_n \mid n \in N\})$  we construct its ultrafilter frame over the set of ultrafilters on  $\mathbb{A}$  as  $\mathfrak{F}^{\mathbb{A}} = (\mathcal{U}(\mathbb{A}), \mathbf{v}_n^{\mathbb{A}})$  where for  $u \in \mathcal{U}(\mathbb{A})$

$$\mathbf{v}_n^{\mathbb{A}}(u) := \{\hat{a} \cap \bigcap_{E_n d \in u} \hat{d} \mid S_n a \in u\}.$$

The sets  $\hat{a} = \{u \in \mathcal{U}(\mathbb{A}) \mid a \in u\}$  constitute the (clopen) basis of a topology on  $\mathcal{U}(\mathbb{A})$ .

For a homomorphism  $h : \mathbb{A} \rightarrow \mathbb{B}$ , the inverse-image map  $h^{-1}[\ ] : \mathcal{U}(\mathbb{B}) \rightarrow \mathcal{U}(\mathbb{A})$  is a (bounded core) morphism from  $\mathfrak{F}^{\mathbb{B}}$  to  $\mathfrak{F}^{\mathbb{A}}$ . This is not immediate to see, and we will hint at the interesting part, namely that it satisfies the (there-n) condition: Assume  $Y \subseteq \mathcal{U}(\mathbb{B})$  with  $Y \in \mathbf{v}_n^{\mathbb{B}}(u)$ . It means that for some  $b \in B$ ,  $Y = \{\hat{b} \cap \bigcap_{E_n c \in u} \hat{c} \mid S_n b \in u\}$ . We want to show that for some  $a \in A$  with  $S_n h(a) \in u$ ,

$$\{h^{-1}[v] \mid v \in \{\hat{b} \cap \bigcap_{E_n c \in u} \hat{c}\}\} = \{\hat{a} \cap \bigcap_{E_n h(d) \in u} \hat{d}\} \in \mathbf{v}_n^{\mathbb{A}}(h^{-1}[u]).$$

By  $S_n b \in u$  we know there is at least one such candidate  $a$ : we can consider  $h^{-1}(b)$  if  $b \in \text{Rng}(h)$ , or  $\top$  otherwise. Now, in both cases,

$$v \in \{\hat{b} \cap \bigcap_{E_n c \in u} \hat{c}\} \equiv \{b\} \cup \{c \mid E_n c \in u\} \subseteq v \equiv h^{-1}[\{b\} \cup \{c \mid E_n c \in u\}] \subseteq h^{-1}[v] \equiv h^{-1}[v] \in \{\hat{a} \cap \bigcap_{E_n h(d) \in u} \hat{d}\}.$$

For the last equivalence, observe that  $h^{-1}(b) = a$  in the first case and that  $\hat{a} = \hat{\top} = \mathcal{U}(\mathbb{A})$  in the second case, and  $h^{-1}[\{c \mid E_n c \in u\}] = \{d \mid E_n h(d) \in u\}$ .

**Complex algebras:** Given a neighborhood frame  $\mathfrak{F} = (W, \mathbf{v}_n)$ , we construct its complex algebra as  $\mathbb{A}^{\mathfrak{F}} = (\mathcal{P}W, \cap, -, \{E_n, S_n \mid n \in N\})$  where

$$\begin{aligned} E_n(X) &:= \{w \mid \forall Y \in \mathbf{v}_n(w) Y \subseteq X\}, \\ S_n(X) &:= \{w \mid \exists Y \in \mathbf{v}_n(w) Y \subseteq X\}. \end{aligned}$$

For a (bounded core) morphism  $f : \mathfrak{F} \rightarrow \mathfrak{G}$ , the inverse-image map  $f^{-1}[\ ] : \mathbb{A}^{\mathfrak{G}} \rightarrow \mathbb{A}^{\mathfrak{F}}$  is a homomorphism of modal algebras with names.

The map  $\hat{\ } : \mathbb{A} \rightarrow \mathbb{A}^{\mathfrak{F}^{\mathbb{A}}}$  assigning  $a \mapsto \hat{a}$  is an embedding (this underlies the completeness proof for the logic if taken without the  $T(S_n)$  axiom). The frame  $\mathfrak{F}^{\mathbb{A}^{\mathfrak{F}}} = ue(\mathfrak{F})$  is the ultrafilter extension of  $\mathfrak{F}$ . With a little extra work one can prove that taking ultrafilter extension reflects frame validity:  $ue(\mathfrak{F}) \models \varphi$  implies  $\mathfrak{F} \models \varphi$ . Clearly, the inverse-image maps injective morphisms to surjective ones and vice versa. Also for  $\mathfrak{F} = \biguplus_{i \in I} \mathfrak{F}_i$ , we can show that  $\mathbb{A}^{\mathfrak{F}} \simeq \prod_{i \in I} \mathbb{A}^{\mathfrak{F}_i}$ . Putting all this to work, one can show literally by the standard argument based on the duality and Birkhoff's theorem (cf. [11, Theorem 7.23]), the following definability theorem:

**Theorem 2.** *Let  $K$  be a class of neighborhood frames over  $N$  which is closed under taking ultrafilter extensions. Then  $K$  is definable in the language  $\mathcal{L}_{\mathcal{N}}$  iff  $K$  is closed under disjoint unions, generated subframes and bounded morphic images, and reflects ultrafilter extensions.*

## 4 Common and distributed knowledge with names

In this section we extend the language of epistemic logic with names with the two most well-known collective epistemic modalities: common and distributed knowledge. These notions are mentioned but explicitly set aside in [10]. [19] and after that [8, pp.213-218] study what they call non-rigid common and distributed knowledge with essentially the same goal as us: formulating a meaningful version of these notions in contexts where there can be uncertainty about who is in the group. As already observed in [10], non-rigid common and distributed knowledge turn out to be related but subtly different from the way we define these two group attitudes. We review these differences below. Common and distributed knowledge have also been studied in term modal logic [29, 20], but in languages with different expressive power than epistemic logic with names.

### 4.1 Common knowledge with names

In the context where group membership may vary from state to state, an appropriate notion of common knowledge for group  $n$  must take into account not only what the agents that are members of  $n$  in the current state consider possible (and what they consider others might consider possible, and so on...), but also who these group members consider might be in the group. Here we extend epistemic logic with names with a common knowledge operator in a way that is as close as possible to the standard definition of common knowledge in multi-agent epistemic logic. After presenting the semantics and returning to our running example we present a sound and complete axiomatization. We discuss briefly at the end of the section the differences between our notion and non-rigid common knowledge defined in [19, 8].

We write  $C_n$  for common knowledge among members of the group *named*  $n$ , and  $\mathcal{L}_{\mathcal{N}\mathcal{E}}$  for the extension of language  $\mathcal{L}_{\mathcal{N}}$  with operator  $C_n$ . Similarly as for the standard common knowledge operators, we can unfold  $C_n$  in terms of  $E_n$ :

$$C_n\varphi := \bigwedge_{k \in \mathbb{N}^*} E_n^k \varphi.$$

where  $E_n^1\varphi := E_n\varphi$  and  $E_n^{k+1}\varphi := E_n E_n^k \varphi$ . The underlying relation  $R_n$  is constructed as follows:  $R_n = \{(w, v) \mid \text{for some } i \in \mu(w, n), (w, v) \in R_i\}$ . We write  $R_{C_n}$  for  $R_n^+$ , the transitive closure of  $R_n$ . The semantics of the operator is given by the following expected clause:

$$M, w \models C_n\varphi \text{ iff for all } v \text{ such that } (w, v) \in R_{C_n}, M, v \models \varphi.$$

As an example, consider again the model depicted in Figure 1. In state  $w$ , it is common knowledge among Trump supporters that he was impeached or lost the election,  $w \models C_n(p \vee q)$ . Note, however, that this is not common knowledge in the standard sense between  $a$  and  $b$ , even though the set of Trump supporters is exactly  $\{a, b\}$  in  $w$ . Conversely, in state  $v$ , it is not common knowledge among trolls that Trump won the election ( $v \not\models C_m \neg q$ ) even though the set of trolls is just  $\{b\}$  and it is common knowledge in set  $\{b\}$  in the standard sense. Despite these differences, semantically our new operator still corresponds to a transitive closure operation, and is therefore captured at the axiomatic level in a way similar to the standard notion, as our axiomatization below reflects.

**Definition 9** (System  $AX_{\mathcal{N}\mathcal{E}}$ ). *The logic  $AX_{\mathcal{N}\mathcal{E}}$  is the logic  $AX_{\mathcal{N}}$  extended with the following axioms and rules for  $C_n$ :*

$$\begin{array}{ll} K(C_n) & C_n(\varphi \rightarrow \psi) \rightarrow (C_n\varphi \rightarrow C_n\psi) \\ FP & C_n\varphi \rightarrow E_n(\varphi \wedge C_n\varphi) \\ Ind & \text{From } \varphi \rightarrow E_n(\varphi \wedge \psi), \text{ infer } \varphi \rightarrow C_n\psi \\ Nec(C_n) & \text{From } \varphi, \text{ infer } C_n\varphi \end{array}$$

**Theorem 3.**  $AX_{\mathcal{N}\mathcal{E}}$  is sound and complete with respect to the class of models.

*Proof.* Soundness is straightforward. To prove completeness, we adapt the usual method of building a canonical model with maximal consistent extensions of the finite closure of a formula to circumvent non-compactness (see for instance [5, Section 7.3]), and combine it with the canonical model construction introduced in [10]. The combination is tedious, in part due to the non-factive nature of operators  $C_n$  (and  $E_n$ ) but works as expected. The heart of the proof is case 6. of Lemma 6 below. We start by defining an appropriate notion of closure.

**Definition 10** (Closure  $cl(\chi)$ ). Let  $\chi \in \mathcal{L}_{\mathcal{N}\mathcal{E}}$  and  $Sub(\varphi)$  be the set of subformulas of  $\varphi$ . The closure  $cl(\chi)$  of  $\chi$  is the smallest set such that:

1.  $\chi \in cl(\chi)$
2. if  $\varphi \in cl(\chi)$ , then  $Sub(\varphi) \in cl(\chi)$ ,
3. if  $\varphi \in cl(\chi)$  and  $\varphi$  is not of the form  $\neg\psi$ , then  $\neg\varphi \in cl(\chi)$ ,
4.  $S_n(p \vee \neg p) \in cl(\chi)$ ,
5.  $E_n(p \wedge \neg p) \in cl(\chi)$ ,
6. if  $E_n\varphi \in cl(\chi)$ , then  $S_n\varphi \in cl(\chi)$ ,
7. if  $C_n\varphi \in cl(\chi)$ , then  $E_n\varphi \in cl(\chi)$ ,
8. if  $C_n\varphi \in cl(\chi)$ , then  $E_nC_n\varphi \in cl(\chi)$ .

**Lemma 4.** For all  $\chi \in \mathcal{L}_{\mathcal{N}\mathcal{E}}$ ,  $cl(\chi)$  is finite.

*Proof.* Standard. □

**Definition 11** (Maximal consistent sets in  $cl(\chi)$ ). A set  $\Gamma$  is maximal consistent in  $cl(\chi)$  when:

1.  $\Gamma \subseteq cl(\chi)$ ,
2.  $\Gamma \not\vdash \perp$ ,
3. there is no  $\Delta$ , such that  $\Gamma \subset \Delta$ ,  $\Delta \subseteq cl(\chi)$ , and  $\Delta \not\vdash \perp$ .

**Lemma 5** (Lindenbaum). For every  $\Gamma \subseteq cl(\chi)$ , if  $\Gamma$  is consistent, then there is a set  $\Delta$ , such that  $\Gamma \subseteq \Delta$  and  $\Delta$  is maximal consistent in  $cl(\chi)$ .

*Proof.* The proof goes via the standard method of enumeration of all formulas in  $cl(\chi)$  and sequential construction. □

**Definition 12** ( $n$ -path,  $n$ -path into  $\varphi$ ). Given a model  $M = (W, A, \mathcal{R}, \pi, \mu)$ , a  $n$ -path from  $w$  is a sequence  $\langle w_0, \dots, w_k \rangle$  with  $w_0, \dots, w_k \in W$  and  $k \in \mathbb{N}^*$ , such that  $w_0 = w$  and for all  $0 \leq m < k$ , there is an agent  $i \in \mu(w_m, n)$  such that  $w_m R_i w_{m+1}$ . A  $n$ -path into  $\varphi$  is a  $n$ -path such that for all  $1 \leq m \leq k$ ,  $\varphi \in w_m$ .

We build the canonical model in a similar manner as [10].

**Definition 13** (Canonical model  $M_{cl(\chi)}$ ). Given a formula  $\chi \in \mathcal{L}_{\mathcal{N}\mathcal{E}}$ , the canonical model for the closure of  $\chi$  is  $M_{cl(\chi)} = (W, A, \mathcal{R}, \pi, \mu)$  where:

- $W$  is the set of all maximal consistent sets within  $cl(\chi)$ ,
- $A$  is the set of agents, constructed as follows:

- for every  $w \in W$ , and every formula  $S_n\phi \in w$ , define  $D_{\phi,w,n} = \{\phi\} \cup \{\psi \mid E_n\psi \in w\}$ ,
- for every such set  $D_{\phi,w,n}$ , the set  $a_{\phi,w,n} = \{v \in W \mid D_{\phi,w,n} \subseteq v\}$  is an agent.
- for all  $w, v \in W$ , and all  $a \in A$ ,  $(w, v) \in R_a$  iff  $w \in a$  and  $v \in a$ ,
- for all  $p \in \Phi$ ,  $\pi(p) = \{w \mid p \in w\}$ ,
- $\mu$  is such that  $\mu(w, n) = \{a_{\phi,w,n} \mid S_n\phi \in w\}$ .<sup>5</sup>

**Lemma 6.** Let  $M_{cl(\chi)} = (W, A, \mathcal{R}, \pi, \mu)$  be the canonical model for  $cl(\chi)$ . We write  $\underline{w}$  for the (finite) conjunction  $\bigwedge_{\phi \in w} \phi$ . For all  $w, v \in W$ :

1. if  $\phi \in cl(\chi)$  and  $\underline{w} \vdash \phi$ , then  $\phi \in w$  (i.e.  $w$  is deductively closed within  $cl(\chi)$ )
2. if  $\neg\phi \in cl(\chi)$ , then  $\neg\phi \in w$  iff  $\phi \notin w$ ,
3. if  $(\phi \wedge \psi) \in cl(\chi)$ , then  $(\phi \wedge \psi) \in w$  iff  $\phi \in w$  and  $\psi \in w$ ,
4. if  $S_n\phi \in cl(\chi)$ , then  $S_n\phi \in w$  iff for all  $v \in W$  such that  $wR_{a_{\phi,w,n}}v$ ,  $\phi \in v$ ,
5. if  $E_n\phi \in cl(\chi)$ , then  $E_n\phi \in w$  iff for all  $\psi$  with  $S_n\psi \in w$ , and all  $v \in W$  such that  $wR_{a_{\psi,w,n}}v$ ,  $\phi \in v$ ,
6. if  $C_n\phi \in cl(\chi)$ , then  $C_n\phi \in w$  iff every  $n$ -path from  $w$  is a path into  $\phi$  and into  $C_n\phi$ .

*Proof.* Other cases are proved in the standard way; we give below the full detail of the crucial case 6. To prove 6., assume  $C_n\phi \in cl(\chi)$ .

For the left-right direction, let  $C_n\phi \in w$ . We show that for all  $n$ -paths  $\langle w, w_1, \dots, w_k \rangle$ , and all  $k \in \mathbb{N}^*$ ,  $\phi \in w_k$  and  $C_n\phi \in w_k$ , by induction on  $k$ .

Base case ( $k = 1$ ): Let  $v$  be such that  $wR_nv$ . By definition of  $R_n$ , there is some agent  $a \in \mu(w, n)$  such that  $wR_av$ . By definition of  $\mu(w, n)$ ,  $a$  is  $a_{\psi,w,n}$  for some  $\psi$  for which  $S_n\psi \in w$ . Since  $C_n\phi \in w$  and  $\vdash C_n\phi \rightarrow E_n(\phi \wedge C_n\phi)$  (FP), by *MP*, *K*( $E_n$ ) and *PL* we obtain  $\underline{w} \vdash E_n\phi$  and  $\underline{w} \vdash E_nC_n\phi$ . Since both  $E_n\phi, E_nC_n\phi \in cl(\chi)$ , we know that  $E_n\phi \in w$  and  $E_nC_n\phi \in w$ .

By definition of  $D_{\psi,w,n}$ , since  $E_n\phi \in w$  and  $E_nC_n\phi \in w$ , then both  $\phi \in D_{\psi,w,n}$  and  $C_n\phi \in D_{\psi,w,n}$ . By definition of  $R_{a_{\psi,w,n}}$ , since  $wR_{a_{\psi,w,n}}v$ , then  $v \in a_{\psi,w,n}$ . By definition of  $a_{\psi,w,n}$ , for all  $u \in a_{\psi,w,n}$ ,  $D_{\psi,w,n} \subseteq u$ . Therefore, in particular,  $D_{\psi,w,n} \subseteq v$ , and since  $\phi \in D_{\psi,w,n}$ , and  $C_n\phi \in D_{\psi,w,n}$ , it follows that  $\phi \in v$  and  $C_n\phi \in v$  as desired.

Step case: ( $k = m + 1$ ). If there is a  $n$ -path from  $w \langle w_0, \dots, w_{m+1} \rangle$ , then there is a  $n$ -path from  $w$  to  $w_m$  and  $w_mR_nw_{m+1}$ . By the induction hypothesis,  $\phi \in w_m$  and  $C_n\phi \in w_m$ . To show that  $\phi \in w_{m+1}$  and  $C_n\phi \in w_{m+1}$ , it suffices to repeat the arguments used for the base case.

For the right-left direction, assume that for all  $n$ -paths  $\langle w, w_1, \dots, w_k \rangle$  and for all  $k \geq 1$ ,  $\phi \in w_k$  and  $C_n\phi \in w_k$ . We show that  $C_n\phi \in w$ . We will write  $W_{n,\phi}$  for the set of states from which all  $n$ -paths are into  $\phi$ . Let  $\omega := \bigvee_{u \in W_{n,\phi}} u$ . By our assumption,  $w \in W_{n,\phi}$  which means that  $\underline{w}$  is a disjunct of  $\omega$ . Therefore,  $\vdash \underline{w} \rightarrow \omega$ . If we show that  $\vdash \omega \rightarrow E_n(\phi \wedge \omega)$ , then applying the induction rule we can infer  $\vdash \omega \rightarrow C_n\phi$ . Since we already have  $\vdash \underline{w} \rightarrow \omega$ , we obtain  $\vdash \underline{w} \rightarrow C_n\phi$ , and therefore, since  $C_n\phi \in cl(\chi)$ , also  $C_n\phi \in w$ .

It remains to show that  $\vdash \omega \rightarrow E_n(\phi \wedge \omega)$ . We will demonstrate separately that  $\vdash \omega \rightarrow E_n\phi$ , and that  $\vdash \omega \rightarrow E_n\omega$ .

To show that  $\vdash \omega \rightarrow E_n\phi$ , suppose for contradiction it is not the case. Then  $\omega \wedge \neg E_n\phi$  is consistent. This implies that there is some disjunct  $u$  of  $\omega$  such that  $u \wedge \neg E_n\phi$  is consistent. Since  $u$  is maximal (and  $\neg E_n\phi \in cl(\chi)$ ),  $\neg E_n\phi \in u$ . There are two cases to consider:

<sup>5</sup>As discussed in [10], this construction guarantees at the same time that all sentences of the form  $S_n\phi$  have a witness (an agent named  $n$  who knows  $\phi$ ), and that this witness also knows whatever sentence everybody with name  $n$  should know.

Case i): Assume  $\mu(u, n) \neq \emptyset$  ( $n$  is not an empty name in  $u$ ). Let  $a$  be an arbitrary member of  $\mu(u, n)$ . By definition of  $R_a$ , we have  $uR_a u$ . By assumption that all  $n$ -paths from  $u$  are into  $C_n\varphi$ , we know that  $C_n\varphi \in u$ . Since  $\vdash C_n\varphi \rightarrow E_n\varphi$  and  $E_n\varphi \in cl(\chi)$ , we know that  $E_n\varphi \in u$ . But then, since both  $E_n\varphi \in u$  and  $\neg E_n\varphi \in u$ ,  $u$  is not consistent. A contradiction.

Case ii) Assume  $\mu(u, n) = \emptyset$  ( $n$  is an empty name in  $u$ ). By definition of  $\mu(u, n)$ , this implies that there is no  $\psi$  such that  $S_n\psi \in u$ . In particular,  $S_n(p \vee \neg p) \notin u$ . Therefore,  $\neg S_n(p \vee \neg p) \in u$ . By axiom *Int*<sub>2</sub>, we obtain  $\vdash \neg S_n(p \vee \neg p) \rightarrow E_n(p \wedge \neg p)$ , and  $E_n(p \wedge \neg p) \in u$ . And since  $\vdash E_n(p \wedge \neg p) \rightarrow E_n(\varphi \wedge \neg \varphi)$  and  $\vdash E_n(\varphi \wedge \neg \varphi) \rightarrow E_n\varphi$ , we have  $E_n\varphi \in u$ . But then, again, both  $E_n\varphi \in u$  and  $\neg E_n\varphi \in u$ , and so  $u$  is not consistent. A contradiction. This concludes the proof that  $\vdash \omega \rightarrow E_n\varphi$ .

We now turn to the proof of  $\vdash \omega \rightarrow E_n\omega$ . Suppose for contradiction it is not the case. Then  $\omega \wedge \neg E_n\omega$  is consistent, and there must be some disjunct  $u_0$  of  $\omega$  such that  $u_0 \wedge \neg E_n\omega$  is consistent. We need to construct  $v \in W$  (a MCS in  $cl(\chi)$ ) such that  $\forall u \in W_{n,\varphi} \exists \psi_u \in u$  with  $\psi_u \notin v$ , and we need to find an agent  $a \in \mu(u_0, n)$  with  $u_0R_a v$ .

Observe first that  $S_n\top \in u_0$ : if not, then  $\neg E_n\perp \notin u_0$  and therefore  $E_n\perp \in u_0$ , and consequently  $u_0 \vdash E_n\omega$  which is a contradiction. Therefore we take  $a_{\top, u_0, n}$  to be the chosen agent in  $\mu(u_0, n)$ . Next, observe that  $D_{\top, u_0, n} \subseteq u_0$ : this follows by  $S_n\top \wedge E_n\gamma \vdash S_n\gamma \vdash \gamma$  for each  $E_n\gamma \in u_0$ . This means we need  $D_{\top, u_0, n} \subseteq v$  to ensure that  $u_0R_a v$ . To sum up,  $v$  needs to include all formulas in  $D_{\top, u_0, n}$ , and for each  $u \in W_{n,\varphi}$  it has to pick a formula  $\psi_u \in u$  to exclude (i.e.  $\neg\psi_u \in v$ ). We need to prove that at least one such choice is consistent: assume for a contradiction that it is not so. This means that for each choice  $C_i = \{\psi_u \mid u \in W_{n,\varphi}\}$  ( $i$  ranges  $1 \dots \prod_{u \in W_{n,\varphi}} |u|$ ), we have that  $D_{\top, u_0, n} \vdash \bigvee C_i$ . Therefore  $D_{\top, u_0, n} \vdash \bigwedge_i \bigvee C_i$  and, by the distributive law,  $D_{\top, u_0, n} \vdash \omega$ . But this entails that  $u_0 \vdash E_n\omega$ , contradicting our original assumption. So, one such choice is possible to make, and by the Lindenbaum lemma we can construct  $v \in W$  as required.

We can deduce that  $v \notin W_{n,\varphi}$  (because for any  $v \in W_{n,\varphi}$ ,  $\vdash v \rightarrow \omega$ ). Since  $v \notin W_{n,\varphi}$ , there is an  $n$ -path from  $v$  into some state  $s$  such that  $\varphi \notin s$ . But any  $n$ -path from  $v$  is an  $n$ -path from  $w$  too, and so there is an  $n$ -path from  $w$  which is not into  $\varphi$ . This contradicts our initial assumption, and concludes the proof that  $\vdash \omega \rightarrow E_n\omega$ .

Therefore, since we have  $\vdash \omega \rightarrow E_n\varphi$  and  $\vdash \omega \rightarrow E_n\omega$ , we obtain  $\vdash \omega \rightarrow E_n(\varphi \wedge \omega)$  (by *PL*, *Nec*( $E_n$ ), and *K*( $E_n$ )). Finally, by inference rule *Ind*,  $\vdash \omega \rightarrow C_n\varphi$  and since  $\vdash \underline{w} \rightarrow \omega$ , we obtain  $\vdash \underline{w} \rightarrow C_n\varphi$ , which in turns implies that  $C_n\varphi \in w$ .  $\square$

**Lemma 7** (Truth Lemma). *Let  $\psi \in \mathcal{L}_{\mathcal{N}\mathcal{E}}$ , and  $M_{cl(\psi)} = (W, A, \mathcal{R}, \pi, \mu)$  canonical for  $cl(\psi)$ . For all  $\varphi \in cl(\psi)$ :*

$$M_{cl(\psi)}, w \models \varphi \text{ if and only if } \varphi \in w.$$

*Proof.* By induction on the length of  $\varphi$ . Boolean cases are standard, and cases for  $E_n\psi$  and  $S_n\psi$  are treated similarly as in [10], we only give the detail for the additional case where  $\varphi := C_n\psi$ .

$$\begin{aligned} M, w \models C_n\psi & \text{ iff (by the semantics of } C_n) \\ \text{for all } v \in R_{C_n}(w), M, v \models \psi & \text{ iff (by IH)} \\ \text{for all } v \in R_{C_n}(w), \psi \in v & \text{ iff (by def of } n\text{-paths)} \\ \text{all } n\text{-paths from } w \text{ are into } \psi & \text{ iff (by Lemma 6)} \\ C_n\psi \in w. & \square \end{aligned}$$

A standard contraposition argument allows to conclude the proof of Theorem 3.  $\square$

We finish this section by comparing briefly the notion of common knowledge studied here with non-rigid common knowledge as defined in [19, 8]. The models used to define non-rigid common knowledge

are essentially the same as the models of epistemic logic with names. The definition of non-rigid common knowledge is, however, based on a different individual modality, written  $B_i^n$ , which is informally described as "agent  $i$  knows/believes that if  $i$  is in  $n$ , then  $\varphi$ ". The original motivation for this reading was to study cases where a process knows that if it is correct, then  $\varphi$  holds. Translated into our semantics this would have the following truth condition.

$$M, w \models B_i^n \varphi \text{ iff } \forall w'. wR_a w' \wedge a \in \mu(w', n) \rightarrow M, w' \models \varphi.$$

An easy argument shows that  $S_n \varphi$  implies  $B_i^n \varphi$  for some agents  $i$  that are members of  $n$  in the current situation, but not the other way around. [10] show that the corresponding  $S'_n$  and  $E'_n$  can be axiomatized using the same axioms as the original  $S_n$  and  $E_n$ . On the basis of this we conjecture that the corresponding notion of common knowledge could be axiomatized as the one defined here, but we leave this for future work.

## 4.2 Distributed knowledge with names

For distributed knowledge we propose a generalization of the  $S_n$  modality. Instead of requiring the existence of one *agent* with name  $n$  that knows  $\varphi$ ,  $D_n \varphi$  is true whenever there exists a (non-empty) *sub-group* of  $n$  such that pooling the information of the agents in that sub-group would entail  $\varphi$ .

**Definition 14** (Satisfaction,  $D_n$ ).

$$M, w \models D_n \varphi \text{ iff } \exists X \subseteq \mu(w, n), X \neq \emptyset \wedge \forall v \in \bigcap_{i \in X} R_i(w), M, v \models \varphi$$

This version of distributed knowledge, unlike  $S_n$ , turns out to be closed under conjunction. The intuitive reason for this is that sub-groups can be merged. If a sub-group  $g$  of  $n$  distributively knows that  $p$  and another sub-group  $g'$  of  $n$  distributively knows that  $q$ , when the union  $g \cup g'$  of these sub-groups effectively pools these two pieces of information, leading to distributed knowledge of  $p \wedge q$ . The logic of  $D_n$  is not completely normal, though, as like for  $S_n$  it does not validate necessitation.

**Definition 15** (System  $AX_{\mathcal{N}\mathcal{D}}$ ). *The logic  $AX_{\mathcal{N}\mathcal{D}}$  is the logic  $AX_{\mathcal{N}}$ , extended with the following axioms:*

$$\begin{array}{ll} D_n \varphi \wedge D_n(\varphi \rightarrow \psi) \rightarrow D_n \psi & K_D \\ S_n \varphi \rightarrow D_n \varphi & \text{Inclusion} \\ D_n \varphi \rightarrow \varphi & T_D \\ D_n \varphi \wedge E_n(\varphi \rightarrow \psi) \rightarrow D_n \psi & \text{Interaction} \end{array}$$

**Theorem 4.** *The logic  $AX_{\mathcal{N}\mathcal{D}}$  is sound and complete w.r.t. the class of frames over  $N$  and the language containing the modality  $D_n$ .*

*Proof.* The proof proceeds by adapting the usual copy-and-splitting technique for completeness with intersection modalities (c.f. [30] for a recent overview).  $\square$

Distributed knowledge with names is, like its standard counterpart, not invariant under the notion of bisimulation in Definition 6. It can however be extended in the standard way to also cover intersections of modalities.

## 5 Conclusion

Adding names to standard epistemic logic lifts one of the most fundamental idealizations, alongside logical omniscience, of classical logic for knowledge and belief: the facts that, on the one hand, the agents' names are common knowledge and, on the other, that groups are defined extensionally. This idealization is perhaps even more problematic than logical omniscience to the extent that it appears unrealistic even if we give a normative interpretation to multi-agents epistemic logic. It seems perfectly possible for perfect reasoners to still be uncertain of the identity of the others, or of who belong to which group.

The results reported here update Grove and Halpern's original formulation, so to speak, by connecting it to known model-theoretic results and to neighborhood semantics, as well as by extending it with the standard notions of common and distributed knowledge. This lays the ground for studying, using modern logical tools, knowledge and beliefs in a much broader class of social situations than what was possible in standard epistemic logic, without having to move to highly complex, explicit first-order extensions of epistemic logic.

The results here suggest many possible avenues for future work. We have mentioned along the way a number of open questions related to model and proof theory, both for the original  $E_n$  and  $S_n$  modalities as well as for the common and distributed knowledge extensions. The joint axiomatization of the latter two notions is also open, as well as the extendability of the results here to more restricted classes of frames, in particular in those cases where the agents know their own names. Another natural next step is to look at dynamic extensions of this language [14, 17], and especially conditions under which agents can learn who are the members of certain groups. One should also study the possibility of postulating certain patterns of interactions between names. In the current framework, there is no such interaction, and one cannot even express the fact that a certain group  $n$  is a sub-group of another group  $m$ . A first question to address here would be what extensions of the language allow for the expressions of such relations while keeping the logic decidable.

## References

- [1] W Lance Bennett, Alexandra Segerberg & Shawn Walker (2014): *Organization in the crowd: peer production in large-scale networked protests*. *Information, Communication & Society* 17(2), pp. 232–260, doi:10.1080/1369118X.2013.870379.
- [2] Cristina Bicchieri (2005): *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, doi:10.1017/CBO9780511616037.
- [3] M.A. Brown (1988): *On the logic of ability*. *Journal Phil. Log.* 17(1), pp. 1–26, doi:10.1007/BF00249673.
- [4] Gabriella Coleman (2014): *Hacker, hoaxer, whistleblower, spy: The many faces of Anonymous*. Verso books.
- [5] Hans van Ditmarsch, Wiebe van der Hoek & Barteld Kooi (2007): *Dynamic epistemic logic*. 337, Springer Science & Business Media, doi:10.1007/978-1-4020-5839-4.
- [6] Barbara Dunin-Keplicz & Rineke Verbrugge (2011): *Teamwork in multi-agent systems: A formal approach*. 21, John Wiley & Sons, doi:10.1002/9780470665237.
- [7] Ronald Fagin & Joseph Y Halpern (1987): *Belief, awareness, and limited reasoning*. *Artificial intelligence* 34(1), pp. 39–76, doi:10.1016/0004-3702(87)90003-8.
- [8] Ronald Fagin, Joseph Y. Halpern, Yoram Moses & Moshe Vardi (2004): *Reasoning about Knowledge*. MIT press, doi:10.7551/mitpress/5803.001.0001.
- [9] Adam J Grove (1995): *Naming and identity in epistemic logic Part II: a first-order logic for naming*. *Artificial Intelligence* 74(2), pp. 311–350, doi:10.1016/0004-3702(95)98593-D.

- [10] Adam J. Grove & Joseph Y. Halpern (1993): *Naming and Identity in Epistemic Logics Part I: The Propositional Case*. *Journal of Logic and Computation* 3(4), pp. 345–378, doi:10.1093/logcom/3.4.345.
- [11] Helle Hvid Hansen (2003): *Monotonic modal logics*. Master’s thesis, ILLC UVA.
- [12] Helle Hvid Hansen, Clemens Kupke & Eric Pacuit (2007): *Bisimulation for neighbourhood structures*. In: *International Conference on Algebra and Coalgebra in Computer Science*, Springer, pp. 279–293, doi:10.1007/978-3-540-73859-6\_19.
- [13] Jaakko Hintikka (1962): *Knowledge and belief: An introduction to the logic of the two notions*. Cornell University Press.
- [14] Barteld Kooi (2007): *Dynamic term-modal logic*. In: *A Meeting of the Minds*, pp. 173–186.
- [15] Björn Lellmann (2019): *Combining Monotone and Normal Modal Logic in Nested Sequents – with Countermodels*. In Serenella Cerrito & Andrei Popescu, editors: *Automated Reasoning with Analytic Tableaux and Related Methods*, Springer, Cham, pp. 203–220, doi:10.1007/978-3-030-29026-9\_12.
- [16] David Lewis (2008): *Convention: A philosophical study*. John Wiley & Sons.
- [17] Andrés Occhipinti Liberman, Andreas Achen & Rasmus Kræmmer Rendsvig (2020): *Dynamic term-modal logics for first-order epistemic planning*. *Artificial Intelligence* 286, p. 103305, doi:10.1016/j.artint.2020.103305.
- [18] Ruth Barcan Marcus (1961): *Modalities and intensional languages*. *Synthese*, pp. 303–322, doi:10.1007/BF00486629.
- [19] Yoram Moses & Mark R Tuttle (1988): *Programming simultaneous actions using common knowledge*. *Algorithmica* 3(1), pp. 121–169, doi:10.1007/BF01762112.
- [20] Pavel Naumov & Jia Tao (2018): *Everyone knows that someone knows: quantifiers over epistemic agents*. *The review of symbolic logic*, doi:10.1017/S1755020318000497.
- [21] Eric Pacuit (2017): *Neighborhood semantics for modal logic*. Springer, doi:10.1007/978-3-319-67149-9.
- [22] Anantha Padmanabha & R Ramanujam (2019): *Propositional modal logic with implicit modal quantification*. In: *Indian Conference on Logic and Its Applications*, Springer, pp. 6–17, doi:10.1007/978-3-662-58771-3\_2.
- [23] Francesca Poggiolesi (2008): *A cut-free simple sequent calculus for modal logic S5*. *The Review of Symbolic Logic* 1(1), pp. 3–15, doi:10.1017/S1755020308080040.
- [24] Olivier Roy & Anne Schwenkenbecher (2019): *Shared intentions, loose groups, and pooled knowledge*. *Synthese*, pp. 1–19, doi:10.1007/s11229-019-02355-x.
- [25] David P. Schweikard & Hans Bernhard Schmid (2020): *Collective Intentionality*. In Edward N. Zalta, editor: *The Stanford Encyclopedia of Philosophy*, winter 2020 edition, Metaphysics Research Lab, Stanford University.
- [26] Gennady Shtakser (2018): *Propositional epistemic logics with quantification over agents of knowledge*. *Studia Logica* 106(2), pp. 311–344, doi:10.1007/s11225-017-9741-0.
- [27] Robert Stalnaker (2006): *On logics of knowledge and belief*. *Philosophical studies* 128(1), pp. 169–199, doi:10.1007/s11098-005-4062-y.
- [28] Georg H Von Wright (1954): *An essay in modal logic*. North Holland.
- [29] Yanjing Wang & Jeremy Seligman (2018): *When names are not commonly known: epistemic logic with assignments*. In Guram Bezhanishvili, Giovanna D’Agostino, George Metcalfe & Thomas Studer, editors: *Proceedings of AiML 2018*.
- [30] Yi N Wáng & Thomas Ågotnes (2020): *Simpler completeness proofs for modal logics with intersection*. In: *International Workshop on Dynamic Logic*, Springer, pp. 259–276, doi:10.1007/978-3-030-65840-3\_16.
- [31] Timothy Williamson (2002): *Knowledge and its Limits*. Oxford University Press, doi:10.1093/019925656X.001.0001.

# Language-based Decisions

Adam Bjorndahl

Department of Philosophy  
Carnegie Mellon University  
Pittsburgh, USA  
abjorn@cmu.edu

Joseph Y. Halpern

Department of Computer Science  
Cornell University  
Ithaca, USA  
halpern@cs.cornell.edu

In Savage’s classic decision-theoretic framework [12], actions are formally defined as functions from states to outcomes. But where do the state space and outcome space come from? Expanding on recent work by Blume, Easley, and Halpern [3], we consider a language-based framework in which actions are identified with (conditional) descriptions in a simple underlying language, while states and outcomes (along with probabilities and utilities) are constructed as part of a representation theorem. Our work expands the role of language from that in [3] by using it not only for the *conditions* that determine which actions are taken, but also the *effects*. More precisely, we take the set of actions to be built from those of the form  $do(\varphi)$ , for formulas  $\varphi$  in the underlying language. This presents a problem: how do we interpret the result of  $do(\varphi)$  when  $\varphi$  is underspecified (i.e., compatible with multiple states)? We answer this using tools familiar from the semantics of counterfactuals [13]: roughly speaking,  $do(\varphi)$  maps each state to the “closest”  $\varphi$ -state. This notion of “closest” is also something we construct as part of the representation theorem; in effect, then, we prove that (under appropriate assumptions) the agent is acting *as if* each underspecified action is first made definite and then evaluated (i.e., by maximizing expected utility). Of course, actions in the real world are often not presented in a fully precise manner, yet agents reason about and form preferences among them all the same. Our work brings the abstract tools of decision theory into closer contact with such real-world scenarios.

## 1 Motivation

In Savage’s classic decision-theoretic framework [12] *actions* are formally defined as functions from *states* to *outcomes*. States are conceptualized as encoding the possible uncertainty the decision-maker may have about the world, while outcomes correspond intuitively to the payoff-relevant ways things might turn out. Thus, an action  $\alpha$  can be viewed as a kind of long list: for each way the world might be (i.e., each state  $s$ ),  $\alpha$  specifies what will happen—namely, the outcome  $\alpha(s)$ —in case action  $\alpha$  is actually performed in state  $s$ .

One might ask: where do the state space and outcome space come from? Is it reasonable to model an agent using a mathematical apparatus they presumably have no access to? Questions like these tap into a long tradition of challenging the idealizations involved in models like Savage’s (see, e.g., [1, 2, 4, 5, 6, 7, 8, 9, 10, 14]). One response might be that we are not trying to *duplicate* the decision-making process going on “in the agent’s head”, but rather to *represent* it, mathematically—to show that under certain conditions it can be tracked with a certain type of formalism (in this case, as a form of expected utility maximization).

Although this reply might assuage some worries about the use of abstract mathematical frameworks for reasoning about decision making in general, it remains problematic that actions—the objects over which agents are supposed to “reveal” their preferences, through concrete, binary choices—cannot themselves be described except by reference to the background state and outcome spaces, which might not

be the states and outcomes that the agent is actually thinking of. In such models, although outcomes are what agents are supposed to ultimately care about, actions are the *means* by which they bring outcomes about. This makes an agent’s preferences regarding actions arguably the closest point of contact that these models have to the empirical, observable reality of choosing between alternatives. Indeed, this interpretation of actions is what underlies many of the intuitions brought to bear to justify the various axioms of decision making that Savage postulates and relies upon to prove his celebrated representation theorem.

The concern with where the states and outcomes are coming from motivated Blume, Easley, and Halpern [3] (henceforth BEH) to consider a model where acts and language are taken to be primary in a sense that we explain shortly, while the state and outcome space are constructed as part of the representation rather than specified exogenously. In more detail, BEH assumed that acts were programs in a simple programming language formed by closing off a set of primitive programs using **if . . . then . . . else . . .**, so that if  $a$  and  $b$  are programs and  $t$  is a test (intuitively, a formula in a propositional language), then **if  $t$  then  $a$  else  $b$**  is a program. Thus, rather than conditioning actions on events (i.e., subsets of a state space), they are conditioned on *descriptions* of events, namely, tests. This approach allows BEH to not only circumvent a fixed, exogenous specification of the state space and outcome space (instead, they are constructed as part of a representation theorem, and programs are identified with maps from from these states to outcomes), but also (as they illustrate with several examples) makes it possible to capture a variety of *framing* effects, which basically derive from a mismatch between how the modeler conceives of the world and how the agent does, as manifested in different ways that descriptions of events might map onto actual events.

Our work is perhaps best understood as an extension of their work in which the role of language is even more central. Specifically, while BEH allowed arbitrary primitive programs, we take the primitive programs to have the form  $do(\varphi)$ , where  $\varphi$  is a formula. The  $do(\varphi)$  notation follows Pearl [11]; intuitively,  $do(\varphi)$  means that the agent somehow makes  $\varphi$  true. Note that this action is somewhat underspecified; it does not say what else becomes true as a result of  $\varphi$  being true; for example, if  $\psi$  is independent of  $\varphi$ , it does not tell us whether  $\psi$  or  $\neg\psi$  is true. In our representation theorem, we assume that the agent has a way of specifying the effects of  $do(\varphi)$ . In more detail, we take states in our state space to be characterized by formulas in the language (this is similar to the canonical model used in BEH’s representation theorem), and take the outcome space to be the same as the state space, so that a program maps states to states. As part of the representation theorem, the agent must decide what state  $do(\varphi)$  maps each state  $\omega$  to. We follow standard approaches to giving semantics to counterfactuals [13] by taking  $do(\varphi)$  to map  $\omega$  to the “closest” state to  $\omega$  (according to some measure of closeness) where  $\varphi$  is true. Of course, what counts as “closest” depends on the agent’s subjective view of the world, and is constructed from their preferences over acts.

This approach allows us to model choices in a way that seems to us closer to how agents perceive and reason about the options available to them. To illustrate, consider a policy-maker trying to decide whether to raise the minimum wage to \$15 or to leave it as is. In our framework, this amounts to comparing the acts  $do(MW = \$15)$  and  $do(true)$  (where  $do(true)$  amounts to doing nothing). Of course, different agents may disagree about the side-effects of increasing the minimum wage (businesses may close, there may be more automation so jobs may be lost, and so on). This amounts to saying that different agents will interpret  $do(MW = \$15)$  differently as a function from states to states, although all will agree that it will result in a state where the minimum wage is \$15.<sup>1</sup> We can also express contingent policies in our

---

<sup>1</sup>We remark that in this paper we consider only the single-agent case, but we find the multi-agent case, and specifically the effect of disagreements about what the closest state is, an exciting direction for future work.

framework, for example, raising the minimum wage if the economy is healthy.

By making both the acts and the test conditions formulas, we can capture framing and coarseness effects not only in the test conditions, but also in the choices. For example, we might imagine agents reacting differently to statements like “we will require that every citizen is paid at least \$15 dollars for each hour they work” versus “we will require every business owner to pay their employees at least \$15 for each hour they work”, even if we can see that these are equivalent statements. Our framework would allow this.

The rest of this paper is organized as follows. We present our approach as an extension of the work of BEH. This has the benefit of allowing us to apply their representation theorem directly and focus our efforts on the novel aspects of our extension. We begin in Section 2 by reviewing the relevant definitions from BEH and augmenting them with the new ones we need to capture language-based, underspecified effects of actions. Then in Section 3 we articulate the representation theorem we are aiming at, introduce decision-theoretic axioms that allow us to achieve it—including axioms from BEH (Section 3.1) as well as several new axioms (Section 3.2)—and finally prove the theorem (Section 3.3). Section 4 concludes with a discussion of future work. Appendix A collects proofs omitted from the main text.

## 2 Language, Actions, and Models

Our first step is to import the relevant definitions from BEH so as to present our extension of their work in context. In order to emphasize the changes that we make and to streamline the presentation, we alter some of their notation and terminology, and focus on the special case of their system without randomization.

Let  $\Phi$  denote a finite set of *primitive propositions*, and  $\mathcal{L} = \mathcal{L}(\Phi)$  the propositional language consisting of all Boolean combinations of these primitives. Although of course it is possible (and interesting) to consider other languages, in this work we focus on languages of this form as the *underlying language of action*—intuitively, the language in which both the conditions and the results of actions are specified.

A **basic model (over  $\mathcal{L}(\Phi)$ )** is a tuple  $M = (\Omega, \llbracket \cdot \rrbracket_M)$  where  $\Omega$  is a nonempty set of *states* and  $\llbracket \cdot \rrbracket_M : \Phi \rightarrow 2^\Omega$  is a *valuation function*. The valuation is recursively extended to all formulas in  $\mathcal{L}$  in the usual way. Intuitively,  $\llbracket \varphi \rrbracket_M$  is the set of states where  $\varphi$  is true. Using  $\llbracket \cdot \rrbracket_M$  allows us to interpret descriptions in the language  $\mathcal{L}$  (what BEH call “tests”) as events:  $\varphi$  is interpreted as the subset  $\llbracket \varphi \rrbracket_M \subseteq \Omega$  of the state space  $\Omega$ . We sometimes drop the subscript when the model is clear from context, and write  $\omega \models \varphi$  for  $\omega \in \llbracket \varphi \rrbracket$ . We say that  $\varphi$  is *satisfiable in  $M$*  if  $\llbracket \varphi \rrbracket_M \neq \emptyset$  and that  $\varphi$  is *valid in  $M$*  if  $\llbracket \varphi \rrbracket_M = \Omega$ , and write  $\models \varphi$  to indicate that  $\varphi$  is valid in all basic models. Finally, we define the *theory of  $\omega$  (in  $M$ )* to be the set of all formulas true at  $\omega$ , denoted  $Th(\omega) = \{\varphi : \omega \models \varphi\}$ , and write  $\omega \equiv \omega'$  iff  $Th(\omega) = Th(\omega')$ .

Up to now, everything we have defined has followed BEH exactly—their “primitive tests” are our primitive propositions  $\Phi$ ; their “tests” are our formulas  $\mathcal{L}(\Phi)$ ; their “test interpretations” are our valuations  $\llbracket \cdot \rrbracket_M$ . Next we define our version of their “primitive choices”. This is where our development begins to diverge, since we take these to be actions of the form  $do(\varphi)$ ; in other words, we specify primitive choices using the same underlying language  $\mathcal{L}(\Phi)$  that corresponds to tests, rather than treating them as a brand new set of primitives.

Formally, given a finite set of formulas  $F \subseteq \mathcal{L}$ , the set of **actions (over  $F$ )**, denoted by  $\mathcal{A}_F$ , is defined recursively as follows: for each  $\varphi \in F$ ,  $do(\varphi)$  is an action (called a *primitive action*), and for all  $\psi \in \mathcal{L}$  and  $\alpha, \beta \in \mathcal{A}_F$ , **if  $\psi$  then  $\alpha$  else  $\beta$**  is an action. Following BEH, we take  $F$  to be finite (who take the set of primitive choices to be finite). It is also convenient because it allows us to exclude logical inconsistencies from  $F$ , obviating the need to interpret actions like  $do(false)$ . For the propositional languages under

consideration in this paper, up to logical equivalence, there are only finitely many formulas in any case.

Naturally, we also wish to *interpret* our actions in a way that respects their connection to the underlying language. This is the topic we turn to next.

## 2.1 Selection models

In a given basic model  $M$ , we want  $do(\varphi)$  to correspond to a function whose range is contained in  $\llbracket \varphi \rrbracket_M$ , the set of  $\varphi$ -states. Thus, we restrict our attention to basic models in which each  $\varphi \in F$  is satisfiable—in this case we say that  $M$  is  **$F$ -rich**. But this is not enough: as discussed,  $do(\varphi)$  is underspecified; it does not in general determine a unique function. In order to interpret such actions and compare them to others, we must in some sense “fill in” the missing details. We formalize this with the concept of a **selection model (for  $F$ )**, which is a basic model  $M = (\Omega, \llbracket \cdot \rrbracket_M)$  together with a *selection function (for  $M$ )*  $c : \Omega \times F \rightarrow \Omega$  satisfying  $c(\omega, \varphi) \in \llbracket \varphi \rrbracket_M$ .

Selection functions were introduced by Stalnaker [13] as a mechanism to interpret counterfactual conditionals. Following this tradition, we think of  $c(\omega, \varphi)$  as representing the “closest” state to  $\omega$  where  $\varphi$  is true. There are many other properties one might insist  $c$  have, aside from  $c(\omega, \varphi) \in \llbracket \varphi \rrbracket$  (which is called **success**). For example, one may require that if  $\omega \in \llbracket \varphi \rrbracket$ , then  $c(\omega, \varphi) = \omega$  (i.e., if  $\varphi$  is true in  $\omega$ , then the closest state to  $\omega$  where  $\varphi$  is true is  $\omega$  itself); this property is called **centering**.

In this paper we will also consider a relatively strong condition on  $c$ , namely, that it is derived from a parametrized family of *well-orders*<sup>2</sup> on the state space, one for each state:  $\leq := \{\leq_\omega : \omega \in \Omega\}$ . Intuitively,  $\omega_1 \leq_\omega \omega_2$  says “ $\omega_1$  is at least as close to  $\omega$  as  $\omega_2$  is”. We say that a selection function  $c$  is **induced by  $\leq$**  if  $c(\omega, \varphi)$  always outputs the  $\leq_\omega$ -minimal element of  $\llbracket \varphi \rrbracket$ . We call  $\leq$  **centered** if, for each  $\omega \in \Omega$ , the  $\leq_\omega$ -minimal element of  $\Omega$  is  $\omega$  (in which case it is also easy to see that the induced selection function satisfies centering). Finally, we say that  $\leq$  is **language-based** if the relations  $\leq_\omega$  on the quotient  $\Omega/\equiv$  given by

$$[\omega_1] \leq_\omega [\omega_2] \text{ iff } \omega_1 \leq_\omega \omega_2$$

are well-defined well-orders, and moreover, whenever  $\omega \equiv \omega'$ , we have  $\leq_\omega = \leq_{\omega'}$ . Note that in this case  $\omega \equiv \omega'$  implies  $c(\omega, \varphi) \equiv c(\omega', \varphi)$ .<sup>3</sup> Intuitively, if  $\leq$  is language-based then what counts as the closest state essentially depends only on the formulas that are true at a state. We cannot have two states  $\omega_1$  and  $\omega_2$  that agree on all formulas (so that  $\omega_1 \equiv \omega_2$ ) and a third state  $\omega_3$  that does not agree with  $\omega_1$  and  $\omega_2$  on all formulas such that  $\omega_3$  is between  $\omega_1$  and  $\omega_2$  in terms of distance from some state  $\omega$  (i.e., we cannot have  $\omega_1 \leq_\omega \omega_3 \leq_\omega \omega_2$ ).

The purpose of the selection function in our models is to take an underspecified transition from states to states and “resolve the ambiguity”. Specifically, given a transition that starts in state  $\omega$  and ends up in a  $\varphi$ -state, the selection function  $c$  can then be applied to specify the exact  $\varphi$ -state, namely  $c(\omega, \varphi)$ , where it actually ends up. In this way, given a basic,  $F$ -rich model  $M$ , each action of the form  $do(\varphi)$  can be interpreted in any selection model  $(M, c)$  based on  $M$  as a function  $\llbracket do(\varphi) \rrbracket_{M,c} : \Omega \rightarrow \Omega$  defined by:

$$\llbracket do(\varphi) \rrbracket_{M,c}(\omega) = c(\omega, \varphi).$$

<sup>2</sup>A binary relation  $\leq$  on a set is called a *linear order* if it is complete, transitive, and antisymmetric (i.e.,  $x \leq y$  and  $y \leq x$  implies  $x = y$ ). A *well-order* is a linear order in which every nonempty subset has a least element.

<sup>3</sup>Here’s why: since  $c(\omega, \varphi)$  is the  $\leq_\omega$ -minimal element of  $\llbracket \varphi \rrbracket$ , it must also be that  $[c(\omega, \varphi)]$  is the  $\leq_\omega$ -minimal element of  $\{[\omega''] : \omega'' \models \varphi\}$ . Similarly,  $[c(\omega', \varphi)]$  is the  $\leq_{\omega'}$ -minimal element of  $\{[\omega''] : \omega'' \models \varphi\}$ . Since  $\leq_\omega = \leq_{\omega'}$ , these must coincide, so we have  $[c(\omega, \varphi)] = [c(\omega', \varphi)]$ .

Of course, we can extend this interpretation to all actions in  $\mathcal{A}_F$  in the obvious way (and exactly as BEH do):

$$\llbracket \text{if } \psi \text{ then } \alpha \text{ else } \beta \rrbracket_{M,c}(\omega) = \begin{cases} \llbracket \alpha \rrbracket_{M,c}(\omega) & \text{if } \omega \in \llbracket \psi \rrbracket \\ \llbracket \beta \rrbracket_{M,c}(\omega) & \text{if } \omega \notin \llbracket \psi \rrbracket. \end{cases}$$

### 3 Representation

We begin as usual with a binary relation  $\succeq$  on  $\mathcal{A}_F$ , where  $\alpha \succeq \beta$  says that  $\alpha$  is “at least as good as”  $\beta$ . Following standard conventions, we define  $\alpha \succ \beta$  as an abbreviation for  $\alpha \succeq \beta$  and  $\beta \not\succeq \alpha$ , and  $\alpha \sim \beta$  for  $\alpha \succeq \beta$  and  $\beta \succeq \alpha$ , representing “strict preference” and “indifference”, respectively. We also assume that  $\succeq$  is *complete*, that is, all elements are comparable, so that for all acts  $\alpha$  and  $\beta$ , either  $\alpha \succeq \beta$  or  $\beta \succeq \alpha$ . Although BEH consider incomplete relations, we focus here on the simpler case of complete relations in order to streamline the presentation and highlight the novel components of our model.

A **language-based SEU (Subjective Expected Utility) representation** for a relation  $\succeq$  on  $\mathcal{A}_F$  is a finite selection model  $(M, c)$  together with a probability measure  $\pi$  on  $\Omega$  and a *utility function*  $u : \Omega \rightarrow \mathbb{R}$  such that, for all  $\alpha, \beta \in \mathcal{A}_F$ ,

$$\alpha \succeq \beta \Leftrightarrow \sum_{\omega \in \Omega} \pi(\omega) \cdot u(\llbracket \alpha \rrbracket_{M,c}(\omega)) \geq \sum_{\omega \in \Omega} \pi(\omega) \cdot u(\llbracket \beta \rrbracket_{M,c}(\omega)). \quad (1)$$

We note the key differences between the representation theorem BEH establish and what we are aiming at. First, their result produces a separate outcome space and state space, whereas for us, these spaces coincide. More importantly, their result treats “primitive choices” (namely, our actions  $do(\varphi)$ , for  $\varphi \in F$ ) as true primitives in the sense that each is assigned to an *arbitrary* function from states to outcomes. By contrast, we want to respect the structure of an action like  $do(\varphi)$ —specifically, its connection to the formula  $\varphi$ —by requiring that  $do(\varphi)$  correspond to a map from  $\Omega$  to  $\Omega$  such that  $\omega \mapsto c(\omega, \varphi)$  for a suitable selection function  $c$ . One of the novel aspects of our proof consists in showing how to determine the selection function from preferences on acts.

Since our framework can be viewed a specialization of the BEH framework (with our actions having additional, language-based structure as described), rather than proving our representation theorem from scratch, we can reuse much of their construction. Thus, we will present the same axioms (adapted to our notation) that BEH present, and subsequently augment them with new principles that allow us to construct the selection function.

#### 3.1 Cancellation

BEH’s main axiom is a *cancellation law*. Explaining this requires a few preliminary definitions, beginning with the notion of a *multiset*, which can be thought of as a set that allows for multiple instances of each of its elements; two multisets are equal just in case they contain the same elements *with the same multiplicities*. For example, the multiset  $\{\{a, a, a, b, b\}\}$  is different from the multiset  $\{\{a, b, b, b, b\}\}$ : both multisets have five elements, but the multiplicity of  $a$  and  $b$  differ.

Given any subset  $X \subseteq \Phi$ , let  $\varphi_X = \bigwedge_{p \in X} p \wedge \bigwedge_{q \notin X} \neg q$ . Intuitively,  $\varphi_X$  is a “complete description” of the truth values of all primitive propositions in the language  $\mathcal{L}(\Phi)$ , namely the description that says for each primitive proposition  $p$  that it is true iff it belongs to  $X$ . An **atom** is any formula of the form  $\varphi_X$ . Since  $\mathcal{L}(\Phi)$  is a propositional language and we use classical semantics for propositional logic, for all formulas  $\varphi \in \mathcal{L}(\Phi)$  and atoms  $\varphi_X$ , the truth of  $\varphi$  is determined by  $\varphi_X$ : either  $\models \varphi_X \rightarrow \varphi$ , or  $\models \varphi_X \rightarrow \neg \varphi$ .

It is therefore not surprising that every action in  $\alpha \in \mathcal{A}_F$  can be identified with a function  $f_\alpha : 2^\Phi \rightarrow F$ , defined recursively as follows:

$$\begin{aligned} f_{do(\varphi)}(X) &= \varphi \\ f_{\text{if } \psi \text{ then } \alpha \text{ else } \beta}(X) &= \begin{cases} f_\alpha(X) & \text{if } \models \varphi_X \rightarrow \psi \\ f_\beta(X) & \text{if } \models \varphi_X \rightarrow \neg\psi. \end{cases} \end{aligned}$$

BEH define atoms in the same way and use them to define functions from atoms to primitive choices just as we did above (replace  $do(\varphi)$  by an arbitrary primitive choice).<sup>4</sup>

Now we can state the central cancellation law that enables us to apply the BEH representation theorem:

**(Canc)** Let  $\alpha_1, \dots, \alpha_n, \beta_1, \dots, \beta_n \in \mathcal{A}_F$ , and suppose that for each  $X \subseteq \Phi$  we have  $\{f_{\alpha_1}(X), \dots, f_{\alpha_n}(X)\} = \{f_{\beta_1}(X), \dots, f_{\beta_n}(X)\}$ . Then, if for all  $i < n$  we have  $\alpha_i \succeq \beta_i$ , it follows that  $\beta_n \succeq \alpha_n$ .

Intuitively, this says that if we get the same collection of outcomes with  $\alpha_1, \dots, \alpha_n$  as with  $\beta_1, \dots, \beta_n$  (taking multiplicity into account) in each state, then we should view the collection  $\{\alpha_1, \dots, \alpha_n\}$  and  $\{\beta_1, \dots, \beta_n\}$  as equally good. Thus, if  $\alpha_i$  is at least as good as  $\beta_i$  for  $i = 1, \dots, n-1$ , then, to balance things out,  $\beta_n$  should be at least as good as  $\alpha_n$ .

As pointed out by BEH, Cancellation is a surprisingly powerful axiom. In particular, BEH show that we can use **(Canc)** to derive many simpler (and more classical) principles of choice: that  $\succeq$  is reflexive and transitive, that *independence* holds,<sup>5</sup> and that if  $\alpha$  and  $\beta$  are *equivalent* in the sense that  $f_\alpha = f_\beta$ , then  $\alpha \sim \beta$ . (However, it should be noted that Cancellation seems stronger than the conjunction of these axioms.)

### 3.2 Selection axioms

To present the new axioms that will allow us to construct an appropriate selection function as part of the representation theorem, it will be helpful to introduce some new notation. To begin, we write **if**  $\varphi$  **then**  $\alpha$  as a shorthand for **if**  $\varphi$  **then**  $\alpha$  **else**  $do(true)$ . Intuitively, the action  $do(true)$  corresponds to doing “nothing”, since *true* is true no matter what, so we might think of “otherwise nothing” as being the default in case no explicit **else...** clause is given. Of course, for this to make sense we must have  $true \in F$ ; we make this assumption henceforth.

Next we define an abbreviation for *conditional preference*, familiar from Savage’s classical development [12]: write  $\alpha \succeq_\varphi \beta$  as an abbreviation for **(if**  $\varphi$  **then**  $\alpha$ )  $\succeq$  **(if**  $\varphi$  **then**  $\beta$ ).<sup>6</sup> When  $\varphi = \varphi_X$ , we write  $\alpha \succeq_X \beta$  for  $\alpha \succeq_{\varphi_X} \beta$ , and we extend this notation to strict conditional preference and conditional indifference in the obvious way.

Our first axiom is related to the centering constraint for selection functions (i.e., that if  $\varphi$  is true at a state, then that state automatically counts the “closest”  $\varphi$ -state):

<sup>4</sup>Technically, we are not mapping atoms to primitive acts, but since there is an obvious bijection  $X \mapsto \varphi_X$  between sets of primitive proposition and atoms, and an obvious bijection  $\varphi \mapsto do(\varphi)$  between elements of  $F$  and primitive acts, we really can be thought of as doing just that.

<sup>5</sup>That is, for all  $\alpha, \beta, \gamma, \gamma' \in \mathcal{A}_F$  and all  $\varphi \in F$ ,

$$(\text{if } \varphi \text{ then } \alpha \text{ else } \gamma \succeq \text{if } \varphi \text{ then } \beta \text{ else } \gamma) \Leftrightarrow (\text{if } \varphi \text{ then } \alpha \text{ else } \gamma' \succeq \text{if } \varphi \text{ then } \beta \text{ else } \gamma').$$

<sup>6</sup>As BEH show, the cancellation law implies independence, so in fact we have  $\alpha \succeq_\varphi \beta$  iff for all  $\gamma$ , **if**  $\varphi$  **then**  $\alpha$  **else**  $\gamma \succeq$  **if**  $\varphi$  **then**  $\beta$  **else**  $\gamma$ .

**(Cent)** If  $\models \psi \rightarrow \phi$ , then **(if  $\psi$  then  $do(\phi)$ )**  $\sim do(true)$ .

To build intuition it's helpful to consider the special case where  $\psi = \phi$ , in which case **(Cent)** just says that doing  $\phi$  precisely when  $\phi$  is already the case (and otherwise doing nothing) is the same as doing nothing. Here of course by “the same” what is really meant is that the agent is indifferent between those two acts. Since we are trying to bootstrap properties of a selection function from the agent's preferences, all our principles will ultimately need to bottom out in statements about what the agent does or does not have a preference between. The general statement of **(Cent)** simply expands this reasoning to cases where the condition  $\psi$  entails the result of the action,  $\phi$ , and so again in this case  $do(\phi)$  happens only in cases where  $\phi$  is already true.

**Lemma 1.** *If  $(M, c)$  is a selection model,  $c$  satisfies centering, and  $\models \psi \rightarrow \phi$ , then*

$$\llbracket \text{if } \psi \text{ then } do(\phi) \rrbracket_{M,c} = id_{\Omega} = \llbracket do(true) \rrbracket_{M,c}.$$

Our second axiom is meant to capture the idea that *sufficiently specific conditions* resolve any ambiguity (expressible in the underlying language) about the effect of an action:

**(SSC)** If  $\models \phi \leftrightarrow (\phi_1 \vee \dots \vee \phi_n)$ , then  $\forall X \subseteq \Phi$ ,  $\exists i \in \{1, \dots, n\}$  such that for all  $\psi$  satisfying  $\models \phi_i \rightarrow \psi$  and  $\models \psi \rightarrow \phi$ , we have  $do(\psi) \sim_X do(\phi_i)$ .

This requires some unpacking. As above, it is illuminating to begin by considering the special case where  $\psi = \phi$ . Then  $\models \psi \rightarrow \phi$  holds trivially and  $\models \phi_i \rightarrow \psi$  is true by assumption, so we can read **(SSC)** intuitively as follows: If  $\phi$  is ambiguous between a variety of (potentially) more precise statements (namely,  $\phi_1, \dots, \phi_n$ ), then for any sufficiently specific condition (i.e., any atom  $\phi_X$ ), there is at least one precisification  $\phi_i$  of  $\phi$  such that, conditional on  $\phi_X$ , doing  $\phi$  is equivalent to doing  $\phi_i$  (from the agent's perspective).

This, as well as the more general statement of **(SSC)**, follows from the assumption that the selection function  $c$  is induced by a language-based family of well-orders.

**Lemma 2.** *If  $(M, c)$  is a selection model where  $c$  is induced by the well-orders  $\leq = \{\leq_{\omega} : \omega \in \Omega\}$ ,  $\leq$  is language-based,  $\models \phi \leftrightarrow (\phi_1 \vee \dots \vee \phi_n)$ , and  $X \subseteq \Phi$ , then  $\exists i \in \{1, \dots, n\}$  such that for all  $\psi$  satisfying  $\models \phi_i \rightarrow \psi$  and  $\models \psi \rightarrow \phi$  and all  $\omega \in \llbracket \phi_X \rrbracket$ , we have  $\llbracket do(\psi) \rrbracket_{M,c}(\omega) = \llbracket do(\phi_i) \rrbracket_{M,c}(\omega)$ .*

The next idea is crucial to the ultimate construction of our selection function. For each atom  $\phi_W$ , we will define a total preorder<sup>7</sup>  $\sqsubseteq_W$  on the set of atoms that will in turn be extended to a linear order and used to specify the selection function. Formally, we define:

$$\phi_X \sqsubseteq_W \phi_Y \text{ iff } do(\phi_X \vee \phi_Y) \sim_W do(\phi_X).$$

Loosely speaking,  $\phi_X \sqsubseteq_W \phi_Y$  says that in  $\phi_W$ -states, the ambiguity inherent in doing  $\phi_X \vee \phi_Y$  is resolved in the agent's mind in favour of doing  $\phi_X$ ; this is why the agent is indifferent (conditional on  $\phi_W$ ) between doing  $\phi_X \vee \phi_Y$  and just doing  $\phi_X$ . In this sense we think of  $\phi_X$  as being at least as “close” to  $\phi_W$  as  $\phi_Y$  is.

Note that the definition above requires  $F$  to contain all atoms as well as all pairwise disjunctions of atoms. This richness in  $F$  is what allows us to use the agent's preferences on actions to define an appropriate preorder. We make this assumption henceforth. It is an interesting question to what extent the ensuing construction can be carried out without this assumption; we return to this point in Section 4.

Now we can state our third axiom, which simply says that this notion of closeness is transitive:

**(Trans)** For all  $W, X, Y, Z \subseteq \Phi$ , if  $\phi_X \sqsubseteq_W \phi_Y$  and  $\phi_Y \sqsubseteq_W \phi_Z$ , then  $\phi_X \sqsubseteq_W \phi_Z$ .

<sup>7</sup>A total preorder is a complete and transitive relation (so, unlike a linear order, it need not be antisymmetric).

**Lemma 3.** (SSC) implies that each  $\sqsubseteq_W$  is complete.

**Lemma 4.** If (SSC) and (Trans) hold, then each  $\sqsubseteq_W$  is a total preorder and can be extended to a well-order  $\leq_W$  on the set of atoms; if, in addition, (Cent) holds, then each  $\leq_W$  can be defined so that  $\varphi_W$  is the  $\leq_W$ -minimal element.

Given a family of well-orders  $\{\leq_W : W \subseteq \Phi\}$  as defined in Lemma 4, let  $\min_{\leq}(W, \varphi)$  denote the unique  $X \subseteq \Phi$  such that  $\varphi_X$  is  $\leq_W$ -minimal in  $\{\varphi_Y : \models \varphi_Y \rightarrow \varphi\}$ . So  $\varphi_X$  is the “closest” atom compatible with  $\varphi$  to  $\varphi_W$ ; intuitively, then, doing  $\varphi$  in a  $\varphi_W$  situation should essentially amount to doing  $\varphi_X$ . This is precisely what the next lemma asserts.

**Lemma 5.** If (SSC) and (Trans) hold, then  $do(\varphi) \sim_W do(\varphi_{\min_{\leq}(W, \varphi)})$ .

### 3.3 The representation theorem

**Theorem 1.** If  $\succeq$  is a complete binary relation on  $\mathcal{A}_F$  satisfying (Canc), (Cent), (SSC), and (Trans), then there is a language-based SEU representation for  $\succeq$ .

*Proof.* We begin by following the proof in [3, Theorem 2] to obtain a state-dependent representation with state space  $2^\Phi$  and outcome space  $F$ .<sup>8</sup> More precisely, we consider the set of functions  $\mathcal{F} = \{f_\alpha : \alpha \in \mathcal{A}_F\}$  defined in Section 3.1, which can be viewed as Savage acts in the classical sense [12]. The relation  $\succeq$  on  $\mathcal{A}_F$  induces a relation  $\succeq^*$  on  $\mathcal{F}$  defined as follows:

$$f_\alpha \succeq^* f_\beta \Leftrightarrow \alpha \succeq \beta.$$

As discussed, (Canc) implies that  $\alpha \sim \alpha'$  whenever  $f_\alpha = f_{\alpha'}$ , so  $\succeq^*$  is well-defined; moreover, as BEH show, (Canc) is strong enough to yield the desired state-dependent representation result for  $\succeq^*$ , namely, that there exists a function  $u^* : 2^\Phi \times F \rightarrow \mathbb{R}$  such that, for all  $f, g \in \mathcal{F}$ ,

$$f \succeq^* g \Leftrightarrow \sum_{X \in 2^\Phi} u^*(X, f(X)) \geq \sum_{X \in 2^\Phi} u^*(X, g(X)).$$

Up to now we have mirrored the proof given by BEH exactly, which has given us a utility function  $u^*$  but also an outcome space that we don't want. Moreover, the utility function is *state-dependent*; it takes as arguments both a state and an outcome. We want a utility function that depends only on states (which for us are the same as outcomes). Thus, our task now is to transform this result into a selection model that we can use to give a language-based SEU representation of  $\succeq$  (including a utility function defined only on states).

Set  $\Omega = 2^\Phi \times 2^\Phi$ ; so our state space is isomorphic to *pairs* of atoms. This is a technical maneuver that allows us to “factor out” probabilities from the state-dependent utility function  $u^*$  we already have. Loosely speaking, given  $(X, Y) \in \Omega$ , the first component  $X$  represents how things are, while the second component  $Y$  represents how things *were*. This intuition should become clearer as we continue.

We define a basic model  $M = (\Omega, \llbracket \cdot \rrbracket_M)$  by specifying the valuation on  $\Omega$  as follows:

$$\llbracket p \rrbracket_M = \{(X, Y) \in \Omega : \models \varphi_X \rightarrow p\}.$$

In other words,  $p$  is true at  $(X, Y)$  just in case  $\varphi_X$  entails  $p$ . Note that the valuation only depends on the first component  $X$  of the state  $(X, Y)$ .

<sup>8</sup>“State-dependent” here means that the utility function constructed will depend not only on outcomes but on states as well.

Next we specify a parametrized family of well-orders on  $\Omega$  that we can use to induce a selection function. First define

$$(X, X') \sqsubseteq_{W, W'} (Y, Y') \text{ iff } \varphi_X \leq_W \varphi_Y.$$

Again, we are ignoring the second component. This is clearly a well-order when restricted to the first component of the state space, but not in general, since by definition we have  $(X, X') \sqsubseteq_{W, W'} (Y, Y')$  and  $(Y, Y') \sqsubseteq_{W, W'} (X, X')$  whenever  $X = Y$ . However, as usual, we can extend these relations to well-orders  $\leq_{W, W'}$  on all of  $\Omega$  simply by choosing a linear order for each set of the form  $\Omega_X := \{(X, Y) : Y \in 2^\Phi\}$ , and in so doing we can insist that for each fixed  $X$ , the state  $(X, W)$  is  $\leq_{W, W'}$ -minimal on the set  $\Omega_X$ .

This is the first time we have paid attention to the second component of the state. Roughly speaking, we are ensuring that the order  $\leq_{W, W'}$  “remembers” the set  $W$ . More perspicuously, it is easy to see that if  $c$  is the selection function induced by the family  $\{\leq_{(W, W')} : (W, W') \in \Omega\}$ , then for each  $(W, W') \in \Omega$  and all  $\varphi \in \mathcal{L}$ , we have

$$\llbracket do(\varphi) \rrbracket_{M, c}(W, W') = c((W, W'), \varphi) = (\min_{\leq}(W, \varphi), W). \quad (2)$$

That is, the closest  $\varphi$ -state to  $(W, W')$  encodes both the closest atom compatible with  $\varphi$  to  $\varphi_W$  (in the first component) *and* the state  $W$  that we started from (in the second component).

Now we can define our utility function and probability measure. Let  $\pi$  be any probability measure on  $\Omega$  satisfying  $\pi(\Omega_X) > 0$  for all  $X$ . Next, define  $u : \Omega \rightarrow \mathbb{R}$  by

$$u(X, W) = \frac{u^*(W, \varphi)}{\pi(\Omega_W)}, \text{ for some } \varphi \text{ such that } \min_{\leq}(W, \varphi) = X.$$

Of course, we need to check that  $u$  is well-defined, and we do so in Lemma 6. But first some intuition is in order. Thinking back to the state-dependent utility function  $u^*$ , a reasonable first gloss of the meaning of  $u^*(W, \varphi)$  might be “the utility of doing  $\varphi$  in  $W$ ”.<sup>9</sup> The point is that  $u^*$  is specifying the utility value not of an action in itself or the “result” of an action, but rather the result of an action *if you started in a certain state*. This is all very informal, but the idea is just to provide some intuition for why, in defining our utility function  $u$  from  $u^*$ , we need to appeal to a rich enough notion of state that can “remember” what the “previous” state was—intuitively, the state we were at before the action was performed.

**Lemma 6.** *The function  $u$  is well-defined.*

The last thing we need to show is that the selection model  $(M, c)$  we have built, along with  $\pi$  and  $u$ , gives us an expected utility representation of  $\succeq$ . So let  $\alpha, \beta \in \mathcal{A}_F$  and suppose that  $\alpha \succeq \beta$ . By definition this is equivalent to  $f_\alpha \succeq^* f_\beta$ , which by the state-dependent representation result is in turn equivalent to

$$\sum_{W \in 2^\Phi} u^*(W, f_\alpha(W)) \geq \sum_{W \in 2^\Phi} u^*(W, f_\beta(W)). \quad (3)$$

Now observe that, for each  $W \in 2^\Phi$ ,

$$\begin{aligned} u^*(W, f_\alpha(W)) &= \pi(\Omega_W) \cdot u(\min_{\leq}(W, f_\alpha(W)), W) && \text{(by definition of } u) \\ &= \pi(\Omega_W) \cdot u(\llbracket do(f_\alpha(W)) \rrbracket_{M, c}(W, W')) && \text{(from (2))} \\ &= \pi(\Omega_W) \cdot u(\llbracket \alpha \rrbracket_{M, c}(W, W')) && \text{(by definition of } f_\alpha \text{ and } (M, c)). \end{aligned}$$

<sup>9</sup>Though this isn't quite right—it's more like the product of that utility with the probability of  $W$ , which is why we have to factor that probability out in defining our utility function.

Note that in the above  $W'$  can be *any* element of  $2^\Phi$ , since it's not taken into account in determining the result of an action. That means we can rewrite the above as

$$u^*(W, f_\alpha(W)) = \sum_{W' \in 2^\Phi} \pi(W, W') \cdot u(\llbracket \alpha \rrbracket_{M,c}(W, W')).$$

Of course, an analogous equation holds for  $u^*(W, f_\beta(W))$ . Thus, (3) is equivalent to:

$$\sum_{W \in 2^\Phi} \sum_{W' \in 2^\Phi} \pi(W, W') \cdot u(\llbracket \alpha \rrbracket_{M,c}(W, W')) \geq \sum_{W \in 2^\Phi} \sum_{W' \in 2^\Phi} \pi(W, W') \cdot u(\llbracket \beta \rrbracket_{M,c}(W, W')),$$

which is exactly the right-hand side of (1), completing the proof.  $\square$

## 4 Discussion

We have considered a framework in which both the conditions for and the results of an action are given by simple descriptions in a fixed language. These descriptions may not be maximally specific, so the results of actions can be underspecified and therefore “open to interpretation”. We have shown that, in this context, agents whose preferences satisfy certain constraints can be represented as if they are expected utility maximizers who interpret each underspecified action using a selection function identical to that employed in standard semantics for counterfactual conditionals.

The representation theorem presented in this extended abstract might be viewed as a sort of “proof of concept”, namely, that such representation results are possible and even natural. This opens the door for a variety of related results connecting different assumptions about the selection function to different constraints on the agent's preferences. As we mentioned above, there are a number of standard assumptions along these lines in the literature on counterfactuals.

The underlying language we chose to work with can also be altered. Perhaps most obviously, we might consider allowing countably-many primitive propositions. In this case, we cannot straightforwardly use atoms as the basis for the state space in the representation theorem, and in general we might need to relax the notion of a “complete description” to something like a “sufficiently detailed description”. Going in the other direction, we might also consider dropping some of the richness constraints we imposed. For instance, we assumed that  $F$  contains all atoms (and all pairwise disjunctions of atoms). Can this assumption be relaxed?

In our framework, because we use the same descriptions for both states and outcomes, we found it convenient to identify the two. This in turn makes it straightforward to extend to a richer language of acts, where we allow *sequential actions*, implemented directly by function composition. That is, we can allow actions of the form  $do(\varphi); do(\psi)$  (“first do  $\varphi$ , then do  $\psi$ ”), or more generally,  $\alpha; \beta$ . Thus, the (underspecified!) results of the first action are directly relevant to the conditions under which the second action is executed, which may allow for entirely new and intriguing ways of encoding modeling features via constraints on preferences.

Finally, generalizing this framework to multiple agents is of interest. Indeed, the original motivation for this work is doubly relevant in multi-agent settings: two different decision-makers might conceive of the same action in different ways, by associating it with different functions. For example, we should be able to model two agents who agree about their values and have the same beliefs about the likelihoods of uncertain events, but still have different preferences over actions—intuitively, because they interpret the “default” way of implementing actions differently (in other words, they have the same utility function and probability measure, but different selection functions).

In short, this area is ripe for further exploration, with many theoretical and practical applications.

## A Proofs

**Lemma 1.** *If  $(M, c)$  is a selection model,  $c$  satisfies centering, and  $\models \psi \rightarrow \phi$ , then*

$$\llbracket \mathbf{if} \ \psi \ \mathbf{then} \ do(\phi) \rrbracket_{M,c} = id_{\Omega} = \llbracket do(true) \rrbracket_{M,c}.$$

*Proof.* By definition, we have

$$\begin{aligned} \llbracket \mathbf{if} \ \psi \ \mathbf{then} \ do(\phi) \rrbracket_{M,c}(\omega) &= \begin{cases} \llbracket do(\phi) \rrbracket_{M,c}(\omega) & \text{if } \omega \in \llbracket \psi \rrbracket \\ \llbracket do(true) \rrbracket_{M,c}(\omega) & \text{if } \omega \notin \llbracket \psi \rrbracket. \end{cases} \\ &= \begin{cases} c(\omega, \phi) & \text{if } \omega \in \llbracket \psi \rrbracket \\ c(\omega, true) & \text{if } \omega \notin \llbracket \psi \rrbracket. \end{cases} \end{aligned}$$

But since  $\llbracket \psi \rrbracket \subseteq \llbracket \phi \rrbracket$  by assumption, in either case, centering applies and guarantees that

$$\llbracket \mathbf{if} \ \psi \ \mathbf{then} \ do(\phi) \rrbracket_{M,c}(\omega) = \omega. \quad \square$$

**Lemma 2.** *If  $(M, c)$  is a selection model where  $c$  is induced by the well-orders  $\leq = \{\leq_{\omega} : \omega \in \Omega\}$ ,  $\leq$  is language-based,  $\models \phi \leftrightarrow (\phi_1 \vee \dots \vee \phi_n)$ , and  $X \subseteq \Phi$ , then  $\exists i \in \{1, \dots, n\}$  such that for all  $\psi$  satisfying  $\models \phi_i \rightarrow \psi$  and  $\models \psi \rightarrow \phi$  and all  $\omega \in \llbracket \phi_X \rrbracket$ , we have  $\llbracket do(\psi) \rrbracket_{M,c}(\omega) = \llbracket do(\phi_i) \rrbracket_{M,c}(\omega)$ .*

*Proof.* Let  $\omega \in \llbracket \phi_X \rrbracket$  and choose  $i$  such that  $c(\omega, \phi) \in \llbracket \phi_i \rrbracket$ . This is possible since we know  $c(\omega, \phi) \in \llbracket \phi \rrbracket$  and, by assumption,  $\llbracket \phi \rrbracket = \llbracket \phi_1 \rrbracket \cup \dots \cup \llbracket \phi_n \rrbracket$ . Since  $c(\omega, \phi)$  is the  $\leq_{\omega}$ -minimal element of  $\llbracket \phi \rrbracket$ , it follows that for any set  $T$  with  $c(\omega, \phi) \in T \subseteq \llbracket \phi \rrbracket$ ,  $c(\omega, \phi)$  is also the  $\leq_{\omega}$ -minimal element of  $T$ . In particular, since  $c(\omega, \phi) \in \llbracket \phi_i \rrbracket \subseteq \llbracket \psi \rrbracket \subseteq \llbracket \phi \rrbracket$ , this implies that  $c(\omega, \phi)$  is the  $\leq_{\omega}$ -minimal element of both  $\llbracket \phi_i \rrbracket$  and  $\llbracket \psi \rrbracket$ . Thus, by definition,  $c(\omega, \phi_i) = c(\omega, \psi)$ , so

$$\llbracket do(\psi) \rrbracket_{M,c}(\omega) = c(\omega, \psi) = c(\omega, \phi_i) = \llbracket do(\phi_i) \rrbracket_{M,c}(\omega).$$

Since  $\omega \models \phi_X$  and this completely determines the theory of  $\omega$ , we know that for any other  $\omega' \in \llbracket \phi_X \rrbracket$ ,  $\omega' \equiv \omega$ , so  $c(\omega', \phi) \equiv c(\omega, \phi)$ . This guarantees that  $c(\omega', \phi) \in \llbracket \phi_i \rrbracket$ ; in other words, the same choice of  $i$  works for all states in  $\llbracket \phi_X \rrbracket$ , which completes the proof.  $\square$

**Lemma 3.** (SSC) *implies that each  $\sqsubseteq_W$  is complete.*

*Proof.* Fix any two atoms  $\phi_X$  and  $\phi_Y$ . We apply (SSC) in the case where  $\phi = \phi_X \vee \phi_Y$ ,  $\phi_1 = \phi_X$ ,  $\phi_2 = \phi_Y$ , and  $\psi = \phi$ . Then we know that given any  $W \subseteq \Phi$ , either  $do(\phi) \sim_W do(\phi_1)$  or  $do(\phi) \sim_W do(\phi_2)$ , that is, either  $do(\phi_X \vee \phi_Y) \sim_W do(\phi_X)$  or  $do(\phi_X \vee \phi_Y) \sim_W do(\phi_Y)$ , which established completeness.  $\square$

**Lemma 4.** *If (SSC) and (Trans) hold, then each  $\sqsubseteq_W$  is a total preorder and can be extended to a well-order  $\leq_W$  on the set of atoms; if, in addition, (Cent) holds, then each  $\leq_W$  can be defined so that  $\phi_W$  is the  $\leq_W$ -minimal element.*

*Proof.* The fact that  $\sqsubseteq_W$  is a total preorder follows immediately from (Trans) and Lemma 3. Moreover, it is easy to see that any total preorder on a finite set can be extended to a well-order (by choosing an arbitrary linear order for each subset of  $\sqsubseteq_W$ -equivalent atoms). To see that this can be done in such a way that  $\phi_W$  is the  $\leq_W$ -minimal element, it suffices to show that for every  $X \subseteq \Phi$ , we have  $\phi_W \sqsubseteq_W \phi_X$ , or in other words,  $do(\phi_W \vee \phi_X) \sim_W do(\phi_W)$ . The result now follows from two applications of (Cent). First we apply it in the case where  $\psi = \phi_W$  and  $\phi = \phi_W \vee \phi_X$  to obtain  $(\mathbf{if} \ \phi_W \ \mathbf{then} \ do(\phi_W \vee \phi_X)) \sim do(true)$ ;

then we apply it in the case where  $\psi = \varphi = \varphi_W$  to obtain  $(\mathbf{if} \varphi_W \mathbf{then} do(\varphi_W)) \sim do(true)$ . Transitivity of  $\sim$  therefore yields

$$(\mathbf{if} \varphi_W \mathbf{then} do(\varphi_W \vee \varphi_X)) \sim (\mathbf{if} \varphi_W \mathbf{then} do(\varphi_W)),$$

which by definition is equivalent to  $do(\varphi_W \vee \varphi_X) \sim_W do(\varphi_W)$ .  $\square$

**Lemma 5.** *If (SSC) and (Trans) hold, then  $do(\varphi) \sim_W do(\varphi_{\min_{\leq}(W, \varphi)})$ .*

*Proof.* Let  $X = \min_{\leq}(W, \varphi)$ , and let  $\varphi_{X_1}, \dots, \varphi_{X_n}$  enumerate all the atoms compatible with  $\varphi$ . Then by definition we know that  $X = X_j$  for some  $j$ . We also clearly have  $\models \varphi \leftrightarrow (\varphi_{X_1} \vee \dots \vee \varphi_{X_n})$ , so we can apply (SSC) (taking  $\psi = \varphi$ ) to find an  $i$  such that  $do(\varphi) \sim_W do(\varphi_{X_i})$ .

By definition of  $X$ , we know that  $\varphi_X \leq_W \varphi_{X_i}$ , which means  $do(\varphi_X \vee \varphi_{X_i}) \sim_W do(\varphi_X)$ . On the other hand, since  $\models \varphi_{X_i} \rightarrow (\varphi_X \vee \varphi_{X_i})$  and  $\models (\varphi_X \vee \varphi_{X_i}) \rightarrow \varphi$ , (SSC) also tells us (taking  $\psi = \varphi_X \vee \varphi_{X_i}$  this time) that  $do(\varphi_X \vee \varphi_{X_i}) \sim_W do(\varphi_{X_i})$ . By transitivity of  $\sim_W$  we therefore have  $do(\varphi_{X_i}) \sim_W do(\varphi_X)$ , and therefore  $do(\varphi) \sim_W do(\varphi_X)$ , as desired.  $\square$

**Lemma 6.** *The function  $u$  is well-defined.*

*Proof.* What we need to show that is that if  $\min_{\leq}(W, \varphi) = X$  and also  $\min_{\leq}(W, \varphi') = X$ , then  $u^*(W, \varphi) = u^*(W, \varphi')$ . By Lemma 5, we know that  $do(\varphi) \sim_W do(\varphi_X)$ , and also that  $do(\varphi') \sim_W do(\varphi_X)$ . Focusing on the first of these two indifferences to begin with, by definition we have

$$\mathbf{if} \varphi_W \mathbf{then} do(\varphi) \sim \mathbf{if} \varphi_W \mathbf{then} do(\varphi_X).$$

Setting  $\alpha = \mathbf{if} \varphi_W \mathbf{then} do(\varphi)$  and  $\beta = \mathbf{if} \varphi_W \mathbf{then} do(\varphi_X)$ , it follows that  $f_\alpha \sim^* f_\beta$  (by definition of  $\succeq^*$ ). Thus, from the state-dependent representation result, we can deduce that

$$\sum_{Z \in 2^\Phi} u^*(Z, f_\alpha(Z)) = \sum_{Z \in 2^\Phi} u^*(Z, f_\beta(Z)).$$

But it's easy to see that whenever  $Z \neq W$ ,  $f_\alpha(Z) = f_\beta(Z)$ , so we can cancel all those terms in the equality above to arrive at  $u^*(W, f_\alpha(W)) = u^*(W, f_\beta(W))$ . This yields  $u^*(W, \varphi) = u^*(W, \varphi_X)$ , since clearly  $f_\alpha(W) = \varphi$  and  $f_\beta(W) = \varphi_X$ . Analogous reasoning starting from the fact that  $do(\varphi') \sim_W do(\varphi_X)$  leads us to  $u^*(W, \varphi') = u^*(W, \varphi_X)$ . Putting these together gives  $u^*(W, \varphi) = u^*(W, \varphi')$ , as desired.  $\square$

## References

- [1] D. Ahn (2008): *Ambiguity without a state space*. *Review of Economic Studies* 71(1), pp. 3–28, doi:10.1111/j.1467-937X.2007.00473.x.
- [2] D. Ahn & H. Ergin (2010): *Framing contingencies*. *Econometrica* 78(2), pp. 655–695, doi:10.3982/ECTA7019.
- [3] L. E. Blume, D. Easley & J. Y. Halpern (2006): *Redoing the Foundations of Decision Theory*. In: *Principles of Knowledge Representation and Reasoning: Proc. Tenth International Conference (KR '06)*, pp. 14–24. A longer version, entitled “Constructive decision theory”, can be found at <http://www.cs.cornell.edu/home/halpern/papers/behfinal.pdf>.
- [4] E. Dekel, B. Lipman & A. Rustichini (2001): *Representing preferences with a unique subjective state space*. *Econometrica* 69, pp. 891–934, doi:10.1111/1468-0262.00224.
- [5] P. Ghirardato (2001): *Coping with ignorance: unforeseen contingencies and non-additive uncertainty*. *Economic Theory* 17, pp. 247–276, doi:10.1007/PL00004108.

- [6] I. Gilboa & D. Schmeidler (2004): *Subjective distributions*. *Theory and Decision* 56, pp. 345–357, doi:10.1007/s11238-004-2596-7.
- [7] E. Karni (2006): *Subjective expected utility theory without states of the world*. *Journal of Mathematical Economics* 42, pp. 325–342, doi:10.1016/j.jmateco.2005.08.007.
- [8] D. Kreps (1992): *Static choice and unforeseen contingencies*. In P. Dasgupta, D. Gale & O. Hart, editors: *Economic Analysis of Markets and Games: Essays in Honor of Frank Hahn*, MIT Press, Cambridge, MA.
- [9] B. L. Lipman (1999): *Decision theory without logical omniscience: Toward an axiomatic framework for bounded rationality*. *Review of Economic Studies* 66, pp. 339–361, doi:10.1111/1467-937X.00090.
- [10] M. Machina (2006): *States of the World and State of Decision Theory*. In D. Meyer, editor: *The Economics of Risk*, W. E. Upjohn Institute.
- [11] J. Pearl (1995): *Causal diagrams for empirical research*. *Biometrika* 82(4), pp. 669–710, doi:10.1093/biomet/82.4.669.
- [12] L. J. Savage (1954): *Foundations of Statistics*. Wiley, New York.
- [13] R. C. Stalnaker (1968): *A theory of conditionals*. In N. Rescher, editor: *Studies in Logical Theory*, Blackwell, Oxford, U.K., pp. 98–112.
- [14] A. Tversky & D. J. Koehler (1994): *Support theory: A nonextensional representation of subjective probability*. *Psychological Review* 101(4), pp. 547–567, doi:10.1037/0033-295X.101.4.547.



# An Awareness Epistemic Framework for Belief, Argumentation and Their Dynamics

Alfredo Burrieza

Antonio Yuste-Ginel

Department of Philosophy,  
University of Málaga, Spain

burrieza@uma.es

antonioyusteginel@gmail.com

The notion of argumentation and the one of belief stand in a problematic relation to one another. On the one hand, argumentation is crucial for belief formation: as the outcome of a process of arguing, an agent might come to (justifiably) believe that something is the case. On the other hand, beliefs are an input for argument evaluation: arguments with believed premisses are to be considered as strictly stronger by the agent to arguments whose premisses are not believed. An awareness epistemic logic that captures qualified versions of both principles was recently proposed in the literature. This paper extends that logic in three different directions. First, we try to improve its conceptual grounds, by depicting its philosophical foundations, critically discussing some of its design choices and exploring further possibilities. Second, we provide a (heretofore missing) completeness theorem for the basic fragment of the logic. Third, we study, using techniques from dynamic epistemic logic, how different forms of information change can be captured in the framework.

## 1 Introduction

Belief and argumentation are two central dimensions of humans' cognitive architecture. They have received attention from antiquity to nowadays, and from a broad range of disciplines. It is then unsurprising that formal researchers have undertaken the task of modelling both phenomena. Regarding beliefs, there is an important amount of options for capturing some of its formal aspects [22]. These models usually capture *what* kind of things are believed (typically, propositions or sentences); *who* believes them (intelligent agents); and, only sometimes, *how* strong or safe these beliefs are (for instance, in probabilistic models of belief or in plausibility structures [7]). However, most of them fail to capture *why* agents do believe certain things. This lack motivates the recent trial within the epistemic logic community of capturing the missing justification component. This enterprise has been approached from a variety of methods: justification logic [3, 4, 1, 2], evidence logics based on neighbourhood semantics [12, 11], and its further topological development [5], amongst others. Yet another natural candidate to model justification consists in using conceptual and technical tools coming from argumentation theory (as done, e.g. in [23, 36, 28, 15]).

As to argumentation theory, it is a well-established, interdisciplinary field of research [19]. Since the last few decades, formal argumentation has gained more and more attention within the field of artificial intelligence, and its general advantages have been highlighted several times [10, 32]. Within formal approaches to argumentation, it is frequent to distinguish between *abstract approaches* (those that consider arguments as primitive, atomic entities) and *structured approaches* (those that explicitly account for the structure of arguments). For expository purposes, we just mention the popular Dung's approach to abstract argumentation [18], based on so-called *abstract argumentation frameworks*, and the ASPIC family of formalisms for structured argumentation, e.g., ASPIC<sup>+</sup> [30, 31], that will be the main argumentative resources used in this paper.

Recently, some works have taken the first steps to explore and exploit the relations among the two different traditions (epistemic logic and formal argumentation). These can be divided in two groups. On the one hand, there are works using epistemic logic tools to reason about argumentation frameworks [35, 34, 33]. On the other hand, there are works using argumentation tools to provide an (argumentatively inspired) notion of justified belief (the already mentioned [23, 36, 28, 15]). The current paper is inserted in the latter group, and it follows the ideas of [15] that, contrarily to [23, 36, 28], and according to more standard ideas in structured argumentation, decides to model arguments as *syntactic entities*.

We start by pointing out that the informal relation between argumentation and belief is itself problematic. Arguably, there is a tension between two intuitive principles governing belief formation and argument evaluation. These principles are:

- P1 *Beliefs are an input for argument evaluation*, meaning that arguments with believed premisses are better to those with contingent or even rejected premisses.<sup>1</sup>
- P2 *Argumentation is an input for belief formation*, meaning that rational agents should believe sentences that are ground in good arguments.

The mentioned tension arises when one tries to embrace both principles without any restriction, leading to an infinite regress. A very similar problem can be found in the root of a long-standing debate about the structure of epistemic justification within contemporary epistemology. *Foundationalist* solutions to such a tension, to which we adhere here, consists in distinguishing between *basic (non-inferred) beliefs* and *non-basic (inferred) beliefs*, where the latter inherit the justification from the former [26]. This implies accepting qualified version of both principles, but giving some sort of priority to P1 over P2. Curiously enough, an analogous distinction can be found as one of the basis of the recent argumentative theory of reason advocated by Mercier and Sperber [29]. In this context, basic beliefs are called *intuitive beliefs* while inferred beliefs are called *reflective beliefs* (see [37] for a detailed exposition of the distinction).

In the rest of this paper, we follow up the work made in [15], by extending it in three different directions. First, and after recalling the logic introduced there, whose language allows talking about basic beliefs and structured arguments, we provide its sound and complete axiomatisation (Section 2). We then explain how to use this logic for reasoning about *explicit basic beliefs* and *argument-based belief*, discussing some of the design choices, as well as depicting some alternatives (Section 3). Finally, we extend the basic fragment of the logic so as to capture different kinds of informational dynamics, illustrating their effects on both types of beliefs (Section 4).

## 2 An awareness logic for belief and argumentation

Let us start by recalling the logic introduced in [15]. We follow the traditional order for presentation: syntax, semantics, and proof theory. We assume a countable set of *propositional letters*  $At$  as fixed from now on. The **language**  $\mathcal{L}$  is defined as the pair  $(F, A)$  of *formulas* and *arguments* which are respectively generated by the following grammars:

$$\begin{aligned} \varphi &::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \Box\varphi \mid \text{aware}(\alpha) \mid \text{conc}(\alpha) = \varphi \mid \\ &\mid \text{strict}(\alpha) \mid \text{undercuts}(\alpha, \alpha) \mid \text{wellshap}(\alpha) \quad p \in At, \alpha \in A. \\ \alpha &::= \langle \varphi \rangle \mid \langle \alpha_1, \dots, \alpha_n \rightarrow \varphi \rangle \mid \langle \alpha_1, \dots, \alpha_n \Rightarrow \varphi \rangle \quad \varphi \in F, n \geq 1. \end{aligned}$$

<sup>1</sup>We use the term *contingent* in its doxastic sense, that is, a sentence is said to be contingent iff it is neither believed nor believed to be false.

The rest of Boolean operators ( $\rightarrow, \vee, \leftrightarrow$ ) and constants ( $\top, \perp$ ), as well as the dual of  $\Box$  (noted  $\Diamond$ ), are defined as usual. Arguments of  $\mathcal{L}$  have the following informal readings.  $\langle \varphi \rangle$  is an *atomic argument*. Note that this kind of arguments are rather strange in real-life examples, since they have one sole premise and conclusion, and there is not a proper inference step. Mathematically, they can be understood as a one-line proof from  $\varphi$  to  $\varphi$ . As for  $\langle \alpha_1, \dots, \alpha_n \rightarrow \varphi \rangle$  (resp.  $\langle \alpha_1, \dots, \alpha_n \Rightarrow \varphi \rangle$ ), it represents an argument claiming that  $\varphi$  follows deductively (resp. defeasibly) from the conclusions of arguments  $\alpha_1, \dots, \alpha_n$ . As an example of a complex argument consider  $\langle \langle \langle \text{Bird} \rangle, \langle \text{Bird} \rightarrow \text{Wings} \rangle \rightarrow \text{Wings} \rangle \Rightarrow \text{Flies} \rangle$  that informally reads “This has wings, because it is a bird and all birds have wings. Moreover, since it has wings, it presumably (defeasibly) flies”.

Regarding formulas, elements of  $\text{At}$  represent factual, atomic propositions.  $\Box\varphi$  means that the agent implicitly (ideally) believes that  $\varphi$ .  $\text{aware}(\alpha)$  reads “the agent is aware of  $\alpha$ ”.  $\text{conc}(\alpha) = \varphi$  reads the “conclusion of  $\alpha$  is  $\varphi$ ”.  $\text{strict}(\alpha)$  means that  $\alpha$  does not contain defeasible inference steps.  $\text{undercuts}(\alpha, \beta)$  means that  $\alpha$  undercuts  $\beta$ , that is,  $\alpha$  attacks some defeasible inference link of  $\beta$ . Finally,  $\text{wellshap}(\alpha)$  means that  $\alpha$  has been constructed properly, that is, all its deductive inference steps are valid and all its defeasible inference steps are accepted by the agent.

We use  $\text{SEQ}(F)$  to denote the *set of all finite sequences over F*. We denote an arbitrary sequence of  $n+1$  elements over  $F$  as  $((\varphi_1, \dots, \varphi_n), \varphi)$ . Sequences of formulas are useful to represent inference steps in the meta-language. Although strongly connected from a conceptual point of view, the sequence  $((\varphi_1, \dots, \varphi_n), \varphi)$  is *not* the same object as, for instance, the object language argument  $\langle \langle \varphi_1 \rangle, \dots, \langle \varphi_n \rangle \Rightarrow \varphi \rangle$ . Let  $R = ((\varphi_1, \dots, \varphi_n), \varphi) \in \text{SEQ}(F)$  we use  $\alpha^R$  as a shorthand for  $\langle \langle \varphi_1 \rangle, \dots, \langle \varphi_n \rangle \Rightarrow \varphi \rangle$ . We can see  $\alpha^R$  as the *simplest argument using R*. As an example, consider the rule  $R_1 = ((\text{Wings}), \text{Flies})$ , we have  $\alpha^{R_1} = \langle \langle \text{Wings} \rangle \Rightarrow \text{Flies} \rangle$ , but note that there are infinitely many other arguments using  $R_1$ , for instance  $\langle \langle \langle \text{Bird} \rangle, \langle \text{Bird} \rightarrow \text{Wings} \rangle \rightarrow \text{Wings} \rangle \Rightarrow \text{Flies} \rangle$ .

Let us define the following meta-syntactic functions for analysing an **argument’s structure**, taken from  $\text{ASPIC}^+$  [30]:

$\text{Prem}(\alpha)$  returns the **premises** of  $\alpha$  and it is defined as follows:  $\text{Prem}(\langle \varphi \rangle) := \{\varphi\}$ ,  $\text{Prem}(\langle \alpha_1, \dots, \alpha_n \leftrightarrow \varphi \rangle) := \text{Prem}(\alpha_1) \cup \dots \cup \text{Prem}(\alpha_n)$  where  $\leftrightarrow \in \{\rightarrow, \Rightarrow\}$ .

$\text{Conc}(\alpha)$  returns the **conclusion** of  $\alpha$  and it is defined as follows  $\text{Conc}(\langle \varphi \rangle) := \{\varphi\}$  and  $\text{Conc}(\langle \alpha_1, \dots, \alpha_n \leftrightarrow \varphi \rangle) := \{\varphi\}$  where  $\leftrightarrow \in \{\rightarrow, \Rightarrow\}$ .

$\text{sub}_A(\alpha)$  returns the **subarguments** of  $\alpha$  and it is defined as follows:  $\text{sub}_A(\langle \varphi \rangle) := \{\langle \varphi \rangle\}$  and  $\text{sub}_A(\langle \alpha_1, \dots, \alpha_n \leftrightarrow \varphi \rangle) := \{\langle \alpha_1, \dots, \alpha_n \leftrightarrow \varphi \rangle\} \cup \text{sub}_A(\alpha_1) \cup \dots \cup \text{sub}_A(\alpha_n)$  where  $\leftrightarrow \in \{\rightarrow, \Rightarrow\}$ .

$\text{TopRule}(\alpha)$  returns the **top rule** of  $\alpha$ , i.e. the last rule applied in the formation of  $\alpha$ . It is defined as follows:  $\text{TopRule}(\langle \varphi \rangle)$  is left undefined,  $\text{TopRule}(\langle \alpha_1, \dots, \alpha_n \rightarrow \varphi \rangle) = \text{TopRule}(\langle \alpha_1, \dots, \alpha_n \Rightarrow \varphi \rangle) := ((\text{Conc}(\alpha_1), \dots, \text{Conc}(\alpha_n)), \varphi)$ .

$\text{DefRule}(\alpha)$  returns the set of **defeasible rules** of  $\alpha$  and it is defined as  $\text{DefRule}(\langle \varphi \rangle) := \emptyset$ ,  $\text{DefRule}(\langle \alpha_1, \dots, \alpha_n \rightarrow \varphi \rangle) := \text{DefRule}(\alpha_1) \cup \dots \cup \text{DefRule}(\alpha_n)$  and  $\text{DefRule}(\langle \alpha_1, \dots, \alpha_n \Rightarrow \varphi \rangle) := \{((\text{Conc}(\alpha_1), \dots, \text{Conc}(\alpha_n)), \varphi)\} \cup \text{DefRule}(\alpha_1) \cup \dots \cup \text{DefRule}(\alpha_n)$ .

Let us also define **semantic propositional negations**, for any  $\varphi, \psi \in F$ :  $\varphi = \sim \psi$  abbreviates  $\text{wellshap}(\langle \langle \varphi \rangle \rightarrow \neg \psi \rangle) \wedge \text{wellshap}(\langle \langle \psi \rangle \rightarrow \neg \varphi \rangle)$ .

Let us now move to semantics. A **model** for  $\mathcal{L}$  is a tuple  $M = (W, \mathcal{B}, \mathcal{O}, \mathcal{D}, n, || \cdot ||)$  where:

- $W \neq \emptyset$  is a set of *possible worlds*.
- $\mathcal{B} \subseteq W$  and  $\mathcal{B} \neq \emptyset$  is the set of *worlds that are doxastically indistinguishable for the agent*.
- $\mathcal{O} \subseteq A$  is the set of *available arguments*, also called the *awareness set of the agent*.

- $\mathcal{D} \subseteq \text{SEQ}(F)$  is a set of *accepted defeasible rules*. Moreover, for every  $((\varphi_1, \dots, \varphi_n), \varphi) \in \mathcal{D}$  we require that:
  - $\{\varphi_1, \dots, \varphi_n, \varphi\} \not\vdash_0 \perp$  (defeasible rules are consistent), where  $\vdash_0$  denotes the consequence relation of classical propositional logic, and
  - $\{\varphi_1, \dots, \varphi_n\} \not\vdash_0 \varphi$  (defeasible rules are not deductively valid).
- $n : \text{SEQ}(F) \rightarrow \text{At}$  is a (possibly partial) *naming function* for rules, where  $n(R)$  informally means “the rule  $R$  is applicable”.
- $\|\cdot\|$  is and an *atomic valuation*, i.e. a function  $\|\cdot\| : \text{At} \rightarrow \wp(W)$ .

**Interpretation.** In a given model  $M = (W, \mathcal{B}, \mathcal{O}, \mathcal{D}, n, \|\cdot\|)$ ,  $\mathcal{O}$  represents the set of arguments that the agent entertains or is aware of. Whenever  $\alpha \in \mathcal{O}$ , we assume that (i) the agent can determine her doxastic attitude toward the premisses of  $\alpha$  through non-inferential methods (for instance, through observations), and (ii) she knows the structure of  $\alpha$  (either because  $\alpha$  has been communicated to her, or because she has gone through the cognitive process of building  $\alpha$ ). Besides this, there is not semantic intuition underlying  $\mathcal{O}$ , so the agent can be perfectly aware of rather silly arguments, as  $\langle\langle p \rangle \rightarrow q \rangle$ , without accepting them in any sense. Moreover, rules in the set  $\mathcal{D}$  are interpreted as rules whose inference strength lies in their content, rather than as purely formal schemas (as deductive rules are). As an example, consider the rule “Peter’s bike is on the bike parking area, therefore he should be in his office”. The term *accepted* means that the agent considers them applicable if there are not good reasons against doing so. Note that  $\alpha^R \in \mathcal{O}$  does not imply  $R \in \mathcal{D}$  (informally corresponding to the intuition that an agent can be aware of a defeasible argument without accepting its rule). There are further restrictions that could be arguably adopted, but that we leave out for the sake of simplicity. For instance, we could require  $\mathcal{O}$  to be closed under subarguments, or that for any accepted defeasible rule, the agent is aware of at least an argument using it.

Let  $M = (W, \mathcal{B}, \mathcal{O}, \mathcal{D}, n, \|\cdot\|)$  be a model for  $\mathcal{L} = (F, A)$ . The **set of well-shaped arguments**  $WS^M \subseteq A$  (depending on  $\mathcal{D}$  in  $M$ ) is the smallest set fulfilling the following conditions:

1.  $\langle \varphi \rangle \in WS^M$  for any  $\varphi \in F$ .
2.  $\langle \alpha_1, \dots, \alpha_n \rightarrow \varphi \rangle \in WS^M$  iff both  $\alpha_i \in WS^M$  for every  $1 \leq i \leq n$  and  $\{\text{Conc}(\alpha_1), \dots, \text{Conc}(\alpha_n)\} \vdash_0 \varphi$ .
3.  $\langle \alpha_1, \dots, \alpha_n \Rightarrow \varphi \rangle \in WS^M$  iff both  $\alpha_i \in WS^M$  for every  $1 \leq i \leq n$  and  $((\text{Conc}(\alpha_1), \dots, \text{Conc}(\alpha_n)), \varphi) \in \mathcal{D}$ .

We drop the superscript  $M$  from  $WS^M$  whenever there is no danger of confusion.

Let  $(M, w)$  be a pointed model for  $\mathcal{L}$ , that is,  $M = (W, \mathcal{B}, \mathcal{O}, \mathcal{D}, n, \|\cdot\|)$  is a model and  $w \in W$ . The **truth** relation, relating pointed models and formulas, is given by:<sup>2</sup>

$$\begin{array}{lll}
 M, w \models \Box \varphi & \text{iff} & \text{for all } w' \in W: w' \in \mathcal{B} \text{ implies } M, w' \models \varphi. \\
 M, w \models \text{aware}(\alpha) & \text{iff} & \alpha \in \mathcal{O}. \\
 M, w \models \text{conc}(\alpha) = \varphi & \text{iff} & \text{Conc}(\alpha) = \varphi. \\
 M, w \models \text{strict}(\alpha) & \text{iff} & \text{DefRule}(\alpha) = \emptyset. \\
 M, w \models \text{undercuts}(\alpha, \beta) & \text{iff} & \beta = \langle \beta_1, \dots, \beta_n \Rightarrow \psi \rangle \text{ and } \text{Conc}(\alpha) = \neg n(\text{TopRule}(\beta)). \\
 M, w \models \text{wellshap}(\alpha) & \text{iff} & \alpha \in WS^M.
 \end{array}$$

<sup>2</sup>Note that we do not need to consider undercuts as a primitive operator, since it could be defined through a (simpler) operator that captures the meaning of  $n$ . We make this choice for the sake of succinctness, as well as for studying the axiomatic behaviour of undercuts.

A formula  $\varphi$  is said to be **valid** (noted  $\models \varphi$ ) iff it is true at all pointed models. We use  $\|\varphi\|_M$  to denote the **truth-set** of  $\varphi$ , i.e., the set of worlds of  $M$  where  $\varphi$  is true, and  $\mathcal{M}$  to denote the class of all models.

We now present a sound and complete axiomatisation of  $\mathcal{L}$  w.r.t.  $\mathcal{M}$ , a topic that was left out in [15], and that constitutes one of the main technical contributions of the current paper. Although our models provide a compact representation of the needed components for reasoning about basic and argument-based beliefs in a single-agent context, they are rather non-standard from a technical point of view. Besides the strongly syntactic character of some of their elements, their modal components are not defined as usual, therefore the definition of its canonical model cannot be extrapolated straightforwardly. Nevertheless, we can provide an indirect completeness proof (see Appendix A1 for details).

**Theorem 1.** *The axiom system  $\mathbb{L}^{\text{BA}}$ , defined in Table 1, is sound and complete for  $\mathcal{L}$  w.r.t.  $\mathcal{M}$ .*

### 3 Basic beliefs and argument-based beliefs

The logic introduced above can be used to study a rich repertoire of doxastic attitudes. We start by discussing *basic beliefs*, informally representing those that are not grounded on inferential processes. As mentioned, they can also be understood in terms of *intuitive beliefs*, i.e., those that the agent extracts from a sort of data-base, seen by her as completely trustworthy [37]. As usual in awareness epistemic logic, we have two versions of this notion. On the one hand, we have the implicit (ideal) version of basic beliefs, modelled through  $\Box\varphi$ , that suffers from the extensively discussed problem of logical omniscience (see e.g. [21, Chapter 9]). On the other hand, we have its explicit counterpart, for which we have chosen  $\Box^e\varphi := \Box\varphi \wedge \text{aware}(\langle\varphi\rangle)$ . Note that, like in other logics for implicit and explicit belief, it holds that  $\models \Box^e\varphi \rightarrow \Box\varphi$ . Moreover, under the current semantics,  $\Box^e\varphi$  is equivalent to a schema that resembles another usual option for modelling explicit beliefs (e.g. [39]):  $\models \Box^e\varphi \leftrightarrow \Box(\varphi \wedge \text{aware}(\langle\varphi\rangle))$ .

Besides basic beliefs, we can also capture in  $\mathcal{L}$  a sort of deductive-explicit belief. Deductive-explicit beliefs are those rooted in a deductive argument s.t. the agent has a basic belief that all its premisses are true. Formally, and following [6], we define **doxastic argument acceptance** as

$$\text{accept}(\alpha) := \bigwedge_{\varphi \in \text{Prem}(\alpha)} \Box\varphi,$$

and **deductive-explicit belief** as

$$\text{B}^{\text{D}}(\alpha, \varphi) := \text{accept}(\alpha) \wedge \text{aware}(\alpha) \wedge \text{conc}(\alpha) = \varphi \wedge \text{strict}(\alpha) \wedge \text{wellshap}(\alpha).$$

Note that  $\models \text{B}^{\text{D}}(\alpha, \varphi) \rightarrow \Box\varphi$  and  $\models \Box^e\varphi \leftrightarrow \text{B}^{\text{D}}(\langle\varphi\rangle, \varphi)$ . The first validity shows that deductive-explicit beliefs are a subset of basic-implicit beliefs. The second one shows that basic-explicit beliefs are an extreme case of deductive-explicit beliefs (those that are rooted in the trivial deduction that goes from  $\varphi$  to  $\varphi$ , i.e., in the atomic argument  $\langle\varphi\rangle$ ).

Up to now, we have not gone far from the kind of attitudes that are usually discussed in the awareness logic literature (e.g. in [20, 13, 24, 25]). We now take a small detour through argumentation theory in order to define argument-based beliefs. Roughly speaking, argument-based beliefs are grounded in arguments that may involve non-deductive steps. They can be understood, at least to some extent, in terms of the *reflective beliefs* of [37]. Recall that we are after formalising the principle P2 presented in the introduction: the beliefs of a rational agent should be grounded in good arguments. But, what does it mean *good* in this context? Following [9], the very notion of argument strength can be analysed in three different layers or dimensions: the *support dimension* (how strong is the reason given by an argument to accept its conclusion), the *dialectic dimension* (how arguments attack and defeat each other), and the *evaluative dimension* (how the former conflicts are to be solved).

<b>Modal core axioms</b>	
(Ax0)	All propositional tautologies
(Ax1)	$KD45$ axioms for $\Box$
<b>Introspection axioms</b>	
(Ax2)	$\text{aware}(\alpha) \rightarrow \Box \text{aware}(\alpha)$
(Ax3)	$\neg \text{aware}(\alpha) \rightarrow \Box \neg \text{aware}(\alpha)$
(Ax4)	$\text{wellshap}(\alpha) \rightarrow \Box \text{wellshap}(\alpha)$
(Ax5)	$\neg \text{wellshap}(\alpha) \rightarrow \Box \neg \text{wellshap}(\alpha)$
(Ax6)	$\text{undercuts}(\alpha, \beta) \rightarrow \Box \text{undercuts}(\alpha, \beta)$
(Ax7)	$\neg \text{undercuts}(\alpha, \beta) \rightarrow \Box \neg \text{undercuts}(\alpha, \beta)$
<b>Axioms for syntactic operators</b>	
(Ax8)	$\text{conc}(\alpha) = \varphi$ whenever $\text{Conc}(\alpha) = \varphi$
(Ax9)	$\neg \text{conc}(\alpha) = \varphi$ whenever $\text{Conc}(\alpha) \neq \varphi$
(Ax10)	$\text{strict}(\alpha)$ whenever $\text{DefRule}(\alpha) = \emptyset$
(Ax11)	$\neg \text{strict}(\alpha)$ whenever $\text{DefRule}(\alpha) \neq \emptyset$
<b>Wellshapedness axioms</b>	
(Ax12)	$\text{wellshap}(\langle \varphi \rangle)$
(Ax13)	$\text{wellshap}(\langle \alpha_1, \dots, \alpha_n \Rightarrow \varphi \rangle) \rightarrow \bigwedge_{1 \leq i \leq n} \text{wellshap}(\alpha_i)$
(Ax14)	$\bigwedge_{1 \leq i \leq n} \text{wellshap}(\alpha_i) \rightarrow \text{wellshap}(\langle \alpha_1, \dots, \alpha_n \Rightarrow \varphi \rangle)$ whenever $\{\text{Conc}(\alpha_i) \mid 1 \leq i \leq n\} \vdash_0 \varphi$
(Ax15)	$\neg \text{wellshap}(\langle \alpha_1, \dots, \alpha_n \Rightarrow \varphi \rangle)$ whenever $\{\text{Conc}(\alpha_i) \mid 1 \leq i \leq n\} \not\vdash_0 \varphi$
(Ax16)	$\left( \bigwedge_{1 \leq i \leq n} \text{wellshap}(\alpha_i) \wedge \text{wellshap}(\langle \langle \text{Conc}(\alpha_1) \rangle, \dots, \langle \text{Conc}(\alpha_n) \rangle \Rightarrow \varphi \rangle) \right)$ $\leftrightarrow \text{wellshap}(\langle \alpha_1, \dots, \alpha_n \Rightarrow \varphi \rangle)$
(Ax17)	$(\text{wellshap}(\langle \alpha_1, \dots, \alpha_n \Rightarrow \varphi \rangle) \rightarrow \neg \text{wellshap}(\langle \langle \text{Conc}(\alpha_1) \rangle, \dots, \langle \text{Conc}(\alpha_n) \rangle, \langle \varphi \rangle \Rightarrow \perp \rangle))$
(Ax18)	$(\text{wellshap}(\langle \alpha_1, \dots, \alpha_n \Rightarrow \varphi \rangle) \rightarrow \neg \text{wellshap}(\langle \alpha_1, \dots, \alpha_n \Rightarrow \varphi \rangle))$
<b>Undercut axioms</b>	
(Ax19)	$\text{undercuts}(\langle \neg p \rangle, \alpha^R) \rightarrow \neg \text{undercuts}(\langle \neg q \rangle, \alpha^R)$ whenever $q \neq p$
(Ax20)	$\neg \text{undercuts}(\alpha, \langle \varphi \rangle)$
(Ax21)	$\neg \text{undercuts}(\alpha, \langle \alpha_1, \dots, \alpha_n \Rightarrow \varphi \rangle)$
(Ax22)	$\neg \text{undercuts}(\alpha, \beta)$ whenever $\text{Conc}(\alpha) \neq \neg p$ for some $p \in \text{At}$
(Ax23)	$(\text{undercuts}(\langle \neg p \rangle, \langle \langle \text{Conc}(\beta_1) \rangle, \dots, \langle \text{Conc}(\beta_n) \rangle \Rightarrow \varphi \rangle) \wedge \text{conc}(\alpha) = \neg p) \rightarrow$ $\text{undercuts}(\alpha, \langle \beta_1, \dots, \beta_n \Rightarrow \varphi \rangle)$
(Ax24)	$(\text{undercuts}(\alpha, \langle \beta_1, \dots, \beta_n \Rightarrow \varphi \rangle) \wedge \text{conc}(\alpha) = \neg p) \rightarrow$ $\text{undercuts}(\langle \neg p \rangle, \langle \langle \text{Conc}(\beta_1) \rangle, \dots, \langle \text{Conc}(\beta_n) \rangle \Rightarrow \varphi \rangle)$
<b>Rules</b>	
(MP)	From $\varphi \rightarrow \psi$ and $\varphi$ infer $\psi$
(Nec)	From $\varphi$ , infer $\Box \varphi$

Table 1: Axiom system.

Hence, the first step is to set up a notion of argument strength regarding the *support dimension*. Formally, we seek to define a preference relation among the arguments of  $\mathcal{L}$  that takes into account P1 (arguments with believed premisses are to be preferred over arguments with premisses that are not believed). In [15], we showed how to do this by splitting all arguments in three preference classes that

were based on the basic doxastic attitude of the agent toward the premisses of the arguments. Here, we take a much simpler view, for the sake of brevity, and directly exclude arguments whose premisses are not believed. Both options makes P2 dependent on P1, since in the process of grounding arguments in beliefs and these in turn in new arguments, we arrive at good arguments that are good just because the agent has a basic belief that all their premisses hold. However, inference links must still play a role when determining the relative strength of two arguments, the simplest principle that can be adopted in this regard is captured by the following binary relation among arguments of  $\mathcal{L}$ :  $\alpha \geq \beta := \text{strict}(\alpha) \vee \neg \text{strict}(\beta)$ . This relation informally corresponds to the idea that, *ceteris paribus*, deductive arguments are to be preferred to non-deductive ones.

Regarding the *dialectic dimension* of argument strength, we capture two forms of argumentative defeat, namely, *undercutting* (attacking a defeasible inference step of any subargument) and *successful rebuttal* (attacking the conclusion of a less or equally preferred subargument). Formally,

- **Undercutting a subargument** undercuts<sup>\*</sup> $(\alpha, \beta) := \bigvee_{\beta' \in \text{sub}_A(\beta)} \text{undercuts}(\alpha, \beta')$ .
- **Unrestricted successful rebuttal**  
Urebutts $(\alpha, \beta) := \neg \text{strict}(\beta) \wedge \bigvee_{\beta' \in \text{sub}_A(\beta)} (\text{conc}(\alpha) = \varphi \wedge \text{conc}(\beta') = \psi \wedge \varphi = \sim \psi \wedge \alpha \geq \beta')$ .
- **Defeat** defeat $(\alpha, \beta) := \text{undercuts}^*(\alpha, \beta) \vee \text{Urebutts}(\alpha, \beta)$ .

As discussed in the formal argumentation literature, there is a more restrictive alternative for the notion of rebuttal, requiring the top rule of the attacked subargument to be defeasible<sup>3</sup>

- **Restricted successful rebuttal**  
Rrebutts $(\alpha, \beta) := \neg \text{strict}(\beta) \wedge \bigvee_{\langle \beta_1, \dots, \beta_n \Rightarrow \varphi \rangle \in \text{sub}_A(\beta)} (\text{conc}(\alpha) = \psi \wedge \varphi = \sim \psi)$ .

Argumentation frameworks and their semantics [18] are the most studied tool to capture the *evaluative dimension* of argument strength. We now explain how to incorporate them in the current approach. Let  $(M, w)$  be a pointed model for  $\mathcal{L} = (F, A)$ , we define its **associated argumentation framework** as  $AF^M := (A^M, \rightsquigarrow)$ , where  $A^M := \{\alpha \in A \mid M, w \models \text{aware}(\alpha) \wedge \text{wellshap}(\alpha) \wedge \text{accept}(\alpha)\}$  and  $\rightsquigarrow \subseteq A^M \times A^M$  is given by  $\alpha \rightsquigarrow \beta$  iff  $M, w \models \text{defeat}(\alpha, \beta)$ . We stress the fact that in the domain of our frameworks (i.e., in  $A^M$ ), basic beliefs act as a filter (in the clause  $\text{accept}(\alpha)$ ), instantiating a qualified, unproblematic version of P1, namely P1': *basic beliefs* are an input for argument evaluation. Given a set of possibly conflicting arguments (an argumentation framework), we need a mechanism for the agent to decide which of the arguments are to be selected (an argumentation semantics). We say that a set  $B \subseteq A^M$  is **conflict-free** iff there are no  $\alpha, \beta \in B$  s.t.  $\alpha \rightsquigarrow \beta$ . Moreover, we say that  $B \subseteq A^M$  **defends**  $\alpha \in A^M$  iff for every  $\gamma \in A^M$ ,  $\gamma \rightsquigarrow \alpha$  implies that there is  $\beta \in B$  s.t.  $\beta \rightsquigarrow \gamma$ . We say that  $B \subseteq A^M$  is a **complete extension** iff it is conflict-free and it contains precisely the elements of  $A^M$  that it defends. We say that  $B \subseteq A^M$  is the **grounded extension** of  $AF^M = (A^M, \rightsquigarrow)$  iff it is the smallest (w.r.t. set inclusion) complete extension. We use  $GE(AF^M)$  to denote the grounded extension of  $AF^M$ . As it is well-known, the grounded extension of an argumentation framework always exists and it is moreover unique [18]. The unfamiliar reader is referred to [8] for an extensive discussion on argumentation semantics.

Finally, we use the grounded extension to define the **argument-based beliefs** of the agent. First, let us extend  $\mathcal{L} = (F, A)$  to  $\mathcal{L}^{\text{AB}} = (F^{\text{AB}}, A)$  by adding a new kind of formulas  $B(\alpha, \varphi)$  where  $\alpha \in A$  and  $\varphi \in F$ .  $B(\alpha, \varphi)$  means that the agent believes that  $\varphi$  based on argument  $\alpha$ . We interpret the new language in the same class of models, by adding the truth clause:

$$M, w \models B(\alpha, \varphi) \quad \text{iff} \quad \alpha \in GE(AF^M) \quad \text{and} \quad \text{Conc}(\alpha) = \varphi.$$

<sup>3</sup>See [41] for a discussion about the two possible design choices. Note moreover that the other customary type of attack, i.e. *undermining* (attacking a premise), makes sense only when non-believed premisses are taken into account.

Note that  $\models \text{B}^{\text{D}}(\alpha, \varphi) \rightarrow \text{B}(\alpha, \varphi)$  and  $\models \Box^e \varphi \rightarrow \text{B}(\langle \varphi \rangle, \varphi)$ .

We close this section by analysing our notion of argument-based belief under the view of [16]’s *rationality postulates*. In a nutshell, if no restrictions are imposed, our agent behaves according to a kind of *minimal rationality* (i.e. she does not explicitly believe in inconsistencies). If, however, we add some ideal assumptions, then she satisfies all [16]’s postulates.

**Proposition 1.** *Let  $(M, w)$  be a pointed model for  $\mathcal{L} = (\text{F}, \text{A})$ , where  $M = (W, \mathcal{B}, \mathcal{O}, \mathcal{D}, \mathbf{n}, \|\cdot\|)$ . Let  $AF^M$  be its associated argumentation framework, then:*

- $AF^M$  satisfies direct consistency, that is, there are no  $\alpha, \beta \in \text{A}$  and  $\varphi, \psi \in \text{F}$  s.t.  $M, w \models \text{B}(\alpha, \varphi) \wedge \text{B}(\beta, \psi) \wedge \varphi \approx \sim \psi$ .
- If restricted rebuttal is assumed and  $\mathcal{O} = \text{A}$ , then  $AF^M$  satisfies direct consistency; indirect consistency (that is,  $\text{Conc}(GE(AF^M)) \not\vdash_0 \perp$ ); sub-argument closure (that is,  $\alpha \in GE(AF^M)$  implies  $\text{sub}_{\text{A}}(\alpha) \subseteq GE(AF^M)$ ); and strict closure (that is,  $\text{Conc}(GE(AF^M)) \vdash_0 \varphi$  implies  $\varphi \in \text{Conc}(GE(AF^M))$ ).

*Proof (sketched).* For the first item, we suppose the contrary, that is, that there are arguments  $\alpha, \beta \in GE(AF^M)$  s.t.  $\text{Conc}(\alpha) = \varphi$ ,  $\text{Conc}(\beta) = \psi$  and  $\varphi$  is propositionally equivalent to the negation of  $\psi$ . Then, we continue by cases on the shape of  $\alpha$  and  $\beta$  (each of them can be either an atomic argument, or an argument whose last inference step is deductive (resp. defeasible)). From the nine different cases, three of them are redundant. From the six remaining cases, it is easy to arrive to  $\alpha \rightsquigarrow \beta$  or  $\beta \rightsquigarrow \alpha$  (which contradicts the assumption that they are in the grounded extension, because it is conflict-free).

For the second item, it suffices to show that under both assumptions (adopting the definition of restricted rebuttal and assuming  $\mathcal{O} = \text{A}$ ), we are just working with an instance of *well-defined ASPIC<sup>+</sup>* frameworks (one constructed over a knowledge base where the set of ordinary premisses is empty), which is guaranteed to satisfy all [16]’s rationality postulates (see [31, Section 3.3] for details).  $\square$

## 4 Dynamics of information

The current framework can throw some light on the relations between dynamics of information, argumentation and doxastic attitudes. We can distinguish several kinds of actions, that have different potential effects on basic and argument-based beliefs. The framework naturally allows for the use of tools imported from *dynamic epistemic logic* (DEL) [17]. In particular, we can describe these actions using dynamic modalities, for which complete axiomatisations can be then provided by finding a full list of *reduction axioms* [27, 17, 40]. In order to do so, one first need to show that the rule of replacement of proved equivalents is sound (it preserves validity) in the extended language (see [27] for details). Although this is *not* the case in  $\mathcal{L}$ , as it happens with other languages containing awareness operators [20, 24], we can restrict the domain of application of the rule, and it still does the job for axiomatizing certain dynamic extensions. More precisely, we will work with the rule:

$$\text{(RE)} \text{ From } \varphi \leftrightarrow \psi, \text{ infer } \delta \leftrightarrow \delta[\varphi/\psi],$$

with  $\delta[\varphi/\psi]$  the result of replacing one or more non- $\star$  occurrences of  $\psi$  in  $\delta$  by  $\varphi$ .<sup>4</sup> Semantically, this amounts to showing that each of the actions act we are about to discuss is well defined, in the sense that whenever we compute  $M^{\text{act}}$  (the result of executing action act in model  $M$ ), we stay in the

<sup>4</sup>A non- $\star$  of  $\psi$  in  $\delta$  is just an occurrence of  $\psi$  in  $\delta$  where  $\psi$  is not inside the scope of  $\star \in \{\text{aware}, \text{conc}, \text{wellshap}, \text{undercuts}\}$ . Note that we assume that  $\varphi$  is inside the scope of  $\text{conc}$  in the formula  $\text{conc}(\alpha) = \varphi$ .

intended class of models. When this does not happen, as it is the case with many DEL actions (e.g. public announcements [17]), one needs to find a set of *preconditions* for the action. Preconditions works as sufficient conditions for the action to be “safe” i.e., to secure that after executing it, we stay in the intended class of models.

Let us start by defining four different actions. Let  $\mathcal{L} = (F, A)$  be given, let  $M = (W, \mathcal{B}, \mathcal{O}, \mathcal{D}, n, \|\cdot\|)$  be an  $\mathcal{L}$ -model, let  $\alpha \in A$ , let  $R \in \text{SEQ}(F)$ , and let  $\varphi \in F$ . We define:

- The act of **acquiring argument**  $\alpha$  (resp. **forgetting argument**  $\alpha$ ) produces the model  $M^{\alpha+!} := (W, \mathcal{B}, \mathcal{O}^{\alpha+!}, \mathcal{D}, n, \|\cdot\|)$ , where  $\mathcal{O}^{\alpha+!} := \mathcal{O} \cup \{\alpha\}$  (resp.  $M^{\alpha-!} := (W, \mathcal{B}, \mathcal{O}^{\alpha-!}, \mathcal{D}, n, \|\cdot\|)$ , where  $\mathcal{O}^{\alpha-!} := \mathcal{O} \setminus \{\alpha\}$ ).
- The act of **accepting the defeasible rule**  $R$  produces the model  $M^{R+!} := (W, \mathcal{B}, \mathcal{O}, \mathcal{D}^{R+!}, n, \|\cdot\|)$ , where  $\mathcal{D}^{R+!} := \mathcal{D} \cup \{R\}$ .
- The act of **publicly announcing**  $\varphi$  produces the model  $M^{\varphi!} := (W^{\varphi!}, \mathcal{B}^{\varphi!}, \mathcal{O}, \mathcal{D}, n, \|\cdot\|)$ , where  $W^{\varphi!} := W \cap \|\varphi\|_M$ ;  $\mathcal{B}^{\varphi!} := \mathcal{B} \cap \|\varphi\|_M$ ; and  $\|p\|_M^{\varphi!} := \|p\|_M \cap \|\varphi\|_M$  for every atom  $p$ .

**Interpretation.** Note that the definition is far from being exhaustive, we analyse them because they are natural adaptations of other actions studied in the literature [24, 17]. The most basic argumentative change we can think of consists in adding an argument into the awareness of the agent. Informally, this can be thought as the result of a communicative event (e.g. an opponent advancing an argument), learning (the agent reading an argument in a book), or as the result of reflection (the own agent constructing an argument). Formally, the action is a direct generalization of the “consider” action defined for sentences in [24, 13]. Its straightforward counterpart is the act of forgetting an argument (i.e. dropping it from the awareness of the agent). As for the action  $(\cdot)^{R+!}$ , defeasible rules can also be learnt in different ways. For instance, an agent can learn the rule  $((\text{Bird}), \text{Flies})$  because an ornithologist told her, because she observed repeatedly that birds fly, or because she read it in a textbook. Finally, public announcements are probably the most studied action in DEL (see e.g. [17, Chapter 4]). This kind of announcements are supposed to be truthful and coming from a completely reliable source.

We now define a **dynamic language**, in order to talk about the different actions. Let  $\mathcal{L} = (F, A)$  be a language, formulas of the extended language  $\mathcal{L}^! = (F^!, A)$  are given by:

$$\varphi ::= \psi \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid [\alpha+!]\varphi \mid [\alpha-!]\varphi \mid [R+!]\varphi \mid [\psi!]\varphi \quad \psi \in F, \alpha \in A, R \in \text{SEQ}(F).$$

Let  $[\text{act}]$  be any of the dynamic modalities we have just defined, we use  $\langle \text{act} \rangle$  as an abbreviation of  $\neg[\text{act}]\neg$ , with  $\langle \text{act} \rangle \varphi$  informally meaning that action  $\text{act}$  can be executed and after executing it,  $\varphi$  holds.

Note that the class of all models  $\mathcal{M}$  is *not* closed under all defined actions. In particular, it is not closed under  $(\cdot)^{R+!}$  nor under  $(\cdot)^{\varphi!}$ . For the former, the reason is that only rules that are consistent and non-deductive can be learnt as defeasible (see the definition of model in Section 2). For the latter, only truthful formulas that do not trivialize the beliefs of the agent (in the sense of making  $\mathcal{B}$  empty), can be announced. This inconvenience is solved by fixing **preconditions** (expressible in  $\mathcal{L}$ ) for both actions.

Let  $R = ((\varphi_1, \dots, \varphi_n), \varphi) \in \text{SEQ}(F)$  and  $\varphi \in F$ , we define:

$$\text{pre}(R) := \neg \text{wellshap}(\langle \langle \varphi_1 \rangle, \dots, \langle \varphi_n \rangle, \langle \varphi \rangle \rightarrow \perp \rangle) \wedge \neg \text{wellshap}(\langle \langle \varphi_1 \rangle, \dots, \langle \varphi_n \rangle \rightarrow \varphi \rangle);$$

$$\text{pre}(\varphi!) := \varphi \wedge \diamond \varphi.$$

It is almost immediate to check that, for any pointed model  $(M, w)$ , any  $R = ((\varphi_1, \dots, \varphi_n), \varphi) \in \text{SEQ}(F)$ , and any  $\varphi \in F$  we have that:

$$\begin{aligned} & \{ \langle \varphi_1, \dots, \varphi_n, \varphi \rangle \not\leq_0 \perp \text{ and } \{ \langle \varphi_1, \dots, \varphi_n \rangle \not\leq_0 \varphi \} \text{ iff } M, w \models \text{pre}(R); \text{ and} \\ & (w \in \|\varphi\|_M \text{ and } \|\varphi\|_M \cap \mathcal{B} \neq \emptyset) \text{ iff } M, w \models \text{pre}(\varphi!). \end{aligned}$$

Moreover, note that  $M, w \models \text{pre}(R)$  iff  $M, u \models \text{pre}(R)$  for every  $u \in W$ . Let  $(M, w)$  be a pointed model with  $M = (W, \mathcal{B}, \mathcal{O}, \mathcal{D}, n, \|\cdot\|)$ , we define the truth clause for the new kind of formulas:

$$\begin{aligned} M, w \models [\sigma+!] \varphi & \text{ iff } M^{\sigma+!}, w \models \varphi, \\ M, w \models [\sigma-!] \varphi & \text{ iff } M^{\sigma-!}, w \models \varphi, \\ M, w \models [R+!] \varphi & \text{ iff } M, w \models \text{pre}(R), \text{ implies } M^{R+!}, w \models \varphi, \\ M, w \models [\varphi!] \psi & \text{ iff } M, w \models \text{pre}(\varphi!) \text{ implies } M^{\varphi!}, w \models \psi. \end{aligned}$$

Finally, we establish a completeness result for  $\mathcal{L}^!$  w.r.t.  $\mathcal{M}$ . Note that in Table 2,  $\pm$  denotes an arbitrary element of  $\{+, -\}$ .

**Proposition 2.** *The proof system  $\mathbb{L}_{\text{BA}}^!$  that extends the one of Table 1 with all axioms of Table 2 and it is closed under (RE) is sound and complete for  $\mathcal{L}^!$  w.r.t.  $\mathcal{M}$ .*

*Proof.* Soundness follows from the validity of all axioms and the validity-preserving character of (RE) in the extended language. Completeness follows from the usual reduction argument. In short, note that in the right-hand side of all axioms of Table 2, either the dynamic operator disappears or it is applied to a less complex formula than in the left-hand side. In the case of reduction axioms for  $[R!+]\text{wellshap}(\alpha)$ , either there are no dynamic modalities occurring in the right-hand side of the equivalence or they are applied to wellshap-formulas with less complex *arguments* than in the right-hand side. Therefore, we can define a meaning-preserving translation from  $F^!$  to  $F$  that, together with Theorem 1, provides the desired result. The validity-preserving character of (RE) in the extended language w.r.t.  $\mathcal{M}$  takes care of formulas with nested dynamic modalities. The reader is referred to [27] for details.  $\square$

We close this section by modelling a toy example, inspired by [38], and illustrating how actions affect argument-based beliefs. Suppose that an agent is wondering whether another agent, Harry, is a British subject (br). Suppose that the only basic-explicit belief she holds at the beginning is that *Harry was born in Bermuda* (be). Other pieces of relevant information are: *Harry's parents are aliens* (a), and that *the rule "If Harry is born in Bermuda, then he is presumably a British subject" is applicable* (r1). Let  $R_1 = ((\text{be}), \text{br})$ . We start with the model  $M_0 = (W, \mathcal{B}, \mathcal{O}, \mathcal{D}, n, \|\cdot\|)$ , where  $W = \mathcal{B} = \{w_0, w_1, w_2, w_3\}$ ,  $\mathcal{D} = \emptyset$ ,  $\mathcal{O} = \{\langle \text{be} \rangle\}$ ,  $n(R_1) = r1$ ,  $\|\text{be}\| = W$ ,  $\|\text{br}\| = \|r1\| = \{w_0, w_2\}$ , and  $\|a\| = \{w_0, w_1\}$ . It is then easy to check that  $M_0, w_0 \models \Box^e \text{be}$ . Moreover, we have that  $M_0, w_0 \models [R1+!][\alpha^{R1+!}]B(\alpha^{R1}, \text{br})$ . In words, after learning the rule  $R_1$  and becoming aware of the simplest argument using it, i.e.  $\langle \langle \text{be} \rangle \Rightarrow \text{br} \rangle$ , the agent has an argument-based belief that Harry is a British subject. If, however the agent learns subsequently from a completely trustworthy source that Harry's parents are alien (a), together with the rule  $R_2 = ((a), \neg r1)$ , and the argument  $\langle \langle a \rangle \Rightarrow \neg r1 \rangle$ , then she revises her argument-based belief about Harry's nationality. In symbols,  $M_0, w_0 \models [R1+!][\alpha^{R1+!}][a!][R2+!][\alpha^{R2+!}]\neg B(\alpha^{R1}, \text{br})$ .

## 5 Concluding remarks

**Closely related work.** From all the works we have commented throughout the paper, it seems that [24, 25] and [36] are the closest one to our approach. Regarding [24, 25], we have somehow generalize their *awareness of rules* to our awareness of arguments (abstracting away from other forms of awareness treated there). As for [36], their choice of modelling arguments semantically (as opens of a topology), permits a transparent axiomatisation of their notion of argument-based beliefs, which is easily guaranteed to be consistent (two of the weaknesses of our approach). On the other hand, we naturally treat arguments as first-class citizens in our language, and the argument-based beliefs of our agent escape from every form of logical omniscience (while the beliefs of [36]'s agent are still closed under equivalent formulas).

$[\alpha\pm!]p \leftrightarrow p$	$[\varphi!]p \leftrightarrow (\text{pre}(\varphi!) \rightarrow p)$
$[\alpha\pm!]\neg\varphi \leftrightarrow \neg[\alpha\pm!]\varphi$	$[\varphi!]\neg\psi \leftrightarrow (\text{pre}(\varphi!) \rightarrow \neg[\varphi!]\psi)$
$[\alpha\pm!](\varphi \wedge \psi) \leftrightarrow ([\alpha\pm!]\varphi \wedge [\alpha\pm!]\psi)$	$[\varphi!](\delta \wedge \psi) \leftrightarrow ([\varphi!]\delta \wedge [\varphi!]\psi)$
$[\alpha\pm!]\Box\varphi \leftrightarrow \Box[\alpha\pm!]\varphi$	$[\varphi!]\Box\psi \leftrightarrow (\text{pre}(\varphi!) \rightarrow \Box[\varphi!]\psi)$
$[\alpha\pm!]\text{aware}(\beta) \leftrightarrow \text{aware}(\beta) \text{ for } \alpha \neq \beta$	$[\varphi!]\text{aware}(\beta) \leftrightarrow (\text{pre}(\varphi!) \rightarrow \text{aware}(\beta))$
$[\alpha+!]\text{aware}(\alpha) \leftrightarrow \top$	
$[\alpha-!]\text{aware}(\alpha) \leftrightarrow \perp$	
$[\alpha\pm!]\text{conc}(\beta) = \varphi \leftrightarrow \text{conc}(\beta) = \varphi$	$[\varphi!]\text{conc}(\beta) = \psi \leftrightarrow (\text{pre}(\varphi!) \rightarrow \text{conc}(\beta) = \psi)$
$[\alpha\pm!]\text{strict}(\beta) \leftrightarrow \text{strict}(\beta)$	$[\varphi!]\text{strict}(\beta) \leftrightarrow (\text{pre}(\varphi!) \rightarrow \text{strict}(\beta))$
$[\alpha\pm!]\text{undercuts}(\beta, \gamma) \leftrightarrow \text{undercuts}(\beta, \gamma)$	$[\varphi!]\text{undercuts}(\beta, \gamma) \leftrightarrow (\text{pre}(\varphi!) \rightarrow \text{undercuts}(\beta, \gamma))$
$[\alpha\pm!]\text{wellshap}(\beta) \leftrightarrow \text{wellshap}(\beta)$	$[\varphi!]\text{wellshap}(\beta) \leftrightarrow (\text{pre}(\varphi!) \rightarrow \text{wellshap}(\beta))$
$[R+!]p \leftrightarrow (\text{pre}(R) \rightarrow p)$	$[R+!]\text{aware}(\alpha) \leftrightarrow (\text{pre}(R) \rightarrow \text{aware}(\alpha))$
$[R+!]\neg\varphi \leftrightarrow (\text{pre}(R) \rightarrow \neg[R+!]\varphi)$	$[R+!]\text{conc}(\alpha) = \varphi \leftrightarrow (\text{pre}(R) \rightarrow \text{conc}(\alpha) = \varphi)$
$[R+!](\varphi \wedge \psi) \leftrightarrow ([R+!]\varphi \wedge [R+!]\psi)$	$[R+!]\text{strict}(\alpha) \leftrightarrow (\text{pre}(R) \rightarrow \text{strict}(\alpha))$
$[R+!]\Box\varphi \leftrightarrow \Box[R+!]\varphi$	$[R+!]\text{undercuts}(\alpha, \beta) \leftrightarrow (\text{pre}(R) \rightarrow \text{undercuts}(\alpha, \beta))$
$[R+!]\text{wellshap}(\langle\varphi\rangle) \leftrightarrow \top$	
$[R+!]\text{wellshap}(\langle\langle\varphi_1\rangle, \dots, \langle\varphi_n\rangle \twoheadrightarrow \varphi\rangle) \leftrightarrow (\text{pre}(R) \rightarrow \text{wellshap}(\langle\langle\varphi_1\rangle, \dots, \langle\varphi_n\rangle \twoheadrightarrow \varphi\rangle))$	
$[R+!]\text{wellshap}(\alpha^R) \leftrightarrow \top$	
$[R+!]\text{wellshap}(\alpha^{R'}) \leftrightarrow (\text{pre}(R) \rightarrow \text{wellshap}(\alpha^{R'})) \quad \text{whenever } R \neq R'$	
$[R+!]\text{wellshap}(\langle\alpha_1, \dots, \alpha_n \hookrightarrow \varphi\rangle) \leftrightarrow$	
$\leftrightarrow (\text{pre}(R) \rightarrow (\bigwedge_{1 \leq i \leq n} [R+!]\text{wellshap}(\alpha_i) \wedge [R+!]\text{wellshap}(\langle\langle\text{Conc}(\alpha_1)\rangle, \dots, \langle\text{Conc}(\alpha_n)\rangle \hookrightarrow \varphi\rangle)))$	

Table 2: Reduction axioms for  $\mathcal{L}^1$ .

**Future work.** There are natural open paths for future work. An urgent task in the development of the logical aspects of the framework consists in axiomatizing (if possible) the argument-based belief operator  $B(\cdot, \cdot)$ . Moreover, the modal semantic apparatus of our models could be extended to plausibility structures [7], so as to model fine-grained preference between arguments, based on the agent's basic epistemic attitudes toward the premisses of the involved arguments (e.g. known premisses are to be preferred to strongly believed premisses, and the latter, in turn, are to be preferred to merely believed premisses). Finally, a multi-agent extension of the current framework could be used to model argument exchange in different kinds of scenarios (e.g. deliberation, persuasion dialogues or inquiry).

## References

- [1] Sergei Artemov (2018): *Justification Awareness Models*. In Sergei Artemov & Anil Nerode, editors: *Logical Foundations of Computer Science, LNCS 10703*, Springer, pp. 22–36, doi:10.1007/978-3-319-72056-2\_2.
- [2] Sergei Artemov & Melvin Fitting (2016): *Justification Logic*. In Edward N. Zalta, editor: *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University.
- [3] Sergei Artemov & Elena Nogina (2005): *Introducing justification into epistemic logic*. *Journal of Logic and Computation* 15(6), pp. 1059–1073, doi:10.1093/logcom/exi053.

- [4] Sergei N Artemov (2012): *The ontology of justifications in the logical setting*. *Studia Logica* 100(1-2), pp. 17–30, doi:10.1007/s11225-012-9387-x.
- [5] Alexandru Baltag, Nick Bezhanishvili, AybÜke Özgün & Sonja Smets (2016): *Justified Belief and the Topology of Evidence*. In Jouko Väänänen, Åsa Hirvonen & Ruy de Queiroz, editors: *Logic, Language, Information, and Computation, LNCS 9803*, Springer, pp. 83–103, doi:10.1007/978-3-662-52921-8\_6.
- [6] Alexandru Baltag, Bryan Renne & Sonja Smets (2012): *The Logic of Justified Belief Change, Soft Evidence and Defeasible Knowledge*. In Luke Ong & Ruy de Queiroz, editors: *Logic, Language, Information and Computation. WoLLIC 2012., LNCS 7456*, Springer, pp. 168–190, doi:10.1007/978-3-642-32621-9\_13.
- [7] Alexandru Baltag & Sonja Smets (2008): *A qualitative theory of dynamic interactive belief revision*. In Wiebe van der Hoek, Giacomo Bonanno & Michael Wooldridge, editors: *Logic and the foundations of game and decision theory (LOFT 7), Texts in Logic and Games 3*, Amsterdam University Press, pp. 9–58.
- [8] Pietro Baroni, Martin Caminada & Massimiliano Giacomin (2018): *Abstract argumentation frameworks and their semantics*. In Pietro Baroni, Dov M. Gabbay, Massimiliano Giacomin & Leendert van der Torre, editors: *Handbook of formal argumentation*, College Publications, pp. 159–236.
- [9] Mathieu Beirlaen, Jesse Heyninck, Pere Pardo & Christian Straßer (2018): *Argument strength in formal argumentation*. *IfCoLog Journal of Logics and their Applications* 5(3), pp. 629–675.
- [10] Trevor JM Bench-Capon & Paul E Dunne (2007): *Argumentation in artificial intelligence*. *Artificial intelligence* 171(10-15), pp. 619–641, doi:10.1016/j.artint.2007.05.001.
- [11] Johan van Benthem, David Fernández-Duque & Eric Pacuit (2014): *Evidence and plausibility in neighborhood structures*. *Annals of Pure and Applied Logic* 165(1), pp. 106–133, doi:10.1016/j.apal.2013.07.007.
- [12] Johan van Benthem, David Fernández-Duque, Eric Pacuit et al. (2012): *Evidence Logic: A New Look at Neighborhood Structures*. *Advances in modal logic* 9, pp. 97–118.
- [13] Johan van Benthem & Fernando R Velázquez-Quesada (2010): *The dynamics of awareness*. *Synthese* 177(1), pp. 5–27, doi:10.1007/s11229-010-9764-9.
- [14] Patrick Blackburn, Maarten De Rijke & Yde Venema (2010): *Modal Logic*. Cambridge University Press, doi:10.1017/CB09781107050884.
- [15] Alfredo Burrieza & Antonio Yuste-Ginel (2020): *Basic beliefs and argument-based beliefs in awareness epistemic logic with structured arguments*. In H. Prakken, S. Bistarelli, F. Santini & C. Taticchi, editors: *Proceedings of the COMMA 2020*, IOS Press, pp. 123–134, doi:10.3233/FAIA200498.
- [16] Martin Caminada & Leila Amgoud (2007): *On the evaluation of argumentation formalisms*. *Artificial Intelligence* 171(5-6), pp. 286–310, doi:10.1016/j.artint.2007.02.003.
- [17] Hans van Ditmarsch, Wiebe van der Hoek & Barteld Kooi (2007): *Dynamic epistemic logic*. Springer, doi:10.1007/978-1-4020-5839-4.
- [18] Phan Minh Dung (1995): *On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games*. *Artificial Intelligence* 77(2), pp. 321–357, doi:10.1016/0004-3702(94)00041-X.
- [19] Frans H. van Eemeren, Bart Garssen, Erik C. W. Krabbe, A. Francisca Snoeck Henkemans, Bart Verheij & Jean H. M. Wagemans (2014): *Handbook of Argumentation Theory*. Springer, doi:10.1007/978-90-481-9473-5.
- [20] Ronald Fagin & Joseph Y Halpern (1987): *Belief, awareness, and limited reasoning*. *Artificial intelligence* 34(1), pp. 39–76, doi:10.1016/0004-3702(87)90003-8.
- [21] Ronald Fagin, Joseph Y Halpern, Yoram Moses & Moshe Vardi (2004): *Reasoning about knowledge*. MIT press, doi:10.7551/mitpress/5803.001.0001.
- [22] Konstantin Genin & Franz Huber (2021): *Formal Representations of Belief*. In Edward N. Zalta, editor: *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University.

- [23] Davide Grossi & Wiebe van der Hoek (2014): *Justified Beliefs by Justified Arguments*. In Chitta Baral, Giuseppe De Giacomo & Thomas Eiter, editors: *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourteenth International Conference*, AAAI Press, pp. 131–140, doi:10.5555/3031929.3031947.
- [24] Davide Grossi & Fernando R. Velázquez-Quesada (2009): *Twelve Angry Men: A Study on the Fine-Grain of Announcements*. In Xiangdong He, John Horty & Eric Pacuit, editors: *Logic, Rationality, and Interaction*, Springer, pp. 147–160, doi:10.1007/978-3-642-04893-7\_12.
- [25] Davide Grossi & Fernando R. Velázquez-Quesada (2015): *Syntactic awareness in logical dynamics*. *Synthese* 192(12), pp. 4071–4105, doi:10.1007/s11229-015-0733-1.
- [26] Ali Hasan & Richard Fumerton (2018): *Foundationalist Theories of Epistemic Justification*. In Edward N. Zalta, editor: *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University.
- [27] Barteld Kooi (2007): *Expressivity and completeness for public update logics via reduction axioms*. *Journal of Applied Non-Classical Logics* 17(2), pp. 231–253, doi:10.3166/janc1.17.231-253.
- [28] Xu Li & Yi N. Wang (2020): *A Logic of Knowledge and Belief Based on Abstract Arguments*. In Mehdi Dastani, Huimin Dong & Leon van der Torre, editors: *Logic and Argumentation*, Springer, pp. 116–130, doi:10.1007/978-3-030-44638-3\_8.
- [29] Hugo Mercier & Dan Sperber (2011): *Why do humans reason? Arguments for an argumentative theory*. *Behavioral and brain sciences* 34(2), pp. 57–74, doi:10.1017/S0140525X10000968.
- [30] Sanjay Modgil & Henry Prakken (2013): *A general account of argumentation with preferences*. *Artificial Intelligence* 195, pp. 361–397, doi:10.1016/j.artint.2012.10.008.
- [31] Sanjay Modgil & Henry Prakken (2018): *Abstract rule-based argumentation*. In Pietro Baroni, Dov M. Gabbay, Massimiliano Giacomin & Leendert van der Torre, editors: *Handbook of formal argumentation*, College Publications, pp. 287–364.
- [32] Sanjay Modgil, Francesca Toni, Floris Bex, Ivan Bratko, Carlos I. Chesñevar, Wolfgang Dvořák, Marcelo A. Falappa, Xiuyi Fan, Sarah Alice Gaggl, Alejandro J. García, María P. González, Thomas F. Gordon, João Leite, Martin Možina, Chris Reed, Guillermo R. Simari, Stefan Szeider, Paolo Torroni & Stefan Woltran (2013): *The Added Value of Argumentation*, pp. 357–403. Springer, doi:10.1007/978-94-007-5583-3\_21.
- [33] Carlo Proietti & Antonio Yuste-Ginel (2021): *Dynamic epistemic logics for abstract argumentation*. *Synthese*, doi:10.1007/s11229-021-03178-5.
- [34] Chiaki Sakama & Tran Cao Son (2020): *Epistemic Argumentation Framework: Theory and Computation*. *Journal of Artificial Intelligence Research* 69, pp. 1103–1126, doi:10.1613/jair.1.12121.
- [35] François Schwarzentruber, Srdjan Vesic & Tjitze Rienstra (2012): *Building an Epistemic Logic for Argumentation*. In Luis Fariñas del Cerro, Andreas Herzig & Jérôme Mengin, editors: *Logics in Artificial Intelligence, LNCS 7519*, Springer, pp. 359–371, doi:10.1007/978-3-642-33353-8\_28.
- [36] Chenwei Shi, Sonja Smets & Fernando R. Velázquez-Quesada (2017): *Argument-based belief in topological structures*. In J. Lang, editor: *Proceedings TARK 2017. EPTCS*, Open Publishing Association, doi:10.4204/EPTCS.251.36.
- [37] Dan Sperber (1997): *Intuitive and reflective beliefs*. *Mind & Language* 12(1), pp. 67–83, doi:10.1111/j.1468-0017.1997.tb00062.x.
- [38] Stephen E. Toulmin ([1958] 2003): *The uses of argument*. Cambridge university press, doi:10.1017/CB09780511840005.
- [39] Fernando R. Velázquez-Quesada (2014): *Dynamic epistemic logic for implicit and explicit beliefs*. *Journal of Logic, Language and Information* 23(2), pp. 107–140, doi:10.1007/s10849-014-9193-0.
- [40] Yanjing Wang & Qinxiang Cao (2013): *On axiomatizations of public announcement logic*. *Synthese* 190(1), pp. 103–134, doi:10.1007/s11229-012-0233-5.

- [41] Zhe Yu, Kang Xu & Beishui Liao (2018): *Structured argumentation: Restricted rebut vs. unrestricted rebut*. *Studies in Logic* 11(3), pp. 3–17.

## Appendix (Proof sketch of Theorem 1)

The outline of the proof is as follows: we first define a new class of (non-standard) models for our language (which are Kripke models where the syntactic components –awareness, accepted rules and names of rule– are maintained throughout the accessibility relation). We then show two things: (i) we can go from pointed Kripke models to its generated submodels without losing  $\mathcal{L}$ -information (just as in the general modal case) and; (ii) we can transform systematically Kripke generated submodels into our models (again, without losing  $\mathcal{L}$ -information). Finally, we prove completeness w.r.t. the class of non-standard models and apply (i) and (ii) to obtain the desired result. Let us unfold some of the details.

First of all, we define a **Kripke model for**  $\mathcal{L} = (F, A)$  as a tuple  $S = (W, \mathcal{R}, \mathcal{O}, \mathcal{D}, n, \|\cdot\|)$  where:  $W \neq \emptyset$  is a set of *possible worlds*;  $\mathcal{R} \subseteq W \times W$  is a serial, transitive and euclidean relation;  $\mathcal{O} : W \rightarrow \wp(A)$  is a function assigning an awareness set  $\mathcal{O}(w)$  to each world  $w$ ;  $\mathcal{D} : W \rightarrow \wp(\text{SEQ}(F))$  (with  $n \in \mathbb{N}$ ) is a function assigning a set of *accepted defeasible rules*  $\mathcal{D}(w)$  to each world  $w$ ;  $n : (W \times \text{SEQ}(F)) \rightarrow \text{At}$  is a (possibly partial) *naming function* for defeasible rules, where  $n(w, R)$  informally means “the defeasible rule  $R$  is applicable at  $w$ ”; and  $\|\cdot\| : \text{At} \rightarrow \wp(W)$  is a valuation function. Moreover, we assume that for every  $w, w' \in W$ ,  $w\mathcal{R}w'$  implies  $\mathcal{O}(w) = \mathcal{O}(w')$ ,  $\mathcal{D}(w) = \mathcal{D}(w')$ , and  $n(w, R) = n(w', R)$ . We also assume that if  $((\varphi_1, \dots, \varphi_n), \varphi) \in \mathcal{D}(w)$ , then  $\{\varphi_1, \dots, \varphi_n, \varphi\} \not\vdash_0 \perp$  and  $\{\varphi_1, \dots, \varphi_n\} \not\vdash_0 \varphi$ .

Note that now  $WS$  sets depend on both the model and the world at which we are looking (since  $\mathcal{D}$  may vary from one world to another). Consequently, we use  $WS^S(w)$  to denote the set of well-shaped arguments at  $(S, w)$ .

Truth w.r.t. pointed Kripke models is denoted by  $\models_k$  and defined as follows (the missing clauses are as expected):

$$\begin{aligned} S, w \models_k \Box \varphi & \text{ iff } w\mathcal{R}v \text{ implies } S, v \models_k \varphi \\ S, w \models_k \text{aware}(\alpha) & \text{ iff } \alpha \in \mathcal{O}(w) \\ S, w \models_k \text{wellshap}(\alpha) & \text{ iff } \alpha \in WS^S(w) \\ S, w \models_k \text{undercuts}(\alpha, \beta) & \text{ iff } \beta = \langle \beta_1, \dots, \beta_n \Rightarrow \varphi \rangle \text{ and } \text{Conc}(\alpha) = \neg n(w, \text{TopRule}(\beta)). \end{aligned}$$

We say that a Kripke model  $S = (W, \mathcal{R}, \mathcal{O}, \mathcal{D}, n, \|\cdot\|)$  is **uniform** iff for every  $w, w' \in W$  it holds that: (i)  $\mathcal{O}(w) = \mathcal{O}(w')$ ; (ii)  $\mathcal{D}(w) = \mathcal{D}(w')$ ; and (iii)  $n(w, R) = n(w', R)$  for every  $R \in \mathcal{D}$ .  $\mathcal{K}$  denotes the class of all pointed Kripke models, and  $\mathcal{K}^u$  denotes the class of all uniform pointed Kripke models. We abuse of notation and use  $\mathcal{M}$  to denote the class of all *pointed* models (the standard ones that we defined in Section 2).

**Transformation lemmas.** Now, we need a couple of lemmas. The first one says that we can go from pointed Kripke models to Kripke uniform pointed models without losing  $\mathcal{L}$ -information, by taking generated submodels. We use  $S^w$  to denote the submodel of  $S$  generated by  $w$  (see [14, Chapter 2]).

**Lemma 1.** *Let  $(S, w) \in \mathcal{K}$ . We have that:*

- i)  $(S^w, w) \in \mathcal{K}^u$ ,  
i.e. each pointed-generated submodel of a Kripke model is a uniform Kripke model.
- ii) For every  $\varphi \in F$ ,  $(S, w) \models_k \varphi$  iff  $(S^w, w) \models_k \varphi$ , i.e. truth is preserved under generated submodels.

Item *i*) follows easily from the definition of generated submodel and uniform Kripke model. Item *ii*) can be proved by induction on  $\varphi$ .

The second lemma says that we can go from Kripke uniform models to our models (the standard ones, defined in Section 2) without losing  $\mathcal{L}$ -information.

**Lemma 2.** *For every uniform pointed Kripke model  $(S, v) \in \mathcal{K}^u$ , there is a pointed model  $(M, w) \in \mathcal{M}$  s.t. for every  $\varphi \in F$ :*

$$S, v \models_k \varphi \quad \text{iff} \quad M, w \models \varphi.$$

Let us define the function  $\tau$  for each uniform Kripke model as follows  $\tau(S, w) = (\tau(S), \tau(w))$  where  $\tau(w) = w$  and  $\tau(S) = (\tau(W), \tau(\mathcal{R}), \tau(\mathcal{O}), \tau(\mathcal{D}), \tau(\mathbf{n}), \tau(\|\cdot\|))$  s.t.:

$$\begin{aligned} \tau(W) &:= \{w\} \cup \mathcal{R}[w], \\ \tau(\mathcal{R}) &:= \mathcal{R}[w], \\ \tau(\mathcal{O}) &:= \mathcal{O}(w), \\ \tau(\mathcal{D}) &:= \mathcal{D}(w), \\ \tau(\mathbf{n}) &:= \{(R, p) \in \text{SEQ}(F) \times \text{At} \mid \mathbf{n}(w, R) = p\}, \\ \tau(\|\cdot\|) &:= \|\cdot\| \cap \tau(W) \text{ for every } p \in \text{At}. \end{aligned}$$

Now, it is easy to check that, for every  $(S, w) \in \mathcal{K}^u$ :  $\tau((S, w)) \in \mathcal{M}$ , that is  $\tau: \mathcal{K}^u \rightarrow \mathcal{M}$ . Once this is done, we can show that, for every  $\varphi \in F$ , it holds that:

$$S, w \models_k \varphi \quad \text{iff} \quad \tau(S, w) \models \varphi.$$

The proof of the last assertion is by induction on  $\varphi$  where the step for  $\varphi = \text{wellshap}(\alpha)$  is another inductive argument (on the construction on  $\alpha$ ).

**Completeness w.r.t. Kripke models.** We can now define the **canonical Kripke model** for  $\mathcal{L}$  as:

$$S^c = (W^c, \mathcal{R}^c, \mathcal{O}^c, \mathcal{D}^c, \mathbf{n}^c, \|\cdot\|^c),$$

where the definition of  $W^c$ ,  $\mathcal{R}^c$  y  $\|\cdot\|^c$  is as usual in modal logic [14]; while the definition of the rest of the elements mimics the one of awareness operators [20]:

$$\begin{aligned} \mathcal{O}^c(\Gamma) &:= \{\alpha \in A \mid \text{aware}(\alpha) \in \Gamma\}, \\ \mathcal{D}^c(\Gamma) &:= \{R \in \text{SEQ}(F) \mid \text{wellshap}(\alpha^R) \in \Gamma\}, \\ ((\Gamma, R), p) \in \mathbf{n}^c &\quad \text{iff} \quad \text{undercuts}(\langle \neg p, \alpha^R \rangle) \in \Gamma. \end{aligned}$$

Now, we need to prove:

**Lemma 3 (Canonicity).**  *$S^c$  is a Kripke model for  $\mathcal{L}$ .*

For showing that  $S^c$  satisfies all conditions, we reason using maximally-consistent set properties and our axiom system. As illustrations: semantic restrictions on the accessibility relations follows from (Ax1) (see e.g. [21] or [14]), while (Ax19) permits showing that  $\mathbf{n}^c$  is a function.

**Lemma 4 (Truth).** *For every  $\varphi \in F$ :  $\varphi \in \Gamma$  iff  $S^c, \Gamma \models_k \varphi$ .*

The proof proceeds by induction on  $\varphi$ . The Boolean and modal cases are standard [14]. The cases for operators aware, conc and strict are straightforward (they actually do not make use of the induction hypothesis, due to their syntactic character). The cases for  $\varphi = \text{undercuts}(\alpha, \beta)$  and  $\varphi = \text{wellshap}(\alpha)$  are slightly more compromised. For the latter, another inductive argument on the structure of  $\alpha$  is required.

**Completeness w.r.t. standard models.** Finally, completeness w.r.t. standard models can be proved as follows. Suppose  $\Gamma \not\models \varphi$ , then  $\Gamma \cup \{\neg\varphi\}$  is consistent. By Lindenbaum, we have that there is a  $\Gamma^+ \in W^c$  s.t.  $\Gamma \cup \{\neg\varphi\} \subseteq \Gamma^+$ . By the Truth Lemma we have that  $S^c, \Gamma^+ \models_k \Gamma \cup \{\neg\varphi\}$ . By item *ii*) of Lemma 1 we have that  $S^{c\Gamma^+}, \Gamma^+ \models_k \Gamma \cup \{\neg\varphi\}$  and by item *i*) we have that  $S^{c\Gamma^+}, \Gamma^+$  is a pointed uniform Kripke model. Then by Lemma 2 we know that  $\tau(S^{c\Gamma^+}, \Gamma^+) \models \Gamma \cup \{\neg\varphi\}$  which implies by definition of semantic logical consequence that  $\Gamma \not\models \varphi$ .



# Local Dominance

Emiliano Catonini

HSE University Moscow

Jingyi Xue

Singapore Management University

We present a local notion of dominance that speaks to the true choice problems among actions in a game tree and does not rely on global planning. When we do not restrict the ability of the players to do contingent reasoning, a reduced strategy is weakly dominant if and only if it prescribes a locally dominant action at every decision node, therefore any dynamic decomposition of a direct mechanism that preserves strategy-proofness is robust to the lack of global planning. Under a form of wishful thinking, we also show that strategy-proofness is robust to the lack of forward-planning. Moreover, from our local perspective, we can identify rough forms of contingent reasoning that are particularly natural. We construct a dynamic game that implements the Top Trading Cycles allocation under a minimal form of contingent reasoning, related to independence of irrelevant alternatives.



# Collective Argumentation: The Case of Aggregating Support-Relations of Bipolar Argumentation Frameworks

Weiwei Chen

Institute of Logic and Cognition and Department of Philosophy  
Sun Yat-sen University  
Guangzhou, China  
chenww26@mail2.sysu.edu.cn

In many real-life situations that involve exchanges of arguments, individuals may differ on their assessment of which supports between the arguments are in fact justified, i.e., they put forward different support-relations. When confronted with such situations, we may wish to aggregate individuals' argumentation views on support-relations into a collective view, which is acceptable to the group. In this paper, we assume that under bipolar argumentation frameworks, individuals are equipped with a set of arguments and a set of attacks between arguments, but with possibly different support-relations. Using the methodology in social choice theory, we analyze what semantic properties of bipolar argumentation frameworks can be preserved by aggregation rules during the aggregation of support-relations.

## 1 Introduction

The attack relation has played a significant role in formal argumentation [2, 11, 23]. However, recent years have seen a revived interest in the support relation between arguments in argumentation systems [4, 5, 6, 7, 26]. In these systems, an argument can not only attack another argument, but it can also support another one. For example, an argument can support another argument by confirming its premise or undermining one of its attackers. The support relation between arguments is vital in modeling debates in real life. Due to the incompleteness of information, or different positions, agents may have different opinions regarding the support relation between arguments. To see this, consider the following example:

**Example 1.** Consider a debate regarding the possible influence of artificial intelligence (AI) to the job market. Suppose that there are two arguments in this debate:

*A*: Artificial intelligence improves the degree of work automation

*B*: More people will lose their jobs due to AI

Given the fact that AI is able to perform more of the tasks done by humans, some occupations will decrease. Therefore, some people hold that argument *A* supports argument *B*. On the other hand, given that AI will improve the quality of the work being done by humans, lower the prices of goods and services, create economic advantages, and allow for the creation of new jobs in new occupations, some people hold that argument *A* does not support argument *B*. △

In many scenarios, such as court debate, parliament debate, policy advisory committee decision-making, agents may have different opinions on which supports between arguments are acceptable, which form argumentative stances of them. When a group of agents are engaged in a debate, we may wish to aggregate stances possessed by agents to obtain a collective decision agreed on by the group. To model the support relation between arguments, we consider the *bipolar argumentation framework (BAF)* [5, 6, 7], a formalism of Dung’s abstract argumentation framework [11]. Given that there is a broad discussion of the aggregation of argumentation systems with the attack relation [10, 25, 12, 9], it is far from being clear what consensus can be achieved when the support relation is involved in this process. The goal of this paper is to investigate the aggregation of views of a group of agents in the context of bipolar argumentation. Given a set of arguments and a set of attack-relations between these arguments, agents might conflict with one another upon supports between arguments, i.e., for every pair of arguments that is being considered in a debate whether the first supports the second. In this scenario, we may wish to aggregate such support-relations.

In this paper, we use the method from *graph aggregation*, a recent discipline of social choice theory that deals with aggregating several graphs into a single output graph that constitutes a good compromise. Following the model introduced by Chen and Endriss [9], we consider the preservation of properties of bipolar argumentation frameworks, i.e., given a property that is satisfied by individual BAFs, we study whether it can be satisfied in the BAF returned by aggregation rules. For some properties, we show that there is an aggregation rule or a family of aggregation rules preserve them. For some others, we show that any aggregation rule that satisfies certain basic axioms and preserves them must be a dictatorship.

**Paper overview** The rest of the paper is organized as follows. In Section 2, we recapitulate the bipolar argumentation framework, along with its semantics. We introduce our model for the aggregation of support-relations of bipolar argumentation frameworks in Section 3, followed by our results of preservation in Section 4. In Section 5, we introduce some work related to our work. Finally, in Section 6, we conclude this work and point out some directions for future work.

## 2 Bipolar argumentation

An abstract bipolar argumentation framework [5, 6, 7] is an extension of Dung’s abstract argumentation framework [11] in which a general support relation between arguments is added. Formally, an abstract bipolar argumentation framework is a triple  $\langle Arg, \rightarrow, \rightsquigarrow \rangle$ , where  $Arg$  is a set of arguments,  $\rightarrow$  is a binary relation on  $Arg$ , which is called the attack relation,  $\rightsquigarrow$  is a binary relation on  $Arg$ , which is called the support relation. Given two arguments  $A, B \in Arg$ , if  $A \rightarrow B$  holds, then we say that  $A$  attacks  $B$ , if  $A \rightsquigarrow B$ , then we say that  $A$  supports  $B$ . The attack relation and the support relation must verify the following consistency constraint:  $\rightarrow \cap \rightsquigarrow = \emptyset$ , which is called *essential constraint*.

**Definition 1.** Let  $A, B \in Arg$ , there is a sequence of supports for  $B$  by  $A$  iff there exists a sequence of elements  $(A_1, \dots, A_n)$  of  $Arg$  such that  $n \geq 2, A = A_1, B = A_n, A_1 \rightsquigarrow A_2, \dots, A_{n-1} \rightsquigarrow A_n$ .

**Definition 2.** Let  $A, B \in Arg$ , a supported attack against  $B$  by  $A$  is a sequence of arguments  $(A_1, \dots, A_n)$  of  $Arg$  such that  $A_1 \rightsquigarrow, \dots, \rightsquigarrow A_{n-1}, A_{n-1} \rightarrow A_n, A = A_1, A_n = B$ , and  $n \geq 3$ .

Note that if  $A \rightarrow B$  is the case, then we say that  $A$  directly attacks  $B$ .

**Definition 3.** A secondary attack against an argument  $B$  by an argument  $A$  is a sequence  $(A_1, \dots, A_n)$  of arguments of  $Arg$  such that  $A_1 \rightarrow A_2, A_2 \rightsquigarrow \dots, \rightsquigarrow A_n, A = A_1, A_n = B$ , and  $n \geq 2$ .

For example, in Figure 1,  $A_1$  supported attacks  $E_1$ , while  $A_2$  secondary attacks  $E_2$ .

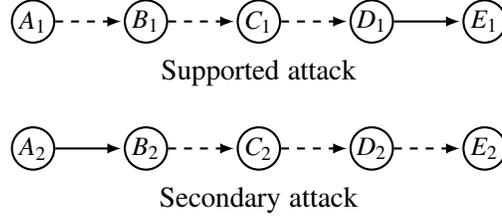


Figure 1: Illustration of supported attack and secondary attack

**Definition 4.** Let  $\Delta \subseteq \text{Arg}$  and  $A \in \text{Arg}$ .  $\Delta$  set-attacks  $A$  iff there exists a supported attack or a secondary attack against  $A$  from an element of  $\Delta$ .  $\Delta$  set-supports  $A$  iff there exists a sequence of supports for  $A$  from an element of  $\Delta$ .

**Definition 5.** Let  $\Delta \subseteq \text{Arg}$  be a set of arguments,  $\Delta$  is conflict-free iff  $\nexists A, B \in \Delta$  such that  $\{A\}$  set-attacks  $B$ .

**Definition 6.** Let  $\Delta \subseteq \text{Arg}$  be a set of arguments,  $\Delta$  is safe iff  $\nexists B \in \text{Arg}$  such that  $\Delta$  set-attacks  $B$  and either  $\Delta$  set-supports  $B$ , or  $B \in \Delta$ .

In the context of bipolar argumentation, admissibility can be translated into d-admissibility, s-admissibility and c-admissibility, based on different lines of coherence. In the following definition, the notion of *defense* is the same as classical defense, namely, we say  $\Delta \subseteq \text{Arg}$  defends the argument  $B \in \text{Arg}$ , then, there is an argument  $C \in \Delta$  with  $C \rightarrow A$  for all arguments  $A \in \text{Arg}$  such that  $A \rightarrow B$ .

**Definition 7.** Let  $\Delta \subseteq \text{Arg}$  be a set of arguments,  $\Delta$  is called d-admissible iff  $\Delta$  is conflict-free and defends all its elements;  $\Delta$  is a d-preferred extension if it is maximal (w.r.t. set-inclusion) among all d-admissible sets.

**Definition 8.** Let  $\Delta \subseteq \text{Arg}$  be a set of arguments,  $\Delta$  is called s-admissible iff  $\Delta$  is safe and defends all its elements;  $\Delta$  is a s-preferred extension if it is maximal (w.r.t. set-inclusion) among all s-admissible sets.

Let the closure of  $\Delta \subseteq \text{Arg}$  be  $\text{CL}(\Delta) = \{A \in \text{Arg} \mid \text{there is a sequence of supports from } B \in \Delta \text{ to } A\}$ , we say  $\Delta$  is closed iff  $\Delta = \text{CL}(\Delta)$ .

**Definition 9.** Let  $\Delta \subseteq \text{Arg}$  be a set of arguments,  $\Delta$  is called c-admissible iff  $\Delta$  is conflict-free, self-defending and closed;  $\Delta$  is a c-preferred extension if it is maximal (w.r.t. set-inclusion) among all c-admissible sets.

We restate a proposition in [5] that demonstrates the relation between safety and conflict-freeness.

**Proposition 1.** Let  $\Delta \subseteq \text{Arg}$  be a set of arguments, if  $\Delta$  is safe, then  $\Delta$  is conflict-free. If  $\Delta$  is conflict-free and closed, then  $\Delta$  is safe.

**Definition 10.** Let  $\Delta \subseteq \text{Arg}$  be a set of arguments,  $\Delta$  is stable if and only if  $\Delta$  is conflict-free and for every argument  $A \in \text{Arg} \setminus \Delta$ ,  $\Delta$  set-attacks  $A$ .

It is worth mentioning that in the original papers, [5, 6] consider a particular set of BAFs, namely acyclic BAFs, showing that such BAFs have some nice features. However, in this paper, we focus on BAFs that are more general, i.e., we remove the restriction on BAFs and consider both acyclic and cyclic BAFs. From a technical point of view, the BAFs that are acyclic have only one stable extension, which is the only preferred extension as well, while the BAFs with cycles could have more than one stable extension and will be more general.

There are several interpretations of support in the literature, including the deductive support, the necessary support, and the evidential support (see an overview in [7]). The deductive support [4] is intended to capture the intuition that given two arguments  $A$  and  $B$ , if  $A$  supports  $B$ , then the acceptance of  $A$  implies the acceptance of  $B$ . The necessary support [19, 20] is intended to capture the intuition that if  $A \rightsquigarrow B$  is the case, then the acceptance of  $B$  implies the acceptance of  $A$ , i.e., the acceptance of  $A$  is necessary to obtain the acceptance of  $B$ . Finally, the evidential support [22, 21] proposes a new type of argument, namely *prima-facie* arguments. Every standard argument is supposed to be supported by at least one *prima-facie* argument, and every *prima-facie* argument does not require support from other arguments.

The supported attack is connected with deductive support. To see this, let us come back to Figure 1, according to the deductive support, the acceptance of  $A_1$  implies the acceptance of  $B_1$ , and so the acceptance of  $C_1$ , the acceptance of  $D_1$ . In the meantime, the acceptance of  $D_1$  implies the non-acceptance of  $E_1$ . Thus, the acceptance of  $A_1$  implies the non-acceptance of  $E_1$ . The necessary support can be taken into account by considering secondary attack. We again consider Figure 1. First, the acceptance of  $A_2$  implies the non-acceptance of  $B_2$ . Then, according to necessary support, the non-acceptance of  $B_2$  implies the non-acceptance of  $C_2$ , and so the non-acceptance of  $D_2$ , the non-acceptance of  $E_2$ . Thus, the acceptance of  $A_2$  implies the non-acceptance of  $E_2$ .

### 3 The model

Fix a finite set  $Arg$  of arguments, a set  $(\rightarrow)$  of attacks between arguments, and a set  $N = \{1, \dots, n\}$  of  $n$  agents. Each agent  $i \in N$  supplies us with a set of supports  $\rightsquigarrow_i$ , which together with  $Arg$  and  $(\rightarrow)$  gives rise to a bipolar argumentation framework  $\langle Arg, \rightarrow, \rightsquigarrow_i \rangle$ , reflecting her individual views on which supports between arguments are acceptable. A *profile* of support-relations  $\rightsquigarrow = (\rightsquigarrow_1, \dots, \rightsquigarrow_n)$  is a set of support-relations provided by agents. An aggregation rule  $F : (2^{Arg \times Arg})^n \rightarrow 2^{Arg \times Arg}$  is a function that maps a given profile of support-relations into a single support-relation. We denote  $N_{sup}^{\rightsquigarrow}$  by the set of agents who accept  $sup$  under profile  $\rightsquigarrow$ , i.e.,  $N_{sup}^{\rightsquigarrow} = \{i \in N \mid sup \in \rightsquigarrow_i\}$ , and  $\#N_{sup}^{\rightsquigarrow}$  denotes the number of such agents.

Here we define desirable properties of aggregation rules. These properties are referred as axioms in the social choice literature. We start with formal definitions, followed by informal descriptions.

**Definition 11.** An aggregation rule  $F$  is unanimous if  $\rightsquigarrow_1 \cap \dots \cap \rightsquigarrow_n \subseteq F(\rightsquigarrow)$ .

**Definition 12.** An aggregation rule  $F$  is grounded if  $F(\rightsquigarrow) \subseteq \rightsquigarrow_1 \cup \dots \cup \rightsquigarrow_n$ .

**Definition 13.** An aggregation rule  $F$  is neutral if for any profile of support-relations  $\rightsquigarrow$ , for any pair of supports  $sup_1, sup_2$ ,  $N_{sup_1}^{\rightsquigarrow} = N_{sup_2}^{\rightsquigarrow}$  then  $sup_1 \in F(\rightsquigarrow)$  iff  $sup_2 \in F(\rightsquigarrow)$ .

**Definition 14.** An aggregation rule  $F$  is independent if for any pair of profiles of support-relations  $\rightsquigarrow_1, \rightsquigarrow_2$ , for any support  $sup$ ,  $N_{sup}^{\rightsquigarrow_1} = N_{sup}^{\rightsquigarrow_2}$  then  $sup \in F(\rightsquigarrow_1)$  iff  $sup \in F(\rightsquigarrow_2)$ .

**Definition 15.** An aggregation rule  $F$  is dictatorial if there is an agent  $i$  such that for any profile of support-relations  $\rightsquigarrow$ ,  $F(\rightsquigarrow) = \rightsquigarrow_i$ .

The *unanimity* axiom states that the support agreed by all agents should be included in the collective BAF. The *groundedness* axiom expresses that all supports in the collective BAF should be supported by at least one agent. The *neutrality* axiom requires that given a profile of support-relations, any pair of supports should be treated equally in this profile. The *independent* axiom states that all support-relations should be treated equally in any profile of support-relations. The *dictatorship* axiom indicates that there is an agent who is dictatorial.

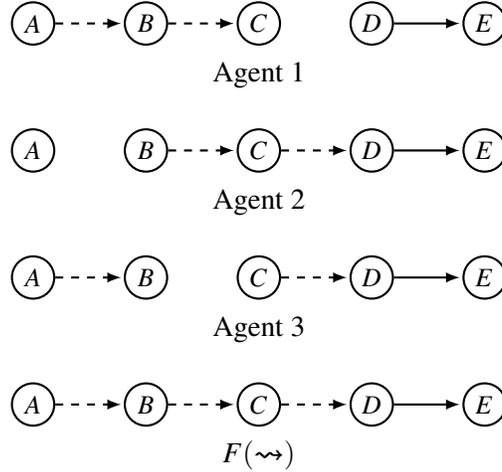


Figure 2: Example for a profile with  $Arg = \{A, B, C, D, E\}$

**Definition 16.** *The unanimity rule is an aggregation rule  $F$  with  $F(\rightsquigarrow) = \{sup \in Arg \times Arg \mid sup \in \rightsquigarrow_1 \cap \dots \cap \rightsquigarrow_n\}$ .*

**Definition 17.** *Let  $i \in N$  be an agent, the dictatorship rule of individual  $i$  is the aggregation rule with  $F(\rightsquigarrow) = \rightsquigarrow_i$ .*

The unanimity rule only accepts those supports approved by all agents: it is a demanding aggregation rule. The dictatorships always return the supports submitted by dictators.

**Example 2.** Suppose that there are three agents have to decide on the acceptance of supports between four arguments. Agent 1 supports  $A \rightsquigarrow B$  and  $B \rightsquigarrow C$ , agent 2 supports  $B \rightsquigarrow C$  and  $C \rightsquigarrow D$ , agent 3 supports  $A \rightsquigarrow B$  and  $C \rightsquigarrow D$ . We assume that the attack relation from  $D$  to  $E$  is accepted by all agents. The scenario is illustrated in Figure 2. If we apply the majority rule, then we obtain a bipolar argumentation framework consisting of the three supports  $A \rightsquigarrow B$ ,  $B \rightsquigarrow C$ , and  $C \rightsquigarrow D$ . We observe that the set  $\{A, E\}$  is conflict-free for all agents. However, it is not conflict-free in the outcome of the majority rule (which returns a set containing only the majoritarian supports) since  $A$  supported attacks  $E$ . So conflict-freeness as a semantic property is not preserved by the majority rule in this specific example.  $\triangle$

But what about the preservation results of other semantic properties? Can they be preserved in general? Before going any further, we introduce more semantic properties of particular interest.

The problem we are considering in this paper is the preservation of semantic properties in the context of bipolar argumentation. Given a property  $P \subseteq 2^{Arg \times Arg}$  that is a set of supports on  $Arg$ , and  $P$  is satisfied by all agents, whether the output of the aggregation rule satisfies  $P$ ? A formal definition is as follows.

**Definition 18.** *An aggregation rule  $F$  preserves a property  $P$  if whenever for every profile  $\rightsquigarrow$  we have that  $P(\rightsquigarrow_i)$  for all  $i \in N$ , then we have  $P(F(\rightsquigarrow))$ .*

The problem of preservation is a special problem of *collective rationality* which has been discussed extensively in other parts of social choice, such as preference aggregation [1], judgment aggregation [16], graph aggregation [13], as well as attack aggregation in the context of abstract argumentation [3, 24, 9].

In the scenario where each agent possesses a BAF, agents might disagree on some details, such as whether a support between two arguments can be justified. Nevertheless, they may agree on some high-level features of BAFs. The *essential constraint* is an example of a high-level feature that requires no

agent accepts both the attack relation and the support relation between a pair of arguments. When we observe that all agents verify such semantic feature, we would like to see what aggregation rule preserves this basic constraint under aggregation.

Given a set of arguments  $\Delta \in Arg$  that is conflict-free in every agent's bipolar argumentation framework, we may wish to preserve its conflict-freeness in the outcome. Therefore, conflict-freeness as a semantic property is of particular interest. Similar definition can be posed to the *preservation of safety and admissibility*. Recall that if a set of arguments  $\Delta$  is conflict-free and closed, then  $\Delta$  is safe (Proposition 1). Thus, the *closedness* is of interest to us as well. Finally, we are also interested in the preservation of semantic extensions. Given a set of arguments  $\Delta \subseteq Arg$  that is an extension of a specific semantics of  $\langle Arg, \rightarrow, \rightsquigarrow_i \rangle$  for all  $i \in N$ , we are interested in under what circumstances  $\Delta$  is an extension of such semantics of  $F(\rightsquigarrow)$  as well. Finally, given an argument that is acceptable under a specific semantics for all agents, we would like to see whether such argument is acceptable in the collective outcome.

## 4 Preservation results

In this section, we present the preservation results for semantic properties. We start with *essential constraint* and *closedness*, two basic requirements of bipolar argumentation frameworks. Then, we turn to consider the preservation of *conflict-freeness*, followed by considering *safety*, followed by considering *d-admissibility*, *s-admissibility*, and *c-admissibility*. Then, we proceed with the study the properties of being an extension, including the property of *being a d-preferred extension*, *being a s-preferred extension*, *being a c-preferred extension* and *being a stable extension*. Finally, we study the preservation of *acceptability of arguments*. Proofs of results in this section can be found in the appendix.

### 4.1 Preservation results for essential constraint, closedness, conflict-freeness, safety and admissibility

Recall that a bipolar AF satisfies essential constraint if it does not contain two arguments for which the first one simultaneously attacks and supports the second one.

**Proposition 2.** *Every aggregation rule  $F$  that is grounded preserves essential constraint.*

The closedness is also an important property. Our result demonstrates that every reasonable rule preserves it.

**Lemma 3.** *Every aggregation rule  $F$  that is grounded preserves closedness.*

For conflict-freeness, we obtain that the unanimity rule, a demanding rule preserves the conflict-freeness of arbitrary sets of arguments.

**Proposition 4.** *The unanimity rule preserves conflict-freeness.*

The preservation of the safety of arbitrary sets of arguments can be accomplished by the unanimity rule.

**Proposition 5.** *The unanimity rule preserves safety.*

*Proof.* This proposition is a consequence of Proposition 4, Lemma 3, and Proposition 1. □

The concepts of d-admissibility and s-admissibility are based on different coherences, but the preservation results for them are similar, as the following proposition demonstrates.

**Proposition 6.** *The unanimity rule preserves either d-admissibility or s-admissibility.*

## 4.2 Preservation results for properties of being an extension

We are going to present preservation results for more demanding properties. Before proceeding, we introduce some necessary terminology and a simple result, as well as a technique developed by Endriss and Grandi for the more general framework of graph aggregation [13]. Let  $sup \in \rightsquigarrow$  be a support, let  $N = \{1, \dots, n\}$  be a finite set of individuals (or agents, we assume that there are two or more agents), and let  $\rightsquigarrow$  be a profile of support-relations. Recall that  $N_{sup}^{\rightsquigarrow}$  is the set of agents who accept  $sup$  under profile  $\rightsquigarrow$ . A *winning coalition*  $\mathcal{W} \subseteq N$  is a set of agents who can decide whether to accept or reject a given support  $sup$ . Given an aggregation rule  $F$ , if  $F$  is *neutral* and *independent*, then  $F$  can be fully determined by a single set  $\mathcal{W}$  of winning coalitions, i.e., for every profile  $\rightsquigarrow$  and every support  $sup$  it is the case that  $sup \in F(\rightsquigarrow) \Leftrightarrow N_{sup}^{\rightsquigarrow} \in \mathcal{W}$ .

In our proofs, we will rely on the concept of *ultrafilter* familiar from set theory [17]. An *ultrafilter* is a collection of subsets of  $N$  satisfying *closure under intersection*, *maximality*, and  $\emptyset \notin \mathcal{W}$ .

**Definition 19.** An ultrafilter  $\mathcal{W}$  on a set  $N$  is a collection of subsets of  $N$  satisfying the following conditions:

- (1)  $\emptyset \notin \mathcal{W}$
- (2) for any pair of sets  $C_1, C_2 \subseteq N$ ,  $C_1, C_2 \in \mathcal{W}$  implies  $C_1 \cap C_2 \in \mathcal{W}$  (closure under intersection)
- (3) for any set  $C$ , one of  $C$  and  $N \setminus C$  is in  $\mathcal{W}$  (maximality)

We restate a simple result, which interprets a well-known fact of ultrafilter in our context:

*Let  $F$  be an independent and neutral aggregation rule and let  $\mathcal{W}$  be the corresponding set of winning coalitions for supports, i.e.,  $sup \in F(\rightsquigarrow) \Leftrightarrow N_{sup}^{\rightsquigarrow} \in \mathcal{W}$  for all  $sup \in \rightsquigarrow$ . Then,  $F$  is dictatorial if and only if  $\mathcal{W}$  is an ultrafilter.*

Besides the properties identified in Section 3, we introduce two meta-properties:

**Definition 20.** A property  $P$  is called **non-simple** if there exists a set  $Sup \subseteq Arg \times Arg$  of supports and three individual supports  $sup_1, sup_2, sup_3 \in Arg \times Arg \setminus Sup$  such that  $\langle Arg, \rightarrow, Sup \cup S \rangle$  with  $S \subseteq \{sup_1, sup_2, sup_3\}$  satisfies  $P$  if and only if  $S \neq \{sup_1, sup_2, sup_3\}$ .

**Definition 21.** A property  $P$  is called **disjunctive** if there exists a set  $Sup \subseteq Arg \times Arg$  of supports and two individual supports  $sup_1, sup_2 \in Arg \times Arg \setminus Sup$  such that  $\langle Arg, \rightarrow, Sup \cup S \rangle$  with  $S \subseteq \{sup_1, sup_2\}$  satisfies  $P$  if and only if  $S \neq \emptyset$ .

Non-simplicity requires that, in the context of  $Sup$ , accepting any proper subset of  $\{sup_1, sup_2, sup_3\}$  is possible, while accepting  $\{sup_1, sup_2, sup_3\}$  is not. Disjunctiveness requires that, in the context of  $Sup$ , we should accept at least one of  $sup_1$  and  $sup_2$ . The term of *simplicity* was introduced by Nehring and Puppe [18] as the *median property*; the term of *disjunctiveness* was introduced by Endriss and Grandi [13] as a graph meta-property. It is worth noting that a meta-property is a class of properties, a property satisfies or does not satisfy a specific meta-property. Even though the definitions of meta-properties are not complicated, deciding whether a given property belongs to a meta-property is still not straightforward.

Meta-properties have connections with properties of BAFs and aggregation rules: on the one hand, meta-properties outline high-level features of properties of BAFs, with which we are able to systematically study the preservation of semantic properties of BAFs with simple proofs; on the other hand, as we will see in the following lemmas, meta-properties allow us to generalize specific result for specific properties by instantiating the general results. To be more specific, if an aggregation rule preserves a property that belongs to a meta-property, then it is a dictatorship.

**Lemma 7.** *Let  $P$  be a property that is non-simple and disjunctive. Then, for  $|Arg| \geq 3$ , any unanimous, grounded, neutral, and independent aggregation rule  $F$  that preserves  $P$  must be a dictatorship.*

If a property we are interested in is non-simple and disjunctive, then we can apply Lemma 7 to obtain an axiomatic result for it.

**Theorem 8.** *For  $|Arg| \geq 5$ , any unanimous, grounded, neutral, and independent aggregation rule  $F$  that preserves either  $d$ -preferred,  $s$ -preferred, or  $c$ -preferred extensions must be a dictatorship.*

For the scenarios when  $|Arg| = 4$ , or even  $|Arg| = 3$ , we are not able to verify whether the theorem can still apply, we conjecture that the bound on the cardinality of  $Arg$  is sharp and we believe that the theorem has covered all cases of practical interest. By comparison, we recall that for the property of being a preferred extension of Dung's argumentation framework, Chen and Endriss have shown that only dictatorships preserve it [9]. They have made the assumption that every agent is equipped with a different set of attack relations while they hold the same set of arguments.

Note that Theorem 8 is an impossibility result that indicates the preservation of specific properties is impossible, unless the aggregation rule under consideration is dictatorial. They relate to generalisations of Arrow's Impossibility Theorem [1] to graph aggregation and attack-relation aggregation. One of the features of the aggregation rules we used in this section is that they accept the axiom of independence. Even though this axiom is attractive in some sense, to escape the impossibilities, a prime direction is to relax it. For example, we can consider distance-based rules and investigate whether we are able to obtain some positive results.

Theorem 8 has assumed that the interpretation of support is deductive support. Even though it is enough for our purposes, namely it is enough to show that only dictatorships preserve  $d$ -preferred ( $s$ -preferred,  $c$ -preferred) extensions, we are still interested in what happens when the interpretation is restricted to necessary support. The bad news is, we still cannot overcome impossibility results.

**Theorem 9.** *If the interpretation of support is necessary support, for  $|Arg| \geq 5$ , any unanimous, grounded, neutral, and independent aggregation rule  $F$  that preserves either  $d$ -preferred,  $s$ -preferred, or  $c$ -preferred extensions must be a dictatorship.*

#### 4.2.1 Preservation result for stable extensions

For stable extensions, by using the same techniques, we obtain a similar impossibility result.

**Theorem 10.** *For  $|Arg| \geq 5$ , any unanimous, grounded, neutral, and independent aggregation rule  $F$  that preserves stable extensions must be a dictatorship.*

By comparison, we recall that the nomination rule preserves stable extensions of Dung's argumentation frameworks [9].

### 4.3 Preservation of argument acceptability

Now, we move to study the preservation of acceptability of arguments. Before proceeding further, it is thus important to keep in mind that our model has assumed that each agent  $i \in N$  reports a set of supports  $\rightsquigarrow_i$  on the same set of arguments and attack relations. Given two BAFs  $\rightsquigarrow_1$  and  $\rightsquigarrow_2$ , if  $\rightsquigarrow_1 \supseteq \rightsquigarrow_2$ , namely the supports of  $\rightsquigarrow_1$  is a superset of  $\rightsquigarrow_2$ , then a  $d$ -admissible ( $s$ -admissible,  $c$ -admissible, respectively) set of  $\rightsquigarrow_1$  remains  $d$ -admissible ( $s$ -admissible,  $c$ -admissible, respectively) in  $\rightsquigarrow_2$ .

**Lemma 11.** *Given two BAFs  $\rightsquigarrow_1$  and  $\rightsquigarrow_2$ , if  $\rightsquigarrow_1 \supseteq \rightsquigarrow_2$ , then every  $d$ -admissible set of arguments of  $\rightsquigarrow_1$  is a  $d$ -admissible set of  $\rightsquigarrow_2$ .*

**Lemma 12.** *Given two BAFs  $\rightsquigarrow_1$  and  $\rightsquigarrow_2$ , if  $\rightsquigarrow_1 \supseteq \rightsquigarrow_2$ , then every s-admissible set of arguments of  $\rightsquigarrow_1$  is a s-admissible set of  $\rightsquigarrow_2$ .*

**Lemma 13.** *Given two BAFs  $\rightsquigarrow_1$  and  $\rightsquigarrow_2$ , if  $\rightsquigarrow_1 \supseteq \rightsquigarrow_2$ , then every c-admissible set of arguments of  $\rightsquigarrow_1$  is a c-admissible set of  $\rightsquigarrow_2$ .*

**Fact 14.** *Given a BAF  $\rightsquigarrow$ , if  $\Delta \subseteq \text{Arg}$  is a d-preferred (s-preferred, c-preferred, respectively) extension of  $\rightsquigarrow$ , then  $\Delta$  is a d-admissible (s-admissible, c-admissible, respectively) set of arguments of  $\rightsquigarrow$ .*

Thus, every d-admissible (s-admissible, c-admissible, respectively) set of arguments is included in a d-preferred (s-preferred, c-preferred, respectively) extension. With this, we are ready to present the preservation results for argument acceptability under preferred semantics, including d-preferred semantics, s-preferred semantics, and c-preferred semantics. Note that in the following we say that an argument under a d-preferred extension, we mean that such argument is a member of a d-preferred extension.

**Proposition 15.** *The unanimity rule preserves the property of argument acceptability under d-preferred semantics.*

**Proposition 16.** *The unanimity rule preserves the property of argument acceptability under either s-preferred or c-preferred semantics.*

The proof is similar to the proof for Theorem 15. The only difference is that every s-admissible (c-admissible) set of  $\rightsquigarrow_i$  is a s-admissible set of  $F(\rightsquigarrow)$  is because of Lemma 12 (Lemma 13).

## 5 Related work

Previous work on obtaining argumentative consensus among a group of agents are mainly focus on abstract argumentation frameworks [10, 25, 9]. Among them, Chen and Endriss [9] study of the preservation of semantic properties during the aggregation of attack-relations of abstract argumentation frameworks. As a potential domain of application for the model they develop, they do not make explicit reference to bipolar argumentation frameworks. In addition, similar to us, they have made use of meta-properties proposed by Endriss and Grandi for graph aggregation [13], which serve as technical devices to obtain preservation results for semantic properties.

The problem of aggregating bipolar opinions has received interests in the literature. The idea of aggregating support-relations of bipolar argumentation frameworks has been outlined in a preliminary version of this paper [8]. Lauren *et al.* [15] consider aggregating bipolar assumption-based argumentation frameworks under the assumption that agents propose the same set of arguments, but propose different sets of attacks and supports. Their focus is quota rules and oligarchic rules. Kontarinis *et al.* [14] study designing mechanisms for “regulating” debates under the setting of each agent in the debate equipped with a bipolar argumentation framework. We note that in their settings, agents report the same set of arguments, but with possibly different attack- and support-relations.

## 6 Conclusion

In this paper, we have studied the aggregation of agents’ view in the context of bipolar argumentation. To be more specific, we have explored the problem of aggregating support-relations of bipolar argumentation frameworks by making use of the methodology of social choice theory. To achieve this, we have designed a model, in which agents are equipped with a set of arguments and a set of attacks, but with possibly different support-relations. We have shown which semantic properties of BAFs can be preserved by

aggregation rules. We have proposed two BAF meta-properties, namely the property of “non-simplicity” and “disjunctiveness”, both of which are high level features of BAFs. We show that such meta-properties can be used to obtain impossibility results, namely for quickly proving what kind of aggregation rules (or no desirable aggregation rule) is collectively rational with respect to BAF properties.

For future work, it is worth having an investigation of further meta-properties. We point out that Lemma 7 is a variant of Theorem 16 in graph aggregation [13]. This indicates that there is space for other variants of impossibility results for different assumptions. The preservation of some desirable properties (for example, the property of being a d-preferred extension) during the aggregation of support-relations is difficult. Thus, it is worth studying whether such properties can be preserved in different settings. For instance, agents might be equipped with the same set of arguments, but with possibly different attack- and support-relations, and we aggregate attack- and support-relations by making use of different quota rules. Finally, recall that there are at least three interpretations of support, in this paper, we focus on the deductive support and necessary support. More interpretations of support, for example, evidential support, should be investigated.

**Acknowledgements.** I would like to thank Ulle Endriss for his generous suggestions on an earlier version and his guidance at the beginning of this work. I also thank three anonymous reviewers for their constructive feedback. This work was supported by the China Postdoctoral Science Foundation (grant no. 2019M663352).

## References

- [1] Kenneth J. Arrow (1963): *Social Choice and Individual Values*, 2nd edition. John Wiley and Sons. First edition published in 1951.
- [2] Philippe Besnard & Anthony Hunter (2008): *Elements of Argumentation*. MIT Press, doi:10.7551/mitpress/9780262026437.001.0001.
- [3] Gustavo M. Bodanza, Fernando A. Tohmé & Marcelo R. Auday (2017): *Collective Argumentation: A Survey of Aggregation Issues around Argumentation Frameworks*. *Argument & Computation* 8(1), pp. 1–34, doi:10.3233/AAC-160014.
- [4] Guido Boella, Dov Gabbay, Leon van der Torre & Serena Villata (2010): *Support in abstract argumentation*. In: *Proceedings of the Third International Conference on Computational Models of Argument (COMMA-10)*, Frontiers in Artificial Intelligence and Applications, IOS Press, pp. 40–51, doi:10.3233/978-1-60750-619-5-111.
- [5] Claudette Cayrol & Marie-Christine Lagasquie-Schiex (2005): *Gradual valuation for bipolar argumentation frameworks*. In: *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, Springer, pp. 366–377, doi:10.1007/11518655\_32.
- [6] Claudette Cayrol & Marie-Christine Lagasquie-Schiex (2005): *On the acceptability of arguments in bipolar argumentation frameworks*. In: *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, Springer, pp. 378–389, doi:10.1007/11518655\_33.
- [7] Claudette Cayrol & Marie-Christine Lagasquie-Schiex (2013): *Bipolarity in argumentation graphs: Towards a better understanding*. *International Journal of Approximate Reasoning* 54(7), pp. 876–899, doi:10.1016/j.ijar.2013.03.001.

- [8] Weiwei Chen (2020): *Aggregation of Support-Relations of Bipolar Argumentation Frameworks (Extended Abstract)*. In: *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-2020)*, IFAAMAS.
- [9] Weiwei Chen & Ulle Endriss (2019): *Preservation of Semantic Properties in Collective Argumentation: The Case of Aggregating Abstract Argumentation Frameworks*. *Artificial Intelligence* 269, pp. 27–48, doi:10.1016/j.artint.2018.10.003.
- [10] Sylvie Coste-Marquis, Caroline Devred, Sébastien Konieczny, Marie-Christine Lagasquie-Schiex & Pierre Marquis (2007): *On the Merging of Dung's Argumentation Systems*. *Artificial Intelligence* 171(10–15), pp. 730–753, doi:10.1016/j.artint.2007.04.012.
- [11] Phan Minh Dung (1995): *On the Acceptability of Arguments and its Fundamental Role in Non-monotonic Reasoning, Logic Programming and n-Person Games*. *Artificial Intelligence* 77(2), pp. 321–358, doi:10.1016/0004-3702(94)00041-X.
- [12] Paul E. Dunne, Pierre Marquis & Michael Wooldridge (2012): *Argument Aggregation: Basic Axioms and Complexity Results*. In: *Proceedings of the 4th International Conference on Computational Models of Argument (COMMA)*, IOS Press, doi:10.3233/978-1-61499-111-3-129.
- [13] Ulle Endriss & Umberto Grandi (2017): *Graph Aggregation*. *Artificial Intelligence* 245, pp. 86–114, doi:10.1016/j.artint.2017.01.001.
- [14] Dionysios Kontarinis, Elise Bonzon, Nicolas Maudet & Pavlos Moraitis (2011): *Regulating Multiparty Persuasion with Bipolar Arguments: Discussion and Examples*. In: *Modèles Formels de l'Interaction*, pp. 119–129.
- [15] Stefan Lauren, Francesco Belardinelli & Francesca Toni (2021): *Aggregating Bipolar Opinions*. In: *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.
- [16] Christian List & Philip Pettit (2002): *Aggregating Sets of Judgments: An Impossibility Result*. *Economics and Philosophy* 18(1), pp. 89–110, doi:10.1017/S0266267102001098.
- [17] Bernard Monjardet (1983): *On the use of ultrafilters in social choice theory*. *Social choice and welfare* 5, pp. 73–78, doi:10.1016/B978-0-444-86487-1.50012-5.
- [18] Klaus Nehring & Clemens Puppe (2007): *The structure of strategy-proof social choice—Part I: General characterization and possibility results on median spaces*. *Journal of Economic Theory* 135(1), pp. 269–305, doi:10.1016/j.jet.2006.04.008.
- [19] Farid Nouioua & Vincent Risch (2010): *Bipolar argumentation frameworks with specialized supports*. In: *2010 22nd IEEE International Conference on Tools with Artificial Intelligence*, 1, IEEE, pp. 215–218, doi:10.1109/ICTAI.2010.37.
- [20] Farid Nouioua & Vincent Risch (2011): *Argumentation frameworks with necessities*. In: *Proceedings of the International Conference on Scalable Uncertainty Management*, Springer, pp. 163–176, doi:10.1007/978-3-642-23963-2\_14.
- [21] Nir Oren, Michael Luck & Chris Reed (2010): *Moving between argumentation frameworks*. In: *Proceedings of the 2010 International Conference on Computational Models of Argument (COMMA-2010)*, IOS Press, doi:10.3233/978-1-60750-619-5-379.
- [22] Nir Oren & Timothy J. Norman (2008): *Semantics for evidence-based argumentation*. *Proceedings of the 2008 International Conference on Computational Models of Argument (COMMA-2008)* 172, p. 276.

- [23] Iyad Rahwan & Guillermo R. Simari (2009): *Argumentation in Artificial Intelligence*. Springer-Verlag, doi:10.1007/978-0-387-98197-0.
- [24] Iyad Rahwan & Fernando A. Tohmé (2010): *Collective Argument Evaluation as Judgement Aggregation*. In: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.
- [25] Fernando A. Tohmé, Gustavo A. Bodanza & Guillermo R. Simari (2008): *Aggregation of Attack Relations: A Social-choice Theoretical Analysis of Defeasibility Criteria*. In: *Proceedings of the 5th International Symposium on Foundations of Information and Knowledge Systems (FoIKS)*, Springer-Verlag, doi:10.1007/978-3-540-77684-0\_4.
- [26] Kristijonas Čyras, Claudia Schulz & Francesca Toni (2017): *Capturing Bipolar Argumentation in Non-flat Assumption-Based Argumentation*. In: *International Conference on Principles and Practice of Multi-Agent Systems*, pp. 386–402, doi:10.1007/978-3-319-69131-2\_23.

## Appendix: Remaining Proofs

In this appendix we present the proofs omitted from the body of the paper.

### Proof of Proposition 2

Let  $\rightsquigarrow = (\rightsquigarrow_1, \dots, \rightsquigarrow_n)$  be a profile of BAFs, in which  $\rightsquigarrow_i$  satisfies the essential constraint for all  $i \in N$ . Let  $F$  be an aggregation rule that is grounded. For the sake of contradiction, we suppose that the essential constraint is violated in  $F(\rightsquigarrow)$ . Without loss of generality, we suppose that both  $A \rightarrow B$  and  $A \rightsquigarrow B$  get accepted in  $F(\rightsquigarrow)$ . From the assumption, we know that every agent agrees  $A \rightarrow B$ . In the meantime, at least one agent accepts  $A \rightsquigarrow B$  under grounded aggregation rules, which cannot be the case. Thus, we have the proposition.  $\square$

### Proof of Lemma 3

We again let  $\rightsquigarrow = (\rightsquigarrow_1, \dots, \rightsquigarrow_n)$  be a profile of BAFs, let  $\Delta \subseteq Arg$  be the set of arguments under consideration, and let  $F$  be an aggregation rule that is grounded. For the sake of contradiction, we suppose that  $\Delta$  is closed in  $\rightsquigarrow_i$  for all  $i \in N$  and not closed in  $F(\rightsquigarrow)$ , i.e., there is an argument  $A \in \Delta$  and an argument  $B \in Arg \setminus \Delta$  such that  $A \rightsquigarrow B$  in  $F(\rightsquigarrow)$ . As rules under considering are grounded, there is at least one agent  $\rightsquigarrow_i$  for which  $A \rightsquigarrow_i B$  is the case. This will lead to the situation that  $\Delta$  not being closed in  $\rightsquigarrow_i$ , contradicting our assumption.  $\square$

### Proof of Proposition 4

Recall that the unanimity rule is the quota rule  $F$  with the quota of  $n$ . Let  $\rightsquigarrow = (\rightsquigarrow_1, \dots, \rightsquigarrow_n)$  be a profile of BAFs. Let  $\Delta \subseteq Arg$  be the set of arguments under consideration. For the sake of contradiction, we suppose that  $\Delta$  is conflict-free in  $\rightsquigarrow_i$  for all  $i \in N$ , and is not conflict-free in  $F(\rightsquigarrow)$ . This means that there are two arguments  $A, B \in \Delta$  such that  $A$  supported or secondary attacks  $B$  in  $F(\rightsquigarrow)$ .

We now show that in the scenario where  $A$  is supported attacking  $B$ , i.e., there is a sequence of arguments in  $F(\rightsquigarrow)$  such that  $A_1 \rightsquigarrow, \dots, \rightsquigarrow A_m, A_m \rightarrow B$  in which  $A_1 = A$ , our proposition holds. From the assumption we know that  $(A_m \rightarrow B) \in \rightarrow_i$  for all  $i \in N$ . In addition, every agent agrees  $A_1 \rightsquigarrow, \dots, \rightsquigarrow A_m$ .

Thus, every agent agrees  $A_1 \rightsquigarrow, \dots, \rightsquigarrow A_m, A_m \rightarrow B$ , i.e.,  $\Delta$  is not conflict-free in  $\rightsquigarrow_i$  for all  $i \in N$ , in contradiction to our earlier assumption.

For the scenario where  $A$  is secondary attacking  $B$ , we note that the proof is similar to the proof of the one where  $A$  is supported attacking  $B$ . Thus, we have the proposition.  $\square$

### Proof of Proposition 6

Let  $F$  be the unanimity rule. Let  $\rightsquigarrow = (\rightsquigarrow_1, \dots, \rightsquigarrow_n)$  be a profile of bipolar argumentation frameworks. Let  $\Delta \subseteq \text{Arg}$  be the set of arguments under consideration.

We suppose that  $\Delta$  is d-admissible in  $\rightsquigarrow_i$  for all  $i \in N$ . Then,  $\Delta$  is conflict-free in  $\rightsquigarrow_i$  for all  $i \in N$  as well. By Proposition 4,  $\Delta$  is conflict-free in  $F(\rightsquigarrow)$ . By the assumption that all agents report the same set of attacks, we get that for each argument  $A \in \Delta$ ,  $A$  is defended by  $\Delta$  in  $\rightsquigarrow_i$  for all  $i \in N$ . It follows that  $A$  is defended by  $\Delta$  in  $F(\rightsquigarrow)$  as well. Thus,  $\Delta$  is d-admissible in  $F(\rightsquigarrow)$ .

We omit the relative easy proof for s-admissibility.  $\square$

### Proof of Lemma 7

Take any property  $P$  that is non-simple and disjunctive. Take any aggregation rule  $F$  that is unanimous, grounded, neutral, independent and preserves  $P$ . By the assumption that  $F$  is neutral and independent,  $F$  is determined by a set of winning coalitions  $\mathcal{W} \subseteq 2^N$ . What we need to do is proving that  $\mathcal{W}$  is an ultrafilter, i.e., to show that  $\mathcal{W}$  is closed under intersection,  $\mathcal{W}$  satisfies maximality, and  $\emptyset \notin \mathcal{W}$ .

$\emptyset \notin \mathcal{W}$  This is a direct consequence of the assumption that  $F$  is grounded.

**Maximality** Take any set of agents  $C \subseteq N$ . Consider a profile in which exactly the individuals in  $C$  propose  $sup_1$  and exactly those in  $N \setminus C$  propose  $sup_2$ . Since  $P$  is disjunctive, we know that one of  $sup_1$  and  $sup_2$  must be part of  $F(\rightsquigarrow)$ . Hence  $C \in \mathcal{W}$  or  $N \setminus C \in \mathcal{W}$ .

**Closure under intersection** Take any two winning coalitions  $C_1, C_2 \in \mathcal{W}$ . Assume toward a contradiction that  $C_1 \cap C_2 \notin \mathcal{W}$ . Consider a profile in which exactly the individuals in  $C_1$  propose  $sup_1$ , exactly the individuals in  $C_2$  propose  $sup_2$ , and exactly the individuals in  $N \setminus (C_1 \cap C_2)$  propose  $sup_3$ . Now, since  $C_1$  and  $C_2$  are winning coalitions,  $sup_1$  and  $sup_2$  must be part of  $F(\rightsquigarrow)$ . Hence, due to  $C_1 \cap C_2 \notin \mathcal{W}$  and  $\mathcal{W}$  satisfying maximality, we have  $N \setminus (C_1 \cap C_2) \in \mathcal{W}$ . Since the individuals in  $N \setminus (C_1 \cap C_2)$  propose  $sup_3$ , we have  $sup_3 \in F(\rightsquigarrow)$ . But we have assumed that  $F$  preserves non-simplicity, i.e.,  $sup_1, sup_2, sup_3$  cannot be accepted together in  $F(\rightsquigarrow)$ . Thus,  $C_1 \cap C_2 \in \mathcal{W}$ .  $\square$

### Proof of Theorem 8

Suppose  $|\text{Arg}| \geq 5$ . Let  $P$  be the properties representing a given set of arguments being either a d-preferred, a s-preferred or a c-preferred extension. Thus, by Lemma 7, we need to show that  $P$  is non-simple and disjunctive in this case.

**Non-simplicity** Let  $\text{Arg} = \{A, B, C, D, E, \dots\}$ , let  $\rightarrow = \{D \rightarrow E, E \rightarrow D, B \rightarrow B, C \rightarrow C\}$ . Now we focus on  $\text{Arg} \setminus \{B, C, D\}$  as the subset of arguments that may (or may not) form either a d-preferred, a s-preferred or a c-preferred extension. We define  $Sup = \emptyset$ ,  $sup_1 = (A \rightsquigarrow B)$ ,  $sup_2 = (B \rightsquigarrow C)$ , and  $sup_3 = (C \rightsquigarrow D)$ . This scenario is depicted in the top part of Figure 3. Consider all BAFs of the form

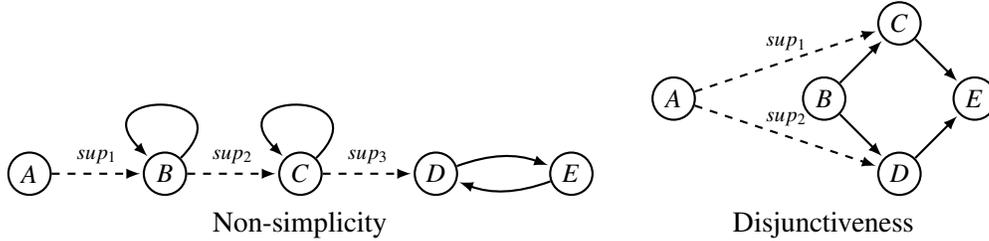


Figure 3: Scenarios used in the proof of Theorem 8

$BAF = \langle Arg, \rightarrow, Sup \cup S \rangle$  with  $S \subseteq \{sup_1, sup_2, sup_3\}$ . It is not difficult for the reader to verify that, for  $S \neq \{sup_1, sup_2, sup_3\}$ ,  $B$  and  $C$  are self-attacking,  $D$  is attacked by  $E$ . Thus, they are unacceptable with respect to  $\{A, E\}$ , i.e.,  $Arg \setminus \{B, C, D\}$  is a d-preferred, a s-preferred and a c-preferred extension. On the other hand, for  $S = \{sup_1, sup_2, sup_3\}$ ,  $Arg \setminus \{B, C, D\}$  is not a d-preferred nor a s-preferred or c-preferred extension as  $E$  is set-attacked by  $A$ . Thus,  $P$  is non-simple.

**Disjunctiveness** Let  $Arg = \{A, B, C, D, E, \dots\}$ , let  $\rightarrow = \{B \rightarrow C, B \rightarrow D, C \rightarrow E, D \rightarrow E\}$ . We focus on  $Arg \setminus \{C, D, E\}$  as the subset of arguments that may (or may not) form a s-preferred extension. We define  $Sup = \emptyset$ ,  $sup_1 = (A \rightsquigarrow C)$ ,  $sup_2 = (A \rightsquigarrow D)$ . This scenario is depicted in the bottom part of Figure 3. Consider all BAFs of the form  $BAF = \langle Arg, \rightarrow, Sup \cup S \rangle$  with  $S \subseteq \{sup_1, sup_2\}$ . For  $S \neq \emptyset$ ,  $Arg \setminus \{C, D, E\}$  is a d-preferred, a s-preferred, and a c-preferred extension. On the other hand, for  $S = \emptyset$ ,  $Arg \setminus \{C, D, E\}$  is not a s-preferred nor a d-preferred or a c-preferred extension as  $E$  is defended by  $B$ . Thus,  $P$  is disjunctive.  $\square$

### Proof of Theorem 9

Similar to Theorem 8, we still need to show that the property of being a d-preferred, a s-preferred, or a c-preferred extension is non-simple and disjunctive when the interpretation of support is necessary support.

**Non-simplicity** Let  $Arg = \{A, B, C, D, E, \dots\}$ , let  $\rightarrow = \{D \rightarrow E, E \rightarrow D, B \rightarrow B, C \rightarrow C\}$ . Now we focus on  $Arg \setminus \{B, C, D\}$  as the subset of arguments that may (or may not) form either a d-preferred, a s-preferred or a c-preferred extension. We define  $Sup = \emptyset$ ,  $sup_1 = (B \rightsquigarrow A)$ ,  $sup_2 = (C \rightsquigarrow B)$ , and  $sup_3 = (D \rightsquigarrow C)$ , as illustrated in the top part of Figure 4. Consider all BAFs of the form  $BAF = \langle Arg, \rightarrow, Sup \cup S \rangle$  with  $S \subseteq \{sup_1, sup_2, sup_3\}$ . It is not difficult for the reader to verify that, for  $S \neq \{sup_1, sup_2, sup_3\}$ ,  $B$  and  $C$  are self-attacking,  $D$  is attacked by  $E$ . Thus, they are unacceptable with respect to  $\{A, E\}$ . In the meantime,  $A$  is not attacked by any other argument,  $E$  defends itself,  $\{A, E\}$  is conflict-free, i.e.,  $Arg \setminus \{B, C, D\}$  is a d-preferred, a s-preferred, and a c-preferred extension. On the other hand, for  $S = \{sup_1, sup_2, sup_3\}$ ,  $Arg \setminus \{B, C, D\}$  is neither a d-preferred nor a s-preferred nor c-preferred extension as  $E$  secondary attacks  $A$ . Thus,  $P$  is non-simple.

**Disjunctiveness** Let  $Arg = \{A, B, C, D, \dots\}$ , let  $\rightarrow = \{B \rightarrow C, B \rightarrow D\}$ . We focus on  $Arg \setminus \{A, C, D\}$  as the subset of arguments that may (or may not) form a s-preferred extension. We define  $Sup = \emptyset$ ,  $sup_1 = (C \rightsquigarrow A)$ ,  $sup_2 = (D \rightsquigarrow A)$ , as illustrated in the bottom part of Figure 4. Consider all BAFs of the form  $BAF = \langle Arg, \rightarrow, Sup \cup S \rangle$  with  $S \subseteq \{sup_1, sup_2\}$ . For  $S \neq \emptyset$ ,  $B$  secondary attacks  $A$  and directly

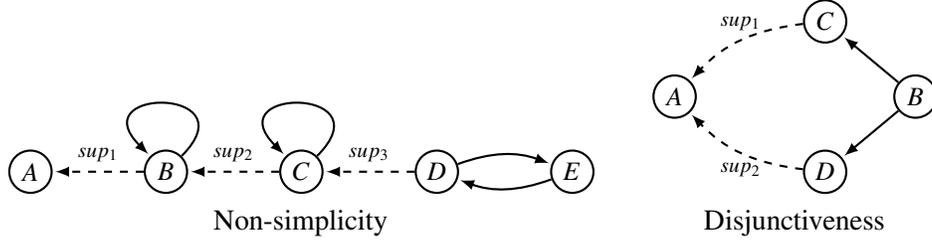


Figure 4: Scenarios used in Theorem 9.

attacks  $C$  and  $D$ . Thus,  $Arg \setminus \{A, C, D\}$  is a d-preferred, a s-preferred, and a c-preferred extension. On the other hand, for  $S = \emptyset$ ,  $Arg \setminus \{A, C, D\}$  is neither a s-preferred nor a d-preferred nor a c-preferred extension as  $A$  is not attacked by any other argument and thus should be included in every d-preferred (s-preferred, c-preferred) extension. Thus,  $P$  is disjunctive.  $\square$

### Proof of Theorem 10

Suppose  $|Arg| \geq 5$ . Let  $P$  be the BAF-properties representing a given set of arguments being a stable extension. We need to demonstrate that  $P$  is non-simple and disjunctive in this case.

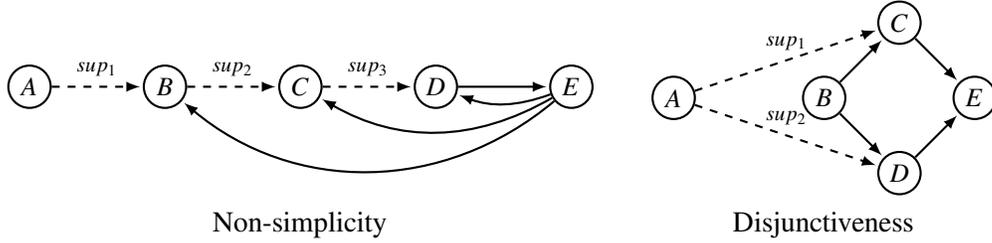


Figure 5: Scenarios used in the proof of Theorem 10

**Non-simplicity** Let  $Arg = \{A, B, C, D, E, \dots\}$ , let  $\rightarrow = \{D \rightarrow E, E \rightarrow B, E \rightarrow C, E \rightarrow D\}$ . We focus on  $Arg \setminus \{B, C, D\}$  as the subset of arguments that may (or may not) form a stable extension. We define  $Sup = \emptyset$ .  $sup_1 = (A \rightsquigarrow B)$ ,  $sup_2 = (B \rightsquigarrow C)$ , and  $sup_3 = (C \rightsquigarrow D)$ . This scenario is depicted in the top part of Figure 5. Consider all BAFs of the form  $BAF = \langle Arg, \rightarrow, Sup \cup S \rangle$  with  $S \subseteq \{sup_1, sup_2, sup_3\}$ . The reader should be able to verify that, indeed, for  $S \neq \{sup_1, sup_2, sup_3\}$ ,  $Arg \setminus \{B, C, D\}$  is a stable extension. For example, for  $S = \{sup_1, sup_2\}$ ,  $B$ ,  $C$ , and  $D$  are attacked by  $E$ . In the meantime,  $\{A, E\}$  is conflict-free, i.e.,  $Arg \setminus \{B, C, D\}$  is a stable extension. On the other hand, for  $S = \{sup_1, sup_2, sup_3\}$ ,  $Arg \setminus \{B, C, D\}$  is not a stable extension as  $E$  is set-attacked by  $A$ . Thus,  $P$  is non-simple.

**Disjunctiveness** Let  $Arg = \{A, B, C, D, E, \dots\}$ , let  $\rightarrow = \{B \rightarrow C, B \rightarrow D, C \rightarrow E, D \rightarrow E\}$ . We focus on  $Arg \setminus \{C, D, E\}$  as the subset of arguments that may (or may not) form a stable extension. We define  $Sup = \emptyset$ ,  $sup_1 = (A \rightsquigarrow C)$ ,  $sup_2 = (A \rightsquigarrow D)$ . This scenario is depicted in the bottom part of Figure 5. Consider all BAFs of the form  $BAF = \langle Arg, \rightarrow, Sup \cup S \rangle$  with  $S \subseteq \{sup_1, sup_2\}$ . The reader should be able to verify that, indeed, for  $S \neq \emptyset$ ,  $Arg \setminus \{C, D, E\}$  is a stable extension. On the other hand, for  $S = \emptyset$ ,  $Arg \setminus \{C, D, E\}$  is not a stable extension as  $E$  is defended by  $B$ . Thus,  $P$  is disjunctive.  $\square$

### Proof of Lemma 11

Let  $\Delta \subseteq Arg$  be a d-admissible set of  $\rightsquigarrow_1$ . We need to show that  $\Delta$  is a d-admissible set of  $\rightsquigarrow_2$ . To achieve this, we need to demonstrate that in  $\rightsquigarrow_2$ , (i)  $\Delta$  is conflict-free, and (ii)  $\Delta$  defends all of its members.

For (i), we need to show that  $\Delta$  is conflict-free in  $\rightsquigarrow_2$ . If not, then there are two arguments  $A, B \in \Delta$  such  $A$  directly, supported, or secondary attacks  $B$ . If  $A$  directly attacks  $B$  in  $\rightsquigarrow_2$ , then  $A$  directly attacks  $B$  in  $\rightsquigarrow_1$  as the pair of BAFs report the same set of attacks, which contradicts the assumption that  $\Delta$  is conflict-free in  $\rightsquigarrow_1$ . If  $A$  supported attacks  $B$  in  $\rightsquigarrow_2$ , then there is a sequence of argument  $(A_1, \dots, A_n)$  such that  $A_1 \rightsquigarrow A_2, \dots, A_{n-1} \rightarrow A_n, A = A_1$ , and  $A_n = B$ . As  $\rightsquigarrow_1 \supseteq \rightsquigarrow_2$  and  $A_{n-1} \rightarrow A_n$  is the case, we know that  $A_1 \rightsquigarrow A_2, \dots, A_{n-1} \rightarrow A_n$  in  $\rightsquigarrow_2$  as well, which means that there are two arguments  $A, B \in \Delta$  such that  $A$  supported attacks  $B$ , contradicting the fact that  $\Delta$  is conflict-free in  $\rightsquigarrow_1$ . If  $A$  supported attacks  $B$  in  $\rightsquigarrow_2$ , this case is similar to the case that  $A$  supported attacks  $B$ , which will lead to that  $\Delta$  failing to satisfy conflict-freeness in  $\rightsquigarrow_1$ . Thus,  $\Delta$  is conflict-free in  $\rightsquigarrow_2$ .

For (ii), we need to show that for every argument  $A \in \Delta$ , if  $B \rightarrow A$ , then there is a  $C \in \Delta$  such that  $C \rightarrow B$ , i.e.,  $\Delta$  defends all its members in  $\rightsquigarrow_2$ . Clearly, this is true as  $\rightsquigarrow_1$  and  $\rightsquigarrow_2$  report the same set of attacks, and  $\Delta$  defends all its members in  $\rightsquigarrow_1$ .  $\square$

### Proof of Lemma 12

We need to show that a s-admissible set of arguments  $\Delta \subseteq Arg$  of  $\rightsquigarrow_1$  is a s-admissible set of  $\rightsquigarrow_2$ . To arrive at this goal, we need to demonstrate that in  $\rightsquigarrow_2$ , (i)  $\Delta$  is conflict-free, (ii)  $\Delta$  defends all of its members, and (iii)  $\Delta$  is safe. For (i) and (ii), the proofs are the same as the ones in Lemma 11. It remains to show that  $\Delta$  is safe in  $\rightsquigarrow_2$ . If not, then there is an argument  $B \in Arg$  such that  $\Delta$  set-attacks  $B$  and  $\Delta$  set-supports  $B$ , or  $B \in \Delta$ . If  $\Delta$  set-supports  $B$ , there are two arguments  $A \in \Delta$  such  $A$  directly, supported, or secondary attacks  $B$ . Using the construction similar to Lemma 11, it is easy to verify that under this assumption,  $\Delta$  set-attacks  $B$  in  $\rightsquigarrow_1$ . According to the assumption that  $\rightsquigarrow_1 \supseteq \rightsquigarrow_2$ ,  $\Delta$  set-supports  $B$  in  $\rightsquigarrow_1$ . Then,  $\Delta$  is not safe in  $\rightsquigarrow_1$ , contradiction.  $\square$

### Proof of Lemma 13

Once again, we need to show that every c-admissible set  $\Delta$  of arguments of  $\rightsquigarrow_1$  is a c-admissible set of  $\rightsquigarrow_2$ . To arrive at this goal, we need to show that in  $\rightsquigarrow_2$  (i),  $\Delta$  is conflict-free, (ii)  $\Delta$  defends all of its members, and (iii)  $\Delta$  is closed. For (i) and (ii), the proofs are the same as the ones in Lemma 11. It remains to show that  $\Delta$  is closed in  $\rightsquigarrow_2$ . If not, then there is an argument  $A \in \Delta$  and an argument  $B \in Arg$  such that  $A$  supports  $B$ , and  $B \notin \Delta$ . According to the assumption that  $\rightsquigarrow_1 \supseteq \rightsquigarrow_2$ ,  $A$  supports  $B$ , and  $B \notin \Delta$  in  $\rightsquigarrow_1$ , we get that  $\Delta$  is not closed in  $\rightsquigarrow_1$ , contradiction.  $\square$

### Proof of Proposition 15

Let  $A \in Arg$  be the argument under consideration, and we suppose that  $A$  is acceptable under a d-preferred extension of  $\rightsquigarrow_i$  for all  $i \in N$ . Let  $F$  be the unanimity rule. Clearly,  $F(\rightsquigarrow)$  is a subset of  $\rightsquigarrow_i$  for all  $i \in N$ . Without loss of generality, we take  $\rightsquigarrow_i$  to be the BAF under consideration. Then,  $\rightsquigarrow_i \supseteq F(\rightsquigarrow)$ . Furthermore,  $A$  is acceptable under a d-preferred extension  $\Delta_1 \subseteq Arg$  of  $\rightsquigarrow_i$  i.e.,  $A \in \Delta_1$ . Note that  $\Delta_1$  is a d-admissible set as well. According to Lemma 11,  $\Delta_1$  is a d-admissible set of  $F(\rightsquigarrow)$  as well. By Fact 14, we know that there is a d-preferred extension  $\Delta_2 \subseteq Arg$  of  $F(\rightsquigarrow)$  such that  $\Delta_2 \supseteq \Delta_1$ , and  $A$  is a member of  $\Delta_2$ . That is to say,  $A$  is acceptable under a d-preferred extension of  $F(\rightsquigarrow)$ . We are done.  $\square$

# De Re Updates

Michael Cohen

Department of Philosophy  
Stanford University, USA  
micohen@stanford.edu

Wen Tang

Department of Philosophy  
Peking University, China  
1800015421@pku.edu.cn

Yanjing Wang\*

Department of Philosophy  
Peking University, China  
y.wang@pku.edu.cn

In this paper, we propose a lightweight yet powerful dynamic epistemic logic that captures not only the distinction between *de dicto* and *de re* knowledge but also the distinction between *de dicto* and *de re* updates. The logic is based on the dynamified version of an epistemic language extended with the assignment operator borrowed from dynamic logic, following the work of Wang and Seligman [35]. We obtain complete axiomatizations for the counterparts of public announcement logic and event-model-based DEL based on new reduction axioms taking care of the interactions between dynamics and assignments.

## 1 Introduction

Epistemic logic is very successful in capturing reasoning patterns of propositional knowledge expressed in terms of *knowing that*. It has been widely applied to formal epistemology, game theory, theoretical computer science, and AI (cf. [9]).

In particular, the development of dynamic epistemic logic (DEL) provides a flexible framework to formally model how propositional knowledge is communicated and updated by concrete actions and events (cf. e.g., [8]). For example, in public announcement logic (PAL), the *knowing that* modality is equipped by its dynamic counterpart of *announcing that* modality, and the implicit assumptions about agents' ability of obtaining new knowledge are reflected by the interaction of these two modalities in terms of the axioms such as perfect recall and no miracles [5, 32]. These axioms together with other axioms about the features of updates also give rise to the so-called *reduction axioms*, which can often be used to eliminate the dynamic modalities within the static epistemic logic.

### 1.1 De re knowledge and updates

Despite the great success of the standard epistemic logic of *knowing that*, there are also other commonly used knowledge expressions such as *knowing what/who/how/why* and so on, which were not well-studied in the standard framework. As already observed by Hintikka in the early days of epistemic logic, such expressions are about *knowledge of objects*, or say *de re* knowledge, compared to the *de dicto* knowledge expressed by *knowing that  $\varphi$*  (cf. e.g., [15]). Hintikka pioneered the approach of using first-order (or higher-order) modal logic to capture such expressions [16], e.g., knowing who murdered Bob can be formalized as  $\exists x \mathcal{K}Kill(x, Bob)$ , in contrast with the *de dicto* knowledge that someone murdered Bob  $\mathcal{K}\exists x Kill(x, Bob)$ .

Inspired by Hintikka's early idea and discussions in philosophy and linguistics about embedded wh-questions [12, 28], Wang proposed to introduce the *bundle modalities* that pack a

---

\*Corresponding author

quantifier and an epistemic modality together to capture each *know-wh* as a whole, instead of breaking it down into smaller components [30]. This leads to a new family of (non-normal) epistemic logics of *know-wh* and new decidable fragments of first-order modal logics [29, 23].

Now a very natural question arises, **how do we capture the dynamics of such *de re* knowledge?** More specifically, **can we repeat the success of DEL with a genuine *de re* counterpart?** We hope the present paper can provide positive answers to such questions by presenting a framework which can handle both *de re* and *de dicto* knowledge and updates.

Let us first understand the technical difficulties in handling the *de re* dynamics in the existing framework. To be more specific, consider a logic of *knowing what* featuring the  $K_v$  modality introduced in the very same paper where Plaza invented the public announcement logic (**PAL**) [24]. Given a non-rigid name  $a$ ,  $K_v i a$  says that agent  $i$  knows (what) the value/reference of  $a$  (is). It has a very intuitive semantics induced by its hidden first-order modal form of  $\exists x K_i(x \approx a)$ . After failing to apply the reduction method of **PAL**, Plaza proposed the question of axiomatizing such a logic with the presence of public announcements. Wang and Fan showed that there is simply no reduction possible in Plaza's language and use a strengthened conditional  $K_v$  modality to axiomatize the logic [33, 34]. Note that although *de dicto* announcements can possibly involve or change *de re* knowledge as nicely shown in [21, 22],<sup>1</sup> it is not the most natural dynamic counterpart of the  $K_v i$  operator, as the table below shows:

	knowledge	dynamics
<i>de dicto</i>	knowing that	announcing that
<i>de re</i>	knowing what	announcing what

Just like one can know a proposition after an announcement, one should be able to know the value of  $a$  after it is announced. However, announcing the value of  $a$  cannot be easily handled by *announcing that*, e.g., suppose the domain of  $a$  is the set of natural numbers, then you need to use infinitely many non-deterministic announcements in the form of  $a \approx k$  for each  $k \in \mathbb{N}$ . Does it mean we need to introduce constants for all the numbers in the language? What if the value domain is uncountable?

Given such concerns, a new *announcing value* modality  $[a]$  was introduced in [11] with its dynamic semantics of eliminating the worlds which do not share the same value of  $a$  as the designated actual world.<sup>2</sup> However, despite some success of axiomatizing  $[a]$  and  $K_v$  in very restricted cases, it remains hard to capture the full logic with know-that, know-what, announce-that, and announce-what. Unlike the announcement operator,  $[a]$  does not obey the *no miracles* axiom,<sup>3</sup> which is one of the pillars behind dynamic epistemic logic (cf. [32]). One reason for this failure is that the *de re* update is *not global*, i.e., the updated effect depends on the value of  $a$  on the world where it is executed. It also leads to the problem of not being able to reduce such a dynamic operator. One way to go around is to introduce some rigid constants and all kinds of conditional knowledge operators as in [2] to eliminate the dynamic operators in a much stronger background logic. However, this seems to be a little bit *ad hoc*, especially when we consider more general *de re* updates which are not entirely public, such as telling  $i$  the passwords of  $c$  and  $d$  but letting the observer  $j$  be uncertain about which is which. **Is there a simpler/natural yet more powerful framework to handle all these dynamics in a uniform way?** Our answer is again affirmative, as to be explained below.

<sup>1</sup>For example, announcing that Bob knows what is the value of the password.

<sup>2</sup>The modality is called the *public inspection* in [11].

<sup>3</sup>A typical *no miracles* axiom is in the shape of  $\langle e \rangle K \phi \rightarrow K[e] \phi$ . It is not valid if  $[e]$  is the announcing-what operator.

## 1.2 Bridging *de re* and *de dicto* by the assignment operator

Our main inspiration comes from the treatment of *de re* knowledge using the *assignment operator*  $[x := t]$  from first-order *dynamic logic* [14]. The intuitive semantics of  $[x := t]$  is simply an imperative one: assigning variable  $x$  the current value of the term  $t$ . As remarked in [19], Pratt in [25] already noticed the connection between the assignment operator and the  $\lambda$ -*abstraction* that is often used to distinguish the *de re* and *de dicto* readings in first-order modal logic [27, 10].<sup>4</sup> Our technical framework follows the quantifier-free epistemic logic with assignments studied in [35], but without considering the termed modalities there. The core idea is that we can use the assignment operator to “store” the actual reference of a certain term, and use it under the right scope to capture various forms of *de re* knowledge. For example,  $[x := a]K_i P x \wedge \neg K_i P a$  says that agent  $i$  knows of  $a$  that  $P$ , but does not know that  $P a$ . As another simple example, note that  $[x := a]K_i(x \approx a)$  actually expresses that  $i$  knows what  $a$  is, exactly as  $K_{v_i} a$  in the knowing value logic that we mentioned [24]. Essentially, the assignment operator can be used as the *bridge* between the *de dicto* and the *de re* knowledge. Now comes the natural question: **can it also bridge the *de dicto* and *de re* updates?**

The answer is positive and the solution is surprisingly simple: we just need to use the usual DEL dynamic operators such as public announcement or event updates together with the assignment operator. For example, the *de dicto* update of announcing that  $x \approx a$  can be turned into the *de re* update of announcing the value of  $a$  by adding the assignment operator  $[x := a]$  in front of the announcement operator  $[\!|x \approx a]$ . As we will show later, the combination of the announcement and the assignment is very powerful and can capture various notions of dependency and conditionals mixing *de re* and *de dicto* updates. The example of telling  $i$  the passwords of  $c$  and  $d$  without letting  $j$  know which is which can also be easily handled by using a two-world event model with  $x \approx c \wedge y \approx d$  and  $x \approx d \wedge y \approx c$  as the preconditions respectively, in the scope of two assignment operators  $[x := c][y := d]$ . This will become clear when we introduce the event update formally later on.

Our treatment of the public announcements and event updates is basically the same as in the standard DEL. Thus the basic properties such as *perfect recall* and *no miracles* between the knowledge operator and dynamic operators stay the same. The combination of the assignment and the dynamics together is responsible for the apparent failure of *no miracles*.<sup>5</sup> As in the standard DEL, we will show that the dynamic operators can be eliminated.

Our contributions in this paper are summarized below:

- We propose a lightweight dynamic epistemic framework with assignment operators, which can handle both *de re* and *de dicto* knowledge and updates in a uniform way.
- The public announcement operator and event model update operators can be eliminated qua expressivity as in the standard dynamic epistemic logic.
- We obtain complete axiomatizations of all the logics introduced in the paper.

The technical results are relatively straightforward. The main point of the paper is to highlight the use of the assignment operators in capturing the *de re* updates, and propose the alternative

<sup>4</sup>There is a large body of research in modal logic trying to distinguish *de re* and *de dicto* readings, cf. e.g., also [13, 7, 17, 26].

<sup>5</sup>The non-global nature of *de re* updates comes from the assignments which only record the local value. This is also related to logics of local announcements [3] and to the study of opaque updates whose result is not always antecedently known to the agent [6].

static epistemic logic which can pre-encode *de re* dynamics. The **magic of the assignment operator** is that it can automatically turn *de dicto* notions into the corresponding *de re* notions almost for free. Therefore we just need to add the assignment operator to a relatively standard *de dicto* epistemic framework to capture all those *de re* knowledge and updates, without introducing various new *ad hoc* modalities.

We hope our framework can also bring new tools for philosophical analysis related to *de re* updates. For example, *de re* updates can be used to analyze scenarios in which an agent has *de dicto* knowledge of every proposition but is still able to learn new *de re* knowledge. Such learning events require *de re* updates. Scenarios in which a propositionally omniscient agent is able to learn something new about their environment play a central role in philosophy of mind (e.g., Frank Jackson's *Mary's room* thought experiment [20, 18]). We leave the philosophical discussion to a future occasion.

**Structure of the paper** In Section 2, we introduce the basic epistemic logic with assignments and its axiomatization. Section 3 adds the public announcement operator and Section 4 discusses the event model updates with and without factual changes. We conclude with future directions in Section 5.

## 2 Epistemic logic with assignments

In this section, we present a language of epistemic logic with assignments. It can be viewed as a simplified version of the language studied in [35] without the term-modalities.

### 2.1 Language and Semantics

**Definition 1 (Language of BELAS)** *Given a set of variables  $X$ , set of names  $N$ , set of agents  $I$  and a set of predicate symbols  $P$ , the language of Basic Epistemic Logic with Assignments (BELAS) is defined as:*

$$t ::= x \mid a$$

$$\varphi ::= t \approx t \mid P\vec{t} \mid (\varphi \wedge \psi) \mid \neg\varphi \mid K_i\varphi \mid [x := t]\varphi$$

where  $x \in X$ ,  $a \in N$ ,  $P \in P$ , and  $i \in I$ .

We call  $t \approx t'$  and  $P\vec{t}$  atomic formulas. We use the usual abbreviations  $\vee, \rightarrow, \widehat{K}_i, \langle x := t \rangle$ , and write  $Kv_i a$  for  $[x := a]K_i(x \approx a)$ . As we discussed in the introduction,  $Kv_i a$  says the agent  $i$  knows the value of  $a$ . Based on the semantics to be given later,  $Kv_i$  is indeed the same know-value modality discussed in [24, 33].

Following [35], we define the free and bound occurrences of a variable in a **BELAS**-formula by viewing  $[x := t]$  in  $[x := t]\varphi$  as a quantifier binding  $x$  in  $\varphi$ . We call  $x$  a *free variable* in  $\varphi$  if there is a *free occurrence* of  $x$  in  $\varphi$ . Formally the set of free variables  $Fv(\varphi)$  in  $\varphi$  is defined as follows:

$$\begin{aligned} Fv(P\vec{t}) &= \text{Var}(\vec{t}) & Fv(t \approx t') &= \text{Var}(t) \cup \text{Var}(t') \\ Fv(\neg\varphi) &= Fv(\varphi) & Fv(\varphi \wedge \psi) &= Fv(\varphi) \cup Fv(\psi) \\ Fv(K_i\varphi) &= Fv(\varphi) & Fv([x := t]\varphi) &= (Fv(\varphi) \setminus \{x\}) \cup \text{Var}(t) \end{aligned}$$

where  $\text{Var}(\vec{t})$  is the set of variables in  $\vec{t}$ . We use  $\varphi[y/x]$  to denote the result of substituting  $y$  for all the free occurrences of  $x$  in  $\varphi$ , and say  $\varphi[y/x]$  is *admissible* if all the occurrences of  $y$  by replacing free occurrences of  $x$  in  $\varphi$  are also free in  $\varphi[y/x]$ . It is showed in [35] that  $[x := t]$  indeed behaves like a quantifier via a translation to a 2-sorted first-order logic.

The models are simply first-order Kripke models.

**Definition 2 (Models)** A (constant domain) Kripke model  $\mathcal{M}$  is a tuple  $\langle W, D, R, \rho, \eta \rangle$  where:

- $W$  is a non-empty set of possible worlds.
- $D$  is a non-empty set of objects, called the domain of  $\mathcal{M}$ .
- $R : I \rightarrow 2^{W \times W}$  assign a binary relation  $R(i)$  (also written  $R_i$ ) between worlds, to each agent  $i$ .
- $\rho : P \times W \rightarrow \bigcup_{n \in \omega} 2^{D^n}$  assigns an  $n$ -ary relation over  $D$  each  $n$ -ary predicate  $P$  at each world.
- $\eta : N \times W \rightarrow D$  assigns an object to each name  $a \in N$  at each world  $w$ .

Given  $\mathcal{M}$ , we refer to its components as  $W_{\mathcal{M}}, D_{\mathcal{M}}, R_{\mathcal{M}}, \rho_{\mathcal{M}}, \eta_{\mathcal{M}}$ . A pointed Kripke model is a triple  $\mathcal{M}, w, \sigma$ , where  $w \in W_{\mathcal{M}}$  and  $\sigma : X \rightarrow D_{\mathcal{M}}$  assigns an object to every variable. Given  $\mathcal{M}$ , and a world  $w$ ,  $\sigma$  can be lifted to  $\sigma_w$  over all the terms  $t$  such that  $\sigma_w(a) = \eta(a, w)$  for names. An epistemic model is a model where the relations are equivalence relations.

Note that the names in  $N$  and predicates are non-rigid designators.

**Definition 3 (Semantics)** The truth conditions are given with respect to  $\mathcal{M}, w, \sigma$ :

$$\begin{array}{l} \hline \mathcal{M}, w, \sigma \models t \approx t' \Leftrightarrow \sigma_w(t) = \sigma_w(t') \\ \mathcal{M}, w, \sigma \models P(t_1 \cdots t_n) \Leftrightarrow (\sigma_w(t_1), \dots, \sigma_w(t_n)) \in \rho(P, w) \\ \mathcal{M}, w, \sigma \models \neg \varphi \Leftrightarrow \mathcal{M}, w, \sigma \not\models \varphi \\ \mathcal{M}, w, \sigma \models (\varphi \wedge \psi) \Leftrightarrow \mathcal{M}, w, \sigma \models \varphi \text{ and } \mathcal{M}, w, \sigma \models \psi \\ \mathcal{M}, w, \sigma \models K_i \varphi \Leftrightarrow \mathcal{M}, v, \sigma \models \varphi \text{ for all } v \text{ s.t. } wR_i v \\ \mathcal{M}, w, \sigma \models [x := t] \varphi \Leftrightarrow \mathcal{M}, w, \sigma[x \mapsto \sigma_w(t)] \models \varphi \\ \hline \end{array}$$

where  $\sigma[x \mapsto \sigma_w(t)]$  denotes an assignment that is the same as  $\sigma$  except for mapping  $x$  to  $\sigma_w(t)$ .

Now we can check the derived semantics for  $Kv_i$ :

$$\begin{array}{l} \mathcal{M}, w, \sigma \models Kv_i a \\ \Leftrightarrow \mathcal{M}, w, \sigma \models [x := a] K_i (x \approx a) \\ \Leftrightarrow \mathcal{M}, v, \sigma[x \mapsto \sigma_w(a)] \models x \approx a \text{ for all } v \text{ s.t. } wR_i v \\ \Leftrightarrow \sigma_w(a) = \sigma_v(a) \text{ for all } v \text{ s.t. } wR_i v \end{array}$$

Over reflexive models we have the semantics in [33]:

$$\mathcal{M}, w, \sigma \models Kv_i a \Leftrightarrow \sigma_w(a) = \sigma_{v'}(a) \text{ for all } v, v' \text{ s.t. } wR_i v \text{ and } wR_i v'.$$

**Example 4** Consider the following model  $\mathcal{M}$  as a simple example with two worlds  $s, t$ , a signature that contains only the unary predicate  $P$ , one agent  $i$ , a domain with two objects  $o_1, o_2$ , a  $\rho$  such that  $\rho(P, s) = \{o_1\}, \rho(P, t) = \{o_2\}$ , and an  $\eta$  depicted below by abusing the symbol  $\approx$ :

$$\begin{array}{c} \begin{array}{ccc} \curvearrowleft & & \curvearrowright \\ s : a \approx o_1, b \approx o_2 & \xleftarrow{i} & t : a \approx o_2, b \approx o_1 \end{array} \end{array}$$

In the above example, agent  $i$  has the *de dicto* knowledge that  $P(a)$ :  $\mathcal{M}, s, \sigma \models K_i P(a)$  for any  $\sigma$ , since the formula does not contain any free variable. However, note that  $\mathcal{M}, s, \sigma \models \neg[x := a]K_i P(x)$ , i.e., in the actual world  $s$  (underlined), agent  $i$  does not have the *de re* knowledge that the object that  $a$  denotes (object  $o_1$ ) has property  $P$ . Although the agent knows all the propositional facts regarding the property  $P$ , it still has the *de re* ignorance. Further note that no closed formula involving just  $P$  can distinguish states  $s$  and  $t$ . A *de re* update is needed for the agent to learn that state  $t$  is not the actual world.

Adapting the proofs in [35, 36], it is not hard to show the following, which we leave for the full version of the paper:

- $[x := t]$  cannot be eliminated in **BELAS** qua expressivity.
- **BELAS** is decidable over arbitrary and reflexive models;
- **BELAS** is undecidable over S5 models.

The undecidability can be shown by coding Fitting's undecidable logic **S5** $\lambda_{\approx}$  where instead of the assignment operator, the  $\lambda$ -abstraction  $\langle \lambda x. \varphi \rangle$  is used to handle the distinction between *de dicto* and *de re*, e.g.,  $\langle \lambda x. \Box \langle \lambda y. y \approx x \rangle (c) \rangle (c)$  says  $c$  is rigid, which is equivalent to our  $[x := c]K[y := c]x \approx y$ . Note that our assignment operator is much easier to read compared to the  $\lambda$ -abstraction.

## 2.2 Axiomatization

Based on the axioms in [35], we proposed the following proof system SBELAS.

Axiom Schemas	
TAUT	all the instances of tautologies
DISTK	$K_i(\varphi \rightarrow \psi) \rightarrow (K_i\varphi \rightarrow K_i\psi)$
ID	$t \approx t$
SYM	$t \approx t' \leftrightarrow t' \approx t$
TRANS	$t \approx t' \wedge t' \approx t'' \rightarrow t \approx t''$
SUBAS	$t \approx t' \rightarrow ([x := t]\varphi \leftrightarrow [x := t']\varphi)$
SUBP	$\vec{t} \approx \vec{t}' \rightarrow (P\vec{t} \leftrightarrow P\vec{t}')$
RIGIDP	$x \approx y \rightarrow K_ix \approx y$
RIGIDN	$x \not\approx y \rightarrow K_ix \not\approx y$
KAS	$[x := t](\varphi \rightarrow \psi) \rightarrow ([x := t]\varphi \rightarrow [x := t]\psi)$
DETAS	$\langle x := t \rangle \varphi \rightarrow [x := t]\varphi$
DAS	$\langle x := t \rangle \top$
EFAS	$[x := t]x \approx t$
SUB2AS	$\varphi[y/x] \rightarrow [x := y]\varphi$ ( $\varphi[y/x]$ is admissible)
Rules	
NECK	$\frac{\varphi}{K_i\varphi}$
MP	$\frac{\varphi, \varphi \rightarrow \psi}{\psi}$
NECAS	$\frac{\psi}{\vdash \varphi \rightarrow \psi} \quad (x \text{ is not free in } \varphi)$

where in SUBP,  $\vec{t} \approx \vec{t}'$  is the abbreviation of the conjunction of point-wise equivalences for sequences of terms  $\vec{t}$  and  $\vec{t}'$  such that  $|\vec{t}| = |\vec{t}'|$ . The system SBELAS5 is defined as SBELAS together with the usual S5 schemata for  $K_i$ : T :  $K_i\varphi \rightarrow \varphi$ , 4 :  $K_i\varphi \rightarrow K_iK_i\varphi$ , and 5 :  $\neg K_i\varphi \rightarrow K_i\neg K_i\varphi$ .

Note that ID, SYM, TRANS captures the nature of equality; SUBP and SUBAS regulate the substitution that we can safely do; RIGIDP and RIGIDN describe that the variables are rigid; KAS, DETAS, DAS capture that the assignment operator is a self-dual normal modality (the necessitation rule for  $[x := t]$  is a special case of NECAS); EFAS characterizes the effect of the assignment operator; SUB2AS and NECAS are the counterparts of the usual axiom and rules of the first-order quantifier.

**Remark 5** Comparing to [35], we do not have the special axioms to handle the term-modalities, and the S5 axioms for the epistemic operators are standard.

Given that  $K_i$  and  $[x := t]$  are normal modalities, we can show that the rule of replacement of equals is an admissible rule in the systems SBELAS and SBELAS5.

**Theorem 6 (Soundness)** System SBELAS is sound over Kripke models and SBELAS5 is sound over epistemic models. The following are handy for the completeness proof.

**Proposition 7** The following are derivable from the system:

$$\begin{array}{l}
\text{DBASEQ} \quad \langle x := t \rangle \varphi \leftrightarrow [x := t] \varphi \\
\text{CNECAS} \quad \frac{\vdash \varphi \rightarrow \psi}{\vdash [x := t] \varphi \rightarrow \psi} \quad (x \notin Fv(\psi)) \\
\text{EAS} \quad [x := t] \varphi \leftrightarrow \varphi (x \notin Fv(\varphi)) \\
\text{SUBASEQ} \quad \varphi[y/x] \leftrightarrow [x := y] \varphi \quad \text{given } \varphi[y/x] \text{ is admissible} \\
\text{NECAS}' \quad \frac{\vdash \varphi}{[x := t] \varphi}
\end{array}$$

PROOF DBASEQ is based on DETAS and DAS. CNECAS is due to NECAS and DBASEQ for contrapositive. EAS is based on NECAS and CNECAS (taking  $\psi = \varphi$ ). SUBASEQ is due to the contrapositive of SUB2AS and DBASEQ. NECAS' is special case for NECAS. ■

We can also rename the bound variables as shown in [35].

**Proposition 8 (Relettering)** Let  $z$  be fresh in  $\varphi$  and  $t$ , then  $\vdash [x := t] \varphi \leftrightarrow [z := t] \varphi[z/x]$ .

The completeness of SBELAS and SBELAS5 can be proved by an adaptation of the corresponding (highly non-trivial) completeness proof in [35] without the treatment for the term-modalities in [35]. We omit the proof and leave it to the full version due to page limitation.

**Theorem 9** SBELAS is strongly complete over arbitrary models and SBELAS5 is strongly complete over epistemic models.

### 3 Adding public announcement

In this section, we develop a public announcement logic based on the language BELAS. We will show that as in the case of standard PAL, the announcement operator can be eliminated.

#### 3.1 Language and semantics

**Definition 10 (Language of PALAS)** The language of Public Announcement Logic with Assignments (PALAS) is defined by adding the announcement operator to BELAS:

$$\begin{array}{l}
t ::= x \mid a \\
\varphi ::= t \approx t \mid P\vec{t} \mid (\varphi \wedge \varphi) \mid \neg \varphi \mid K_i \varphi \mid [x := t] \varphi \mid [! \varphi] \varphi
\end{array}$$

where  $x \in X$ ,  $a \in N$ ,  $P \in P$  and  $i \in I$ .

As in the standard **PAL**,  $[\psi]\phi$  intuitively says that if  $\psi$  can be truthfully announced then afterwards  $\phi$  holds. Besides the usual abbreviations, we also write  $\langle !\phi \rangle$  for  $\neg[!\phi]\neg$ .

With this simple addition, we can capture the *de re* updates of publicly announcing the actual value of  $a$  by  $[x := a][!x \approx a]$ . In the following, we will also write  $[!a]\phi$  for  $[x := a][!x \approx a]\phi$  when  $x$  is not free in  $\phi$ . This is essentially the  $[a]$  operator introduced in [11].

We can actually do *much more* beyond announcing the value of  $a$ . For example, the *de re* announcement that the reference of  $a$  does have the property  $P$  is expressed by  $[x := a][!Px]$ . This will give the *de re* knowledge that object  $o_1$  has property  $P$  to the agent in Example 4.

The semantics of the announcement operator is essentially the same as in standard **PAL**.

### Definition 11 (Semantics)

$$\overline{\mathcal{M}, w, \sigma \models [!\psi]\phi \Leftrightarrow \mathcal{M}, w, \sigma \models \psi \text{ implies } \mathcal{M}|_{\psi}^{\sigma}, w, \sigma \models \phi}$$

where  $\mathcal{M}|_{\psi}^{\sigma}$  is the submodel of  $\mathcal{M}$  restricted to the  $\psi$  worlds in  $\mathcal{M}$ , i.e.,  $\mathcal{M}|_{\psi}^{\sigma} = \{W', D_{\mathcal{M}}, R', \eta'\}$  such that  $W' = \langle v \mid \mathcal{M}, v, \sigma \models \psi \rangle$ ,  $R'_i = R_i|_{W' \times W'}$  and  $\eta' = \eta|_{W'}$ .

Now we can check the induced semantics of  $[!a]$ :

$$\mathcal{M}, w, \sigma \models [!a]\phi \Leftrightarrow \mathcal{M}, w, \sigma \models [x := a][!x \approx a]\phi \Leftrightarrow \mathcal{M}, w, \sigma[x \mapsto \sigma_w(a)] \models [!x \approx a]\phi \Leftrightarrow \mathcal{M}|_{x \approx a}^{\sigma[x \mapsto \sigma_w(a)]} \models \phi$$

where  $\mathcal{M}|_{x \approx a}^{\sigma[x \mapsto \sigma_w(a)]}$  is the submodel of  $\mathcal{M}$  with all the worlds that share the same value of  $a$  as the actual world. This is indeed the semantics given to the *public inspection operator* in [11].

With the announcement operator, we can also define the conditional operators introduced in [34, 2] over epistemic models. For example:

- $Kv_i(\phi, c) := K_i[!\phi]Kv_i c$ : Agent  $i$  would know the value of  $c$  given  $\phi$ .
- $Kv_i(c, d) := K_i[!c]Kv_i d$ : Agent  $i$  would know the value of  $d$  given the value of  $c$ , namely, agent  $i$  knows how the value of  $d$  functionally depends on the value of  $c$ ;
- $Kv_i(c, \phi) := K_i[!c](K_i \phi \vee K_i \neg \phi)$ : Agent  $i$  would know the truth value of  $\phi$  given the value of  $c$ , i.e., agent  $i$  knows how the truth value of  $\phi$  depends on the value of  $c$ ;
- $Kv_i(\psi, \phi) := K_i([!\psi](K_i \phi \vee K_i \neg \phi) \wedge [!\neg \psi](K_i \phi \vee K_i \neg \phi))$ : Agent  $i$  knows how the truth value of  $\phi$  depends on the truth value of  $\psi$ .

Based on the semantics, it is not hard to show the axioms of perfect recall and no miracles are still valid, which form the foundation for the reduction of the announcement operator to be introduced (cf. [32]).

**Proposition 12** *The following are valid:*

$$\begin{array}{l} \text{PR} \quad K_i[!\psi]\phi \rightarrow [!\psi]K_i\phi \\ \text{NM} \quad \langle !\psi \rangle K_i\phi \rightarrow K_i[!\psi]\phi \end{array}$$

## 3.2 Axiomatization

We define the proof system SPALAS (SPALAS5) as the proof system obtained by extending SBELAS (SBELAS5) with the following reduction axioms, which help us to eliminate the announcement operator in **PALAS**. AK is essentially the combination of the axioms PR and NM (cf. [32]). Besides the usual reduction axioms for **PAL**, we have a new axiom AASSI.

Axiom Schemas	
AATOM	$[\! \psi ]p \leftrightarrow (\psi \rightarrow p)$ (if $p$ is atomic)
ANEG	$[\! \psi ]\neg\phi \leftrightarrow (\psi \rightarrow \neg[\! \psi ]\phi)$
ACON	$[\! \psi ](\phi \wedge \chi) \leftrightarrow ([\! \psi ]\phi \wedge [\! \psi ]\chi)$
AK	$[\! \psi ]K_i\phi \leftrightarrow (\psi \rightarrow K_i[\! \psi ]\phi)$
ACOM	$[\! \psi ][\! \chi ]\phi \leftrightarrow [!(\psi \wedge [\! \psi ]\chi)]\phi$
AASSI	$[\! \psi ][x := t]\phi \leftrightarrow [z := x][x := t][\! \psi[z/x] ]\phi$ ( $z$ does not occur in $[\! \psi ][x := t]\phi$ )

**Theorem 13** *SPALAS is sound over arbitrary models.*

PROOF The validity of the first five reduction axioms is as in the standard **PAL**. We only focus on the last one, AASSI, which is about switching the assignment operator and the announcement operator. Note that in AASSI,  $z$  is fresh in  $\phi, \psi$  and  $z \neq t$ , therefore  $\psi[z/x]$  is always admissible. We first prove the following claim:

**Claim 13.1** *For any  $v$  in  $\mathcal{M}$ :*

$$\mathcal{M}, v, \sigma \models \psi \iff \mathcal{M}, v, \sigma \models [z := x][x := t]\psi[z/x]$$

*Proof of Claim 13.1:* Since  $z$  is fresh, and there is no free  $x$  in  $\psi[z/x]$ , we have for any  $v, u$  in  $\mathcal{M}$ :

$$\begin{aligned} & \mathcal{M}, v, \sigma \models \psi \\ \iff & \mathcal{M}, v, \sigma[z \mapsto \sigma(x)] \models \psi[z/x] \\ \iff & \mathcal{M}, v, \sigma[z \mapsto \sigma(x)][x \mapsto \sigma_u(t)] \models \psi[z/x] \quad (\star). \end{aligned}$$

Let  $\sigma^* = \sigma[z \mapsto \sigma(x)]$ . Since  $t \neq z$ , and since changing  $\sigma$  does not affect  $\eta$  on  $u$ , we have for any  $u$  in  $\mathcal{M}$

$$\sigma_u(t) = \sigma_u^*(t) \quad (\dagger)$$

no matter whether  $t$  is a variable or a name. Therefore we have for any  $u$  in  $\mathcal{M}$ :

$$\sigma[z \mapsto \sigma(x)][x \mapsto \sigma_u(t)] = \sigma[z \mapsto \sigma(x)][x \mapsto \sigma_u^*(t)]$$

Now from  $(\star)$  we have :

$$\mathcal{M}, v, \sigma \models \psi \iff \mathcal{M}, v, \sigma[z \mapsto \sigma(x)][x \mapsto \sigma_u^*(t)] \models \psi[z/x] \quad (\ddagger)$$

In particular, taking  $u = v$  in  $(\ddagger)$  gives us the proof for the claim according to the semantics. ■

Now consider the following two cases:

**(Case I)** If  $\mathcal{M}, w, \sigma \not\models \psi$ , then  $\mathcal{M}, w, \sigma \models [\!|\psi|][x := t]\phi$  is trivially true. By the above claim,  $\mathcal{M}, w, \sigma \not\models [z := x][x := t]\psi[z/x]$ . Thus  $\mathcal{M}, w, \sigma \models [z := x][x := t][\!|\psi[z/x]|]\phi$ .

**(Case II)** If  $\mathcal{M}, w, \sigma \models \psi$ , by the above claim,  $\mathcal{M}, w, \sigma \models [z := x][x := t]\psi[z/x]$ . According to the semantics we need to show (1) iff (2) below :

$$(1) \mathcal{M} \mid_{\psi}^{\sigma}, w, \sigma[x \mapsto \sigma_w(t)] \models \phi$$

$$(2) \mathcal{M} \Big|_{\psi[z/x]}^{\sigma[z \mapsto \sigma_w(x)][x \mapsto \sigma_w^*(t)]}, w, \sigma[z \mapsto \sigma_w(x)][x \mapsto \sigma_w^*(t)] \models \varphi$$

Note that  $\sigma(x) = \sigma_w(x)$  by definition. Now taking  $u = w$  in  $(\ddagger)$  immediately shows that  $\mathcal{M} \Big|_{\psi[z/x]}^{\sigma[z \mapsto \sigma_w(x)][x \mapsto \sigma_w^*(t)]}$  is exactly the same model as  $\mathcal{M} \Big|_{\psi}^{\sigma}$ . Now we only need to consider whether the difference between  $\sigma[x \mapsto \sigma_w(t)]$  and  $\sigma[z \mapsto \sigma_w(x)][x \mapsto \sigma_w^*(t)]$  matters for the truth value of  $\varphi$ . Note that  $z$  does not occur in  $\varphi$  and by  $(\dagger)$   $\sigma_w(t) = \sigma_w^*(t)$ , therefore the above difference in  $\sigma$  does not affect the truth value of  $\varphi$ . It follows that (1) iff (2), and this completes the proof. ■

With the formulas above, we can translate **PALAS**-formulas to **BELAS**-formulas and eliminate the public announcement operators.

Based on the above theorem, we can define a translation  $\text{tr}$  to eliminate the announcement operators as in the standard **PAL** using the left-to-right direction of the reduction axioms [8, 32], and the following extra clause (where  $z$  is fresh):

$$\text{tr}([\!|\psi|][x := t]\varphi) = [z := x][x := t]\text{tr}([\!|\psi[z/x]|]\varphi)$$

It is not hard to show that the translation preserves the equivalence of formulas.

**Proposition 14** For all  $\varphi \in \text{PALAS}$ :  $\models \varphi \leftrightarrow \text{tr}(\varphi)$

**Theorem 15** *SPALAS is complete over arbitrary models.*

**PROOF** The proof is done by the following reduction.

$$\models \varphi \implies \models \text{tr}(\varphi) \implies \vdash_{\text{SBELAS}} \text{tr}(\varphi) \implies \vdash_{\text{SPALAS}} \text{tr}(\varphi) \implies \vdash_{\text{SPALAS}} \varphi$$

The first step is due to Proposition 14. The second step is due to Theorem 9. The third step is due to the fact that SPALAS is an extension of SBELAS, and the last step is due to the reduction axioms in the system that you can show  $\vdash_{\text{SPALAS}} \text{tr}(\varphi) \leftrightarrow \varphi$ . ■

Similarly we can show that:

**Theorem 16** *SPALAS5 is complete over epistemic models.*

## 4 Adding Event Models

In this section, we generalize the public announcements to event models proposed in [1]. We first consider the event models without factual changes.

### 4.1 Language and Semantics

**Definition 17 (Event model)** An event model  $\mathcal{E}$  with respect to a given language  $\mathbf{L}$  is a triple:  $\langle E, \succ, \text{Pre} \rangle$  where:

- $E$  is a finite non-empty set of events;
- $\succ: I \rightarrow 2^{E \times E}$  assigns a relation to each agent;
- $\text{Pre}: E \rightarrow \mathbf{L}$  assigns each event a precondition formula.

A pointed event model  $\mathcal{E}, e$  is an event model with a designated event. An epistemic event model is an event model where the accessibility relations are equivalence relations.

We often write  $\succrightarrow_i$  for  $\succrightarrow(i)$  to denote the relation for  $i$ .

**Definition 18 (Update product  $\otimes$ )** Given  $N, I$ , a model  $\mathcal{M} = \langle W, D, R, \eta \rangle$ , an assignment  $\sigma$ , and an event model  $\mathcal{E} = \langle E, \succrightarrow, Pre \rangle$  with respect to a given language  $\mathbf{L}$ , the updated model  $(\mathcal{M} \otimes \mathcal{E})^\sigma$  is a tuple  $\langle W', D, R', \rho', \eta' \rangle$  where:

- $W' = \{(w, e) \mid \mathcal{M}, w, \sigma \models Pre(e)\}$ ;
- $(s, e)R'_i(s', e')$  iff  $sR_i s'$  and  $e \succrightarrow_i e'$ ;
- $\eta'(a, (w, e)) = \eta(a, w)$ ;
- $\rho'(P, (w, e)) = \rho(P, w)$ .

Note that  $\sigma$  is necessary in defining the updated model.

**Definition 19 (Language of DELAS)** The language of Dynamic Epistemic Logic with Assignments (*DELAS*) is defined below:

$$t ::= x \mid a$$

$$\varphi ::= t \approx t \mid P\vec{t} \mid (\varphi \wedge \varphi) \mid \neg\varphi \mid K_i\varphi \mid [x := t]\varphi \mid [\mathcal{E}, e]\varphi$$

where  $x \in \mathbf{X}$ ,  $a \in \mathbf{N}$ ,  $P \in \mathbf{P}$ ,  $i \in \mathbf{I}$ , and  $\mathcal{E}, e$  is a pointed event model w.r.t. *DELAS*.<sup>6</sup>

As in [1], we can compose two event models into one.

**Definition 20 (Composition of event models)** Let  $\mathcal{E} = \langle E, \succrightarrow, Pre \rangle$  and  $\mathcal{E}' = \langle E', \succrightarrow', Pre' \rangle$  be two event models. Then the composition of  $\mathcal{E}$  and  $\mathcal{E}'$  is  $\mathcal{E} \circ \mathcal{E}' = \langle E'', \succrightarrow'', Pre'' \rangle$  where

- $E'' = E \times E'$
- $(e, e') \succrightarrow''_i (f, f') \iff e \succrightarrow_i f$  and  $e' \succrightarrow'_i f'$
- $Pre''(e, e') = Pre(e) \wedge [\mathcal{E}, e]Pre'(e')$

The composition of two pointed model  $\mathcal{E}, e$  and  $\mathcal{E}', e'$  (denoted as  $(\mathcal{E}, e) \circ (\mathcal{E}', e')$ ) is defined as the pointed model  $\mathcal{E} \circ \mathcal{E}', (e, e')$ .

The semantics is as in the standard event-model *DEL*.

**Definition 21 (Semantics)** We give the truth condition for the event updates.

$$\boxed{\mathcal{M}, w, \sigma \models [\mathcal{E}, e]\varphi \iff \mathcal{M}, w, \sigma \models Pre(e) \Rightarrow \mathcal{M} \otimes \mathcal{E}, (s, e), \sigma \models \varphi}$$

With event models, we can capture non-trivial *de re* dynamics. For example, agent 1 is told a password with agent 2 around, but agent 2 is not sure whose password it is: it could be Cindy's (c) or Dave's (d). The following event model captures such an event  $\mathcal{E}$  (with precondition specified):

$$\begin{array}{ccc} \downarrow^{1,2} & & \downarrow^{1,2} \\ \underline{e : x \approx c} & \longleftarrow 2 \longrightarrow & f : x \approx d \end{array}$$

The underlining event is the real event. Suppose initially agent 1 and agent 2 have no idea about the possible passwords of  $c$  and  $d$  (thus think all the natural numbers are possible), the (infinite) initial model  $\mathcal{M}, s$  may look like below:

<sup>6</sup>The event model and formulas are defined by a mutual induction, cf. [1].

$$\frac{\sqrt[1,2]}{s : c \approx 12, d \approx 34} \leftarrow_{1,2} \frac{\sqrt[1,2]}{t : c \approx 4, d \approx 12} \leftarrow \dots$$

According to the semantics, we can verify

$$\mathcal{M}, s \models [x := c][\mathcal{E}, e](Kv_1c \wedge \neg Kv_1d \wedge \neg Kv_2c \wedge \neg Kv_2d \wedge K_2(Kv_1c \vee Kv_1d)).$$

As a variant mentioned in the introduction, we can capture the event where agent 1 and agent 2 are told two numbers (the passwords of  $c$  and  $d$ ) such that agent 1 knows which is which but agent 2 does not know it.

$$\frac{\sqrt[1,2]}{e : x \approx c, y \approx d} \leftarrow_2 \frac{\sqrt[1,2]}{f : x \approx d, y \approx c}$$

We can verify:

$$\mathcal{M}, s \models [x := c][y := d][\mathcal{E}, e](Kv_1c \wedge Kv_1d \wedge \neg Kv_2c \wedge \neg Kv_2d \wedge K_2(Kv_1c \wedge Kv_1d)).$$

## 4.2 Axiomatization

We define the proof system SDELAS (SDELAS5) as the proof system obtained by extending SBELAS (SBELAS5) with the following reduction axioms:

Axiom Schemas	
UATOM	$[\mathcal{E}, e]p \leftrightarrow (Pre(e) \rightarrow p)$ ( $p$ is atomic)
UNEG	$[\mathcal{E}, e]\neg\varphi \leftrightarrow (Pre(e) \rightarrow \neg[\mathcal{E}, e]\varphi)$
UCON	$[\mathcal{E}, e](\varphi \wedge \psi) \leftrightarrow ([\mathcal{E}, e]\varphi \wedge [\mathcal{E}, e]\psi)$
UK	$[\mathcal{E}, e]K_i\varphi \leftrightarrow (Pre(e) \rightarrow \bigwedge_{e \rightarrow_i f} [\mathcal{E}, f]\varphi)$
UCOM	$[\mathcal{E}, e][\mathcal{E}', e']\varphi \leftrightarrow [\mathcal{E} \circ \mathcal{E}', (e, e')]\varphi$
UASSI	$[\mathcal{E}, e][x := t]\varphi \leftrightarrow [z := x][x := t][\mathcal{E}', e]\varphi$

where in UASSI,  $z$  does not occur in  $\varphi$ ,  $t$  or in  $Pre_{\mathcal{E}}(e)$  for all  $e \in \mathcal{E}$ .  $\mathcal{E}'$  is an event model with the same domain and relations as  $\mathcal{E}$  and for all  $f$  in  $\mathcal{E}$ ,  $Pre_{\mathcal{E}'}(f) = Pre_{\mathcal{E}}(f)[z/x]$ .

Note that when  $\mathcal{E}$  is a singleton model with precondition  $\psi$  then UASSI coincides with AASSI.

**Theorem 22** *SPALAS is sound over arbitrary models.*

**PROOF** We only need to check UASSI. The proof is similar to the proof of the validity of AASSI. The idea is to use a fresh variable to store the initial value of  $x$ , and replace the free occurrences of  $x$  by  $z$  in the preconditions in  $\mathcal{E}$  such that all these preconditions will be evaluated according to the initial value of  $x$ . Note that the reduction also depends on the fact that the update itself does not change the value of  $t$  or  $x$ . ■

Similarly, as in the previous section, we have the completeness:

**Theorem 23** *SDELAS is complete over arbitrary models and SDELAS5 is complete over epistemic models.*

### 4.3 Adding factual changes

Finally, let us consider event models with factual changes inspired by [4].

**Definition 24 (Event model with factual changes)** *An event model with factual changes  $\mathcal{E}$  w.r.t. language  $\mathbf{L}$  is a tuple:  $\langle E, \succ, Pre, Pos \rangle$  where  $E, \succ, Pre$  are defined as before, and*

- $Pos : \mathbf{N} \times E \rightarrow \mathbf{N}$  is a function that maps all but finite names to themselves.

Intuitively, a post condition changes the value of one name to the value of another one, e.g., an event may switch the value of  $c$  and  $d$  by setting  $Pos(c, e) = d$  and  $Pos(d, e) = c$ .<sup>7</sup>

Accordingly, we also incorporate the factual change in the definition of the update:

**Definition 25 (Update product with factual change)** *Given  $\mathbf{N}, \mathbf{I}$ , a model  $\mathcal{M} = \langle W, D, R, \rho, \eta \rangle$ , an assignment  $\sigma$ , and an event model  $\mathcal{E} = \langle E, \succ, Pre, Pos \rangle$  w.r.t.  $\mathbf{L}$ , the updated Kripke model  $(\mathcal{M} \otimes \mathcal{E})^\sigma$  is a tuple  $\langle W', D, R', \rho', \eta' \rangle$  where:  $W', D, R', \rho'$  are as before, and*

- $\eta'(a, (w, e)) = \eta(Pos(a, e), w)$

We will show that with factual changes,  $[\mathcal{E}, e]$  can still be eliminated.

Given an event model with postconditions  $\mathcal{E}$ , we first lift the  $Pos_{\mathcal{E}}$  to the function  $Pos_{\mathcal{E}}^+ : (\mathbf{N} \cup \mathbf{X}) \times E \rightarrow \mathbf{N} \cup \mathbf{X}$  where for all  $x \in \mathbf{X}$  and  $e \in E$ ,  $Pos_{\mathcal{E}}^+(x, e) = x$  and for all  $a \in \mathbf{N}$ ,  $Pos_{\mathcal{E}}^+(a, e) = Pos_{\mathcal{E}}(a, e)$ .

Now we can state the new reduction axiom.

$$\text{UASSI}' \quad \frac{}{[\mathcal{E}, e][x := t]\varphi \leftrightarrow [z := x][x := Pos_{\mathcal{E}}^+(t, e)][\mathcal{E}', e]\varphi}$$

where  $z$  is fresh.  $\mathcal{E}'$  is defined as before with the new component  $Pos_{\mathcal{E}'}(a, e) = Pos_{\mathcal{E}}(a, e)$ , i.e., the postcondition is unchanged.

It is not hard to verify that UASSI' is valid, and we can use it to give a complete axiomatization as before given the event models with factual changes.

## 5 Future directions

In this work, we propose a lightweight dynamic epistemic framework to capture *de re* knowledge and updates. In particular, **BELAS** can be viewed as a more powerful alternative to the standard epistemic logic, which can also pre-encode *de re* dynamics. There are numerous directions for further work, inspired by the large body of research on the standard (dynamic) epistemic logic. Here we just list a few.

- As in [36], we can try to add function symbols and allow varying domain in the model.
- Adding the usual common knowledge operator will create complications in axiomatization as in the case of **DEL**, but the *de re* common knowledge comes for free.
- As in [4], we can try to build a more general framework based on *dynamic logic*. It makes particular sense since the assignment operator is actually from dynamic logic.
- We can develop the counterparts of the non-reductive axiomatizations in [32, 31].
- We can try to find the boundary of the decidability given different frame conditions. We know S5 is bad but T is good [36], what about S4?

<sup>7</sup>It is more interesting to have function symbols in the language to change the value of  $a$  to the value of a term, which we leave for future work.

- It is also interesting to see whether our framework can capture all the intuitive *de re* updates. We think the answer is negative, e.g., it seems hard to capture the private announcement of some value using finite event models. It may lead to further extensions of the framework.

**Acknowledgement** Michael Cohen thanks the Pre-Doctoral Fellowship Program of Stanford Center at Peking University, which made this joint work possible. Yanjing Wang acknowledges the support of NSSF grant 19BZX135.

## References

- [1] A Baltag, L Moss & S Solecki (1998): *The logic of public announcements, common knowledge, and private suspicions*. In: *Proceedings of TARK '98*, Morgan Kaufmann Publishers Inc., pp. 43–56.
- [2] Alexandru Baltag (2016): *To Know is to Know the Value of a Variable*. In: *Advances in Modal Logic Vol. 11*, pp. 135–155. Available at <http://www.aiml.net/volumes/volume11/Baltag.pdf>.
- [3] Francesco Belardinelli, Hans van Ditmarsch & Wiebe van der Hoek (2017): *A Logic for Global and Local Announcements*. *Electronic Proceedings in Theoretical Computer Science* 251, p. 28–42, doi:10.4204/eptcs.251.3.
- [4] J van Benthem, J van Eijck & B Kooi (2006): *Logics of communication and change*. *Information and Computation* 204(11), pp. 1620–1662, doi:10.1016/j.ic.2006.04.006.
- [5] J van Benthem, J Gerbrandy, T Hoshi & E Pacuit (2009): *Merging Frameworks for Interaction*. *Journal of Philosophical Logic* 38(5), pp. 491–526, doi:10.1007/s10992-008-9099-x.
- [6] Michael Cohen (2020): *Opaque Updates*. *Journal of Philosophical Logic*, pp. 1–24, doi:10.1007/s10992-020-09571-8.
- [7] Giovanna Corsi & Eugenio Orlandelli (2013): *Free quantified epistemic logics*. *Studia Logica* 101(6), pp. 1159–1183, doi:10.1007/s11225-013-9528-x.
- [8] H van Ditmarsch, W van der Hoek & B Kooi (2007): *Dynamic Epistemic Logic*. Springer, doi:10.1007/978-1-4020-5839-4.
- [9] Hans van Ditmarsch, Joseph Halpern, Wiebe van der Hoek & Barteld Kooi, editors (2015): *Handbook of Epistemic Logic*. College Publications.
- [10] Melvin Fitting & Richard L Mendelsohn (1998): *First-order modal logic*. Synthese Library, Springer, doi:10.1007/978-94-011-5292-1.
- [11] Malvin Gattinger, Jan van Eijck & Yanjing Wang (2017): *Knowing Values and Public Inspection*. In: *Proceedings of ICLA'17*, pp. 77–90, doi:10.1007/978-3-662-54069-5\_7.
- [12] Jeroen Groenendijk & Martin Stokhof (1982): *Semantic Analysis of "Wh"-Complements*. *Linguistics and Philosophy* 5(2), pp. 175–233, doi:10.1007/BF00351052.
- [13] Adam J Grove & Joseph Y Halpern (1993): *Naming and identity in epistemic logics part I: the propositional case*. *Journal of Logic and Computation* 3(4), pp. 345–378, doi:10.1093/logcom/3.4.345.
- [14] D Harel, D Kozen & J Tiuryn (2000): *Dynamic Logic*. MIT Press, doi:10.7551/mitpress/2516.001.0001.
- [15] Jaakko Hintikka (1996): *Knowledge Acknowledged: Knowledge of Propositions vs. Knowledge of Objects*. *Philosophy and Phenomenological Research* 56(2), pp. 251–275, doi:10.2307/2108519.
- [16] Jaakko Hintikka (2003): *A Second Generation Epistemic Logic and its General Significance*. In Vincent F Hendricks, Klaus Frovin Jørgensen & Stig Andur Pedersen, editors: *Knowledge Contributors*, Springer, pp. 33–55, doi:10.1007/978-94-007-1001-6\_3.

- [17] Wesley H Holliday & John Perry (2014): *Roles, Rigidity, and Quantification in Epistemic Logic*. In Alexandru Baltag & Sonja Smets, editors: *Johan van Benthem on Logic and Information Dynamics*, Springer, pp. 591–629, doi:10.1007/978-3-319-06025-5\_22.
- [18] Frank Jackson (1986): *What Mary Didn't Know*. *Journal of Philosophy* 83(5), pp. 291–295, doi:10.2307/2026143.
- [19] Barteld Kooi (2007): *Dynamic term-modal logic*. In: *Proceedings of LORI-I*.
- [20] David Lewis (1979): *Attitudes de Dicto and de Se*. *Philosophical Review* 88(4), pp. 513–543, doi:10.2307/2184843.
- [21] Andrés Occhipinti Liberman & Rasmus K. Rendsvig (2019): *Dynamic Term-Modal Logic for Epistemic Social Network Dynamics*. In: *Proceedings of LORI-VII*, Springer, pp. 168–182, doi:10.1007/978-3-662-60292-8\_13.
- [22] Andrés Occhipinti Liberman, Andreas Achen & Rasmus Kræmmer Rendsvig (2020): *Dynamic term-modal logics for first-order epistemic planning*. *Artificial Intelligence* 286, p. 103305, doi:10.1016/j.artint.2020.103305.
- [23] Anantha Padmanabha, R. Ramanujam & Yanjing Wang (2018): *Bundled fragments of first-order modal logic: (un)decidability*. In: *Proceedings of FSTTCS '18*, doi:10.4230/LIPIcs.FSTTCS.2018.43.
- [24] J A Plaza (1989): *Logics of public communications*. In: *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, pp. 201–216.
- [25] V R Pratt (1976): *Semantical considerations on Floyd-Hoare Logic*. Technical Report, Cambridge, MA, USA.
- [26] Rasmus K. Rendsvig (2010): *Epistemic Term-Modal Logic*. In Marija Slavkovic, editor: *Proceedings of Student Session of ESSLLI 2010*, pp. 37–46.
- [27] Robert Stalnaker & Richmond Thomason (1968): *Abstraction in First-Order Modal Logic*. *Theoria* 34(3), pp. 203–207, doi:10.1111/j.1755-2567.1968.tb00351.x.
- [28] Jason Stanley & Timothy Williamson (2001): *Knowing how*. *The Journal of Philosophy*, pp. 411–444, doi:10.2307/2678403.
- [29] Yanjing Wang (2017): *A New Modal Framework for Epistemic Logic*. In: *Proceedings of TARK '07*, pp. 515–534, doi:10.4204/EPTCS.251.38.
- [30] Yanjing Wang (2018): *Beyond knowing that: A new generation of epistemic logics*. In: *Jaakko Hintikka on Knowledge and Game-Theoretical Semantics*, Springer, pp. 499–533, doi:10.1007/978-3-319-62864-6\_21.
- [31] Yanjing Wang & Guillaume Aucher (2013): *An Alternative Axiomatization of DEL and Its Applications*. In: *IJCAI*, pp. 1147–1154. Available at <https://www.ijcai.org/Proceedings/13/Papers/172.pdf>.
- [32] Yanjing Wang & Qinxiang Cao (2013): *On axiomatizations of public announcement logic*. *Synthese* 190(1S), pp. 103–134, doi:10.1007/s11229-012-0233-5.
- [33] Yanjing Wang & Jie Fan (2013): *Knowing That, Knowing What, and Public Communication: Public Announcement Logic with  $K_v$  Operators*. In: *Proceedings of IJCAI'13*, pp. 1139–1146. Available at <https://www.ijcai.org/Proceedings/13/Papers/173.pdf>.
- [34] Yanjing Wang & Jie Fan (2014): *Conditionally knowing what*. In: *Advances in Modal Logic Vol.10*, pp. 569–587. Available at <http://www.aiml.net/volumes/volume10/Wang-Fan.pdf>.
- [35] Yanjing Wang & Jeremy Seligman (2018): *When Names Are Not Commonly Known: Epistemic Logic with Assignments*. In: *Advances in Modal Logic 12*, College Publications, pp. 611–628. Available at <http://www.aiml.net/volumes/volume12/Wang-Seligman.pdf>.
- [36] Yanjing Wang, Yu Wei & Jeremy Seligman (2021): *Quantifier-free Epistemic Term-modal Logic with Assignments*. Manuscript, under submission.



# Dynamically Rational Judgment Aggregation: A Summary

Franz Dietrich

Paris School of Economics & CNRS

fd@franzdietrich.net

Christian List

LMU Munich

c.list@lmu.de

Judgment aggregation theory traditionally aims for collective judgments that are rational. So far, rationality has been understood in purely static terms: as coherence of judgments at a given time, where ‘coherence’ could for instance mean consistency, or completeness, or deductive closure, or combinations thereof. By contrast, this paper, which summarises results from Dietrich and List (2021), asks the novel question of whether collective judgments can be dynamically rational: whether they can respond well to new information, i.e., change rationally when information is learnt by everyone. Formally, we call a judgment aggregation rule dynamically rational with respect to a given revision operator if, whenever all individuals revise their judgments in light of some information (a proposition), then the new aggregate judgments are the old ones revised in light of this information. In short, aggregation and revision commute. A general impossibility theorem holds: as long as the propositions on the agenda are sufficiently interconnected, no judgment aggregation rule with standard properties is dynamically rational with respect to any revision operator satisfying mild conditions (familiar from belief revision theory). The theorem is a counterpart for dynamic rationality of known impossibility theorems for static rationality. Relaxation of the theorem’s conditions opens the door to interesting aggregation rules generating dynamically rational judgments, including certain premise-based rules, as we briefly discuss (see Dietrich and List 2020 for details).

## 1 Introduction

Suppose a group of individuals – say, a committee, expert panel, multi-member court, or other decision-making body – makes collective judgments on some propositions by aggregating its members’ individual judgments on those propositions. And now suppose some new information – in the form of the truth of some proposition – is learnt. All individuals rationally revise their judgments. Aggregating the new individual judgments yields new collective judgments. If the group is to be a rational agent, then it should incorporate new information rationally, and so the new aggregate judgments should coincide with the old ones revised in light of the information. Technically, this means that the operations of aggregation and revision commute: aggregating judgments and then revising the result yields the same as revising individual judgments and then aggregating.

In this paper, we investigate whether we can find reasonable aggregation rules that enable a group to achieve such dynamic rationality: aggregation rules which commute with reasonable revision methods. Surprisingly, this question has not been studied in the judgment-aggregation framework, where judgments are binary verdicts on some propositions: “yes”/“no”, “true”/“false”, “accept”/“reject”. (On judgment-aggregation theory, see List and Pettit 2002, Dietrich and List 2007, Nehring and Puppe 2010, Dokow and Holzman 2010, List and Puppe 2009.) The focus in judgment-aggregation theory has generally been on static rationality, namely on whether properties such as consistency, completeness, or deductive closure are preserved when individual judgments are aggregated into collective ones at a given point in time.<sup>1</sup>

---

<sup>1</sup>The revision of judgments has been investigated only in a different sense in judgment aggregation theory, namely in peer-disagreement contexts, where individuals do not learn a proposition but learn the judgments of others (Pettit 2006, List 2011).

By contrast, the question of dynamic rationality has received much attention in the distinct setting of probability aggregation, where judgments aren't binary but take the form of subjective probability assignments to the elements of some algebra. In that context, a mix of possibility and impossibility results has been obtained (e.g., Madansky 1964, Genest 1984, Genest et al. 1986, Dietrich 2010, 2019, Russell et al. 2015). These show that some familiar methods of aggregation – notably, the arithmetic averaging of probabilities – fail to commute with belief revision, while other methods – particularly geometric averaging – do commute with revision. An investigation of the parallel question in the case of binary judgments is therefore overdue.

We present a negative result: for a large class of familiar judgment aggregation rules, dynamic rationality is unachievable relative to a large class of reasonable judgment revision methods. However, if we relax some of our main theorem's conditions on the aggregation rule, dynamically rational aggregation becomes possible. In particular, "premise-based" aggregation can be dynamically rational relative to certain "premise-based" revision methods. This extended abstract focuses on the impossibility finding, for reasons of space. Possibilities are discussed in Dietrich and List (2021), which also contains all proofs.

## 2 The formal setup

We begin with the basic setup from judgment-aggregation theory (following List and Pettit 2002 and Dietrich 2007). We assume that there is a set of individuals who hold judgments on some set of propositions, and we are looking for a method of aggregating these judgments into resulting collective judgments. The key elements of this setup are the following:

**Individuals.** These are represented by a finite and non-empty set  $N$ . Its members are labelled  $1, 2, \dots, n$ . We assume  $n \geq 2$ .

**Propositions.** These are represented in formal logic. For our purposes, a thin notion of "logic" will suffice. Specifically, a logic,  $\mathbf{L}$ , is a non-empty set of formal objects called "propositions", which is endowed with two things: a negation operator, denoted  $\neg$ , so that, for every proposition  $p$  in  $\mathbf{L}$  there is also its negation  $\neg p$  in  $\mathbf{L}$ ; and a well-behaved notion of consistency, which specifies, for each set of propositions  $A \subseteq \mathbf{L}$ , whether  $S$  is consistent or inconsistent.<sup>2</sup> Standard propositional, predicate, modal, and conditional logics all fall under this definition, as do Boolean algebras.<sup>3</sup> A proposition  $p$  is *contradictory* if  $\{p\}$  is inconsistent, *tautological* if  $\{\neg p\}$  is inconsistent, and *contingent* if  $p$  is non-contradictory and non-tautological.

**Agenda.** The agenda is the set of those propositions from  $\mathbf{L}$  on which judgments are to be made. Formally, this is a finite non-empty subset  $X$  of  $\mathbf{L}$  which can be partitioned into proposition-negation pairs  $\{p, \neg p\}$ , abbreviated  $\{\pm p\}$ . Sometimes it is useful to make this partition explicit. We write  $\mathcal{X}$  to denote the set of these proposition-negation pairs of  $X$ . The elements of  $\mathcal{X}$  can be interpreted as the binary

---

<sup>2</sup>Well-behavedness is a three-part requirement: (i) any proposition-negation pair  $\{p, \neg p\}$  is inconsistent; (ii) any subset of any consistent set is still consistent; and (iii) the empty set is consistent, and any consistent set  $S$  has a consistent superset  $S'$  which contains a member of every proposition-negation pair  $\{p, \neg p\}$ .

<sup>3</sup>Readers familiar with probability theory could take  $\mathbf{L}$  to be a Boolean algebra on a non-empty set  $\Omega$  of possible worlds (e.g., the power set  $\mathbf{L} = 2^\Omega$ ), with negation defined as set-theoretic complementation and consistency of a set defined as non-empty intersection. The Boolean algebra could also be an abstract rather than set-theoretic Boolean algebra.

issues under consideration. Then the agenda  $X$  is their disjoint union, formally  $X = \cup_{Z \in \mathcal{Z}} Z$ . Throughout this paper, we assume that double-negations cancel out in agenda propositions.<sup>4</sup>

Our focus will be on agendas satisfying a non-triviality condition. To define it, call a set of propositions *minimal inconsistent* if it is inconsistent but all its proper subsets are consistent. Proposition-negation pairs of the form  $\{p, \neg p\}$  (with  $p$  contingent) are minimal inconsistent, and so are sets of the form  $\{p, q, \neg(p \wedge q)\}$  (with  $p$  and  $q$  contingent), where “ $\wedge$ ” stands for logical conjunction (“and”). We call an agenda *non-simple* if it has at least one minimal inconsistent subset of size greater than two. An example of a non-simple agenda is the set  $X = \{\pm p, \pm(p \rightarrow q), \pm q\}$ , where  $p$  might be the proposition “Current atmospheric CO2 is above 407 ppm”,  $p \rightarrow q$  might be the proposition “If current atmospheric CO2 is above 407 ppm, then the Arctic iceshield will melt by 2050”, and  $q$  might be the proposition “The Arctic iceshield will melt by 2050”. The conditional  $p \rightarrow q$  can be formalized in standard propositional logic or in a suitable logic for conditionals. A three-member minimal inconsistent subset of this agenda is  $\{p, p \rightarrow q, \neg q\}$ .

**Judgments.** Each individual’s (and subsequently the group’s) judgments on the given propositions are represented by a judgment set, which is a subset  $J \subseteq X$ , consisting of all those propositions from  $X$  that its bearer “accepts” (e.g., affirms or judges to be true). A judgment set  $J$  is

- *complete* if it contains a member of each proposition-negation pair from  $X$ ,
- *consistent* if it is a consistent set in the sense of the given logic, and
- *classically rational* if it has both of these properties.

We write  $\mathcal{J}$  to denote the set of all classically rational judgment sets on the agenda  $X$ . A list of judgment sets  $(J_1, \dots, J_n)$  across the individuals in  $N$  is called a *profile* (of individual judgment sets).

**Aggregation rule.** A (*judgment*) *aggregation rule* is a function,  $F$ , which maps each profile  $(J_1, \dots, J_n)$  in some domain  $\mathcal{D}$  of admissible profiles (often  $\mathcal{D} = \mathcal{J}^n$ ) to a collective judgment set  $J = F(J_1, \dots, J_n)$ . A standard example is majority rule, which is defined as follows: for each  $(J_1, \dots, J_n) \in \mathcal{J}^n$ ,

$$F(J_1, \dots, J_n) = \{p \in X : |\{i : p \in J_i\}| > n/2\}.$$

A typical research question in judgment aggregation theory is whether we can find aggregation rules that satisfy certain requirements of democratic responsiveness to the individual judgments and collective rationality.

### 3 Judgment revision

The idea we wish to capture is that whenever any individual (or subsequently the group) learns some new information, in the form of the truth of some proposition, this individual (or the group) must incorporate the learnt information in the judgments held – an idea familiar from belief revision theory in the tradition of Alchourrón, Gärdenfors and Makinson (1985) (see also Rott 2001 and Peppas 2008). Our central concept is that of a *judgment revision operator*. This is a function which assigns to any pair  $(J, p)$  of an initial judgment set  $J \subseteq X$  and a learnt proposition  $p \in X$  a new judgment set  $J|p$ , the revised judgment set, given  $p$ . Formally, the revision operator is any function from  $2^X \times X$  to  $2^X$ . We call it *regular* if it satisfies the following two minimal conditions:

<sup>4</sup>To be precise, henceforth, by the negation of any proposition  $q \in X$  we shall mean the *agenda-internal* negation of  $q$ , i.e., the opposite proposition in the binary issue  $\{p, \neg p\}$  to which  $q$  belongs. This is logically equivalent to the ordinary negation of  $q$  and will again be denoted  $\neg q$ , for simplicity. This convention ensures that  $\neg\neg q = q$ .

- (i) it is *successful*, i.e.,  $p \in J|p$  for any pair  $(J, p)$  (“accept what you learn”), and
- (ii) it is *conservative*, i.e.,  $J|p = J$  for any pair  $(J, p)$  such that  $p \in J$  (“no news, no change”).

We further call a revision operator *rationality-preserving* if whenever  $J \in \mathcal{J}$ , we have  $J|p \in \mathcal{J}$  for all non-contradictory propositions  $p \in X$ . These definitions are well-illustrated by the class of distance-based revision operators, familiar from belief revision theory. Such operators require that when a judgment set is revised in light of some new information, the post-revision judgments remain as “close” as possible to the pre-revision judgments, subject to the constraint that the learnt information be incorporated and no inconsistencies be introduced. Different distance-based operators spell out the notion of “closeness” in different ways (different metrics have been introduced in the area of judgment aggregation by Konieczny and Pino-Pérez 2002 and Pigozzi 2006).

## 4 Can aggregation and revision commute?

We are now ready to turn to this paper’s question. As noted, we would ideally want any decision-making group to employ a judgment aggregation rule and a revision operator that generate the same collective judgments irrespective of whether revision takes place before or after aggregation. This requirement (an analogue of the classic “external Bayesianity” condition in probability aggregation theory, as in Madansky 1964, Genest 1984, and Genest et al. 1986) is captured by the following condition on the aggregation rule  $F$  and the revision operator  $|$ :

**Dynamic rationality.** For any profile  $(J_1, \dots, J_n)$  in the domain of  $F$  and any learnt proposition  $p \in X$  where the revised profile  $(J_1|p, \dots, J_n|p)$  is also in the domain of  $F$ ,  $F(J_1|p, \dots, J_n|p) = F(J_1, \dots, J_n)|p$ .

To see that this condition is surprisingly hard to satisfy, consider an example. Suppose a three-member group is making judgments on the agenda  $X = \{\pm p, \pm(p \rightarrow q), \pm q\}$ , where  $p \rightarrow q$  is understood as a subjunctive conditional. That is, apart from the subsets of  $X$  that include a proposition-negation pair, the only inconsistent subset of  $X$  is  $\{p, p \rightarrow q, \neg q\}$ .<sup>5</sup> Suppose, further, members’ initial judgments are

	Before learning $p$			After learning $p$		
	$p$ ?	$p > q$ ?	$q$ ?	$p$ ?	$p > q$ ?	$q$ ?
Individual 1	N	N	Y	Y	N	Y
Individual 2	N	Y	N	Y	Y	Y
Individual 3	N	N	N	Y	N	N
Collective	N	N	N	Y	N	Y

Table 1: Majority rule and revision

as shown on the left-hand side Table 1, where “yes” stands for the acceptance of a proposition and “no” for the acceptance of its negation. Suppose the group uses majority rule, and the revision operator is based on the Hamming distance, with some tie-breaking provision such that, in the case of a tie, one is more ready to change one’s judgment on  $p$  or  $p \rightarrow q$  (which represent “premises”) than on  $q$

<sup>5</sup>This subjunctive understanding of  $p \rightarrow q$  contrasts with the material one, where  $p \rightarrow q$  is understood less realistically as  $\neg p \vee q$ . On the material understanding, the subsets  $\{p, \neg(p \rightarrow q), q\}$ ,  $\{\neg p, \neg(p \rightarrow q), q\}$ , and  $\{\neg p, \neg(p \rightarrow q), \neg q\}$  would also be deemed inconsistent.

(which represents a “conclusion”). If the individuals learn the truth of  $p$  and revise their judgments, they arrive at the post-revision judgments shown on the right-hand side of Table 1. The aggregate (majority) judgments, before and after information arrival, are as shown in Table 1. Crucially, the post-information group judgment set,  $\{p, \neg(p \rightarrow q), q\}$ , differs from the revision in light of  $p$  of the pre-information group judgment set, because  $\{\neg p, \neg(p \rightarrow q), \neg q\} \uparrow p = \{p, \neg(p \rightarrow q), \neg q\}$ . That is, the group replaces  $\neg q$  with  $q$  in its judgment set, although learning  $p$  did not force the group to revise its position on  $q$  (recall that  $\{p, \neg(p \rightarrow q), \neg q\}$  is perfectly consistent, given that  $\rightarrow$  is a subjunctive conditional). Thus the group’s (majority) judgment set does not evolve rationally.

At first sight, one might think that this problem is just an artifact of majority rule or our specific distance-based revision operator, or that it is somehow unique to our example. However, the following formal result – a simplified (‘anonymous’) version of our impossibility theorem – shows that the problem is more general. Define a *uniform quota rule*, with acceptance threshold  $m \in \{1, \dots, n\}$ , as the aggregation rule with domain  $\mathcal{J}^n$  such that, for each  $(J_1, \dots, J_n) \in \mathcal{J}^n$ ,

$$F(J_1, \dots, J_n) = \{p \in X : |\{i : p \in J_i\}| \geq m\}.$$

Majority rule is a special case of a uniform quota rule, namely the one where  $m$  is the smallest integer greater than  $n/2$ . We have:

**Theorem 1** *If the agenda  $X$  is non-simple, then no uniform quota rule whose threshold is not the unanimity threshold  $n$  is dynamically rational with respect to any regular rationality-preserving revision operator.*

In short, replacing majority rule with some other uniform quota rule with threshold less than  $n$  wouldn’t solve our problem of dynamic irrationality, and neither would replacing our distance-based revision operator with some other regular rationality-preserving revision operator. In fact, the problem generalizes further, as shown in the next section.

## 5 A general impossibility theorem

We will now abstract away from the details of any particular aggregation rule, and suppose instead we are looking for an aggregation rule  $F$  that satisfies the following general conditions:

**Universal domain:** The domain of admissible inputs to the aggregation rule  $F$  is the set of all classically rational profiles, i.e.,  $\mathcal{D} = \mathcal{J}^n$ .

**Non-imposition:**  $F$  does not always deliver the same antecedently fixed output judgment set  $J$ , irrespective of the individual inputs, i.e.,  $F$  is not a constant function.

**Monotonicity:** Additional individual support for an accepted proposition does not overturn the proposition’s acceptance, i.e., for any profile  $(J_1, \dots, J_n) \in \mathcal{D}$  and any proposition  $p \in F(J_1, \dots, J_n)$ , if any  $J_i$  not containing  $p$  is replaced by some  $J'_i$  containing  $p$  and the modified profile  $(J_1, \dots, J'_i, \dots, J_n)$  remains in  $\mathcal{D}$ , then  $p \in F(J_1, \dots, J'_i, \dots, J_n)$ .

**Non-oligarchy:** There is no non-empty set of individuals  $M \subseteq N$  (a set of “oligarchs”) such that, for every profile  $(J_1, \dots, J_n) \in \mathcal{D}$ ,  $F(J_1, \dots, J_n) = \bigcap_{i \in M} J_i$ .

**Systematicity:** The collective judgment on each proposition is determined fully and neutrally by individual judgments on that proposition. Formally, for any propositions  $p, p' \in X$  and any profiles  $(J_1, \dots, J_n), (J'_1, \dots, J'_n) \in \mathcal{D}$ , if, for all  $i \in N$ ,  $p \in J_i \Leftrightarrow p' \in J'_i$ , then  $p \in F(J_1, \dots, J_n) \Leftrightarrow p' \in F(J'_1, \dots, J'_n)$ .

Why are these conditions initially plausible? The reason is that, for each of them, a violation would entail a cost. Violating universal domain would mean that the aggregation rule is not fully robust to pluralism in its inputs; it would be undefined for some classically rational judgment profiles. Violating non-imposition would mean that the collective judgments are totally unresponsive to the individual judgments, which is completely undemocratic. Violating monotonicity could make the aggregation rule erratic in some respect: an individual could come to accept a particular collectively accepted proposition and thereby overturn its acceptance. Violating non-oligarchy would mean two things. First, the collective judgments would depend only on the judgments of the “oligarchs”, which is undemocratic (unless  $M = N$ ); and second, the collective judgments would be incomplete with respect to any binary issue on which there is the slightest disagreement among the oligarchs, which would lead to widespread indecision (except when  $M$  is singleton, so that the rule is dictatorial). Violating systematicity, finally, would mean that the collective judgment on each proposition is no longer determined as a proposition-independent function of individual judgments on that proposition. It may then either depend on individual judgments on other propositions too (a lack of propositionwise independence), or the pattern of dependence may vary from proposition to proposition (a lack of neutrality). Systematicity – the conjunction of propositionwise independence and neutrality – is the most controversial condition among the five. But it is worth noting that it is satisfied by majority rule and all uniform quota rules. Indeed, majority rule and uniform quota rules (except the unanimity rule) satisfy all five conditions.

Our main theorem shows that, for non-simple agendas, the present five conditions are incompatible with dynamic rationality:

**Theorem 2** *If the agenda  $X$  is non-simple, then no aggregation rule satisfying universal domain, non-imposition, monotonicity, non-oligarchy, and systematicity is dynamically rational with respect to any regular rationality-preserving revision operator.*

Interestingly, Theorem 2 does not impose any condition of static rationality. The theorem does not require that collective judgment sets are consistent or complete or deductively closed. The impossibility of dynamic inconsistency is thus independent of classic impossibilities of static rationality. In fact, Theorem 2 would continue to hold if its condition of dynamic rationality were replaced by static rationality in the form of consistency and completeness of collective judgment sets.

By Theorem 2, the problem identified by Theorem 1 is not restricted to uniform quota rules, but extends to all aggregation rules satisfying our conditions. Moreover, since practically all non-trivial agendas are non-simple, the impossibility applies very widely.

The natural follow-up question is that of whether any of the conditions in the theorem is redundant, i.e., could be dropped, and if not what sort of dynamically rational aggregation rules become possible after dropping any of these conditions. This question goes beyond the scope of this summary and is treated in Dietrich and List (2021). Four remarks should however be given:

Firstly, none of the theorem’s conditions on the aggregation rule, the revision operator, or the agenda is redundant. That is, whenever we drop the agenda condition (non-simplicity) or any one of the aggregation conditions (universal domain, non-imposition, monotonicity, non-oligarchy, and systematicity) or any of the revision conditions (successfulness, conservativeness, and rationality preservation), there exist dynamically rational aggregation rules such that the remaining conditions hold.

Secondly, abandoning exactly one condition on the aggregation rule leads to rather degenerate dynamically rational possibilities, in the form of ‘peculiar’ aggregation rules and/or revision operators (with the exception of universal domain, whose relaxation allows for interesting dynamically rational possibilities). One of the conditions on aggregation seems very strong: systematicity. An important difference between static and dynamic rationality is that dropping systematicity or even independence makes it easy (indeed, too easy) to satisfy static rationality – for instance by using distance-based rules or prioritarian rules or scoring rules – whereas dynamic rationality remains hard to achieve without systematicity, as illustrated by the degenerate nature of the non-systematic escape route constructed in Dietrich and List (2021). It thus seems inappropriate to blame systematicity for being the main culprit for the impossibility of dynamic rationality.

Thirdly, let us give examples of dynamically rational aggregation rules that become possible if we give up any one of the three conditions on the revision operator while preserving all other conditions on revision or aggregation.

- *Non-successful revision.* Consider the constant revision operator, defined by

$$J|p = J \text{ for all } (J, p).$$

This operator is only conservative and rationality-preserving. All aggregation rules are trivially dynamically rational with respect to it.

- *Non-conservative revision.* For each proposition  $p \in X$ , fix a judgment set  $J_p$  which contains  $p$  and moreover is rational (i.e., in  $\mathcal{J}$ ) as long as  $p$  is non-contradictory. Consider the revision operator given by

$$J|p = J_p \text{ for all } (J, p).$$

This operator is only successful and rationality-preserving. As one can show, every unanimity-preserving aggregation rule is dynamically rational with respect to it.

- *Non-rationality-preserving revision.* Consider a revision operator such that  $J|p$  is  $J$  whenever  $p \in J$  and is any irrational judgment set containing  $p$  if  $p \notin J$ . This operator is only conservative and successful. As one can show, every aggregation rule satisfying universal domain and propositionwise unanimity preservation is dynamically rational with respect to it. Here, propositionwise unanimity-preservation means that, for all profiles  $(J_1, \dots, J_n)$  in the domain and all propositions  $p \in X$ , if  $p \in J_i$  for all  $i$ , then  $p \in F(J_1, \dots, J_n)$ .

Finally, on a more positive note, in Dietrich and List (2021) we explore an interesting class of dynamically rational aggregation rules, which simultaneously relax multiple of Theorem 2’s conditions on aggregation and revision, notably systematicity. In a nutshell, *premise-based* aggregation rules are dynamically rational with respect to *premise-based* revision operators. Presenting these rules goes beyond the scope of this summary.

Proofs of theorems and other technical details are given in Dietrich and List (2021).

## References

- Alchourrón, C. E., Gärdenfors, P., and Makinson, D. (1985): *On the logic of theory change: Partial meet contraction and revision functions.* *Journal of Symbolic Logic* 50(2), pp. 510–530. DOI: 10.2307/2274239

- Dietrich, F. (2007): *A generalised model of judgment aggregation*. *Social Choice and Welfare* 28(4), pp. 529–565. DOI: 10.1007/s00355-006-0187-y
- Dietrich, F. (2010): *Bayesian group belief*. *Social Choice and Welfare* 35(4), pp. 595–626. DOI: 10.1007/s00355-010-0453-x
- Dietrich, F. (2019): *A theory of Bayesian groups*. *Noûs* 53(3), pp. 708–736. DOI: 10.1111/nous.12233
- Dietrich, F., and List, C. (2007): *Arrow's theorem in judgment aggregation*. *Social Choice and Welfare* 29(1), pp. 19–33. DOI: 10.1007/s00355-006-0196-x
- Dietrich, F., and List, C. (2021): *Dynamically Rational Judgment Aggregation*. Working paper, see <https://philpapers.org/rec/DIEDRJ>
- Dokow, E., and Holzman, R. (2010): *Aggregation of binary evaluations*. *Journal of Economic Theory* 145(2), pp. 495–511. DOI: 10.1016/j.jet.2007.10.004
- Genest, C. (1984): *A characterization theorem for externally Bayesian groups*. *Annals of Statistics* 12(3), pp. 1100–1105. DOI: 10.1214/aos/1176346726
- Genest, C., McConway, K. J., and Schervish, M. J. (1986): *Characterization of externally Bayesian pooling operators*. *Annals of Statistics* 14(2), pp. 487–501. DOI: 10.1007/BF02562628
- Konieczny, S., and Pino-Pérez, R. (2002): *Merging information under constraints: A logical framework*. *Journal of Logic and Computation* 12(5), pp. 773–808. DOI: 10.1093/logcom/12.5.773
- List, C. (2011): *Group Communication and the Transformation of Judgments: An Impossibility Result*. *Journal of Political Philosophy* 19(1), pp. 1–27. DOI: 10.1111/j.1467-9760.2010.00369.x
- List, C., and Pettit, P. (2002): *Aggregating sets of judgments: An impossibility result*. *Economics and Philosophy* 18(1), pp. 89–110. DOI: 10.1023/B:SYNT.0000029950.50517.59
- List, C., and Pettit, P. (2011): *Group Agency: The Design, Possibility, and Status of Corporate Agents*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199591565.001.0001
- List, C., and Puppe, C. (2009): *Judgment aggregation: A survey*. In P. Anand, C. Puppe, and P. Pattanaik, *Oxford Handbook of Rational and Social Choice*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199290420.001.0001
- Madansky, A. (1964): *Externally Bayesian Groups*. Technical Report RM-4141-PR, RAND Corporation.
- Nehring, K., and Puppe, C. (2010): *Abstract Arrovian aggregation*. *Journal of Economic Theory* 145(2), pp. 467–494. DOI: 10.1016/j.jet.2010.01.010
- Peppas, P. (2008): *Belief Revision*. In F. van Harmelen, V. Lifschitz and B. Porter, *Handbook of Knowledge Representation*, Elsevier, pp. 317–359.
- Pettit, P. (2006): *When to defer to majority testimony – and when not*. *Analysis* 66(3), pp. 179–187.
- Pigozzi, G. (2006): *Belief merging and the discursive dilemma: An argument-based account to paradoxes of judgment aggregation*. *Synthese* 152(2), pp. 285–298. DOI: 10.1007/s11229-006-9063-7
- Rott, H. (2001): *Change, Choice and Inference: A Study of Belief Revision and Non-monotonic Reasoning*. Oxford: Oxford University Press.
- Russell, J. S., Hawthorne, J., and Buchak, L. (2015): *Groupthink*. *Philosophical Studies* 172(5), pp. 1287–1309. DOI: 10.1007/s11098-014-0350-8

# Deliberation and Epistemic Democracy

Huihui Ding

CY Cergy Paris University

Marcus Pivato

CY Cergy Paris University

We study the effects of deliberation on epistemic social choice, in two settings. In the first setting, the group faces a binary epistemic decision analogous to the Condorcet Jury Theorem. In the second setting, group members have probabilistic beliefs arising from their private information, and the group wants to aggregate these beliefs in a way that makes optimal use of this information. During deliberation, each agent discloses private information to persuade the other agents of her current views. But her views may also evolve over time, as she learns from other agents. This process will improve the performance of the group, but only under certain conditions; these involve the nature of the social decision rule, the group size, and also the presence of neutral agents whom the other agents try to persuade.



# No Finite Model Property for Logics of Quantified Announcements

Hans van Ditmarsch

Open University of the Netherlands  
Heerlen, The Netherlands

`hans.vanditmarsch@ou.nl`

Tim French

University of Western Australia  
Perth, Australia

`tim.french@uwa.edu.au`

Rustam Galimullin

University of Bergen  
Bergen, Norway

`rustam.galimullin@uib.no`

Quantification over public announcements shifts the perspective from reasoning strictly about the results of a particular announcement to reasoning about the existence of an announcement that achieves some certain epistemic goal. Depending on the type of the quantification, we get different formalisms, the most known of which are arbitrary public announcement logic (*APAL*), group announcement logic (*GAL*), and coalition announcement logic (*CAL*). It has been an open question whether the logics have the finite model property, and in the paper we answer the question negatively. We also discuss how this result is connected to other open questions in the field.

## 1 Introduction

One of the most well-known ways to introduce communication into the multi-agent epistemic logic (*EL*) [18] is by extending the logic with public announcements, which, being dynamic operators, model the situation of all agents publicly and simultaneously receiving the same piece of information. Epistemic logic with such operators is called *public announcement logic (PAL)* [21], and it extends *EL* with constructs  $[\varphi]\psi$  that mean ‘after public announcement of  $\varphi$ ,  $\psi$ ’.

A natural way to generalise *PAL*, with epistemic planning [5] flavour to it, is to consider quantification over public announcements<sup>1</sup>. Such an extension allows us to reason about the *existence* of an announcement, or a sequence thereof, that reaches certain epistemic goal. There are several ways to quantify over public announcements, and we call the resulting logics *quantified public announcement logics (QPAL)*. In the paper, we consider the Big Three of *QPAL*’s, namely *arbitrary public announcement logic (APAL)*, *group announcement logic (GAL)*, and *coalition announcement logic (CAL)*.

*APAL* [4] extends *PAL* with constructs  $\Box\varphi$  that are read as ‘after any public announcement,  $\varphi$  is true’. While *APAL* modalities do not take into account who announces a formula or whether the formula can be truthfully announced by any of the agents in a system, *GAL* constructs  $[G]\varphi$  restrict the quantification to the formulas that agents actually know [1]. Thus,  $[G]\varphi$  is read as ‘after any joint truthful announcement by agents from group  $G$ ,  $\varphi$  is true’. ‘Truthful’ here denotes the fact that the agents know the formulas they announce. Finally, *CAL* is somewhat similar to *GAL* with a crucial difference that agents outside of the selected group can also make a simultaneous announcement [2, 12]. *CAL* extends *PAL* with constructs  $\langle\langle G \rangle\rangle\varphi$  that are read as ‘whatever agents from  $G$  announce, there is a simultaneous announcement by the agents outside of  $G$ , such that  $\varphi$  is true after the joint announcement’. Given that the modalities of *CAL* were inspired by coalition logic [20], it is not surprising that they are game-theoretic in nature, and, in particular, express the property of  $\beta$ -effectivity<sup>2</sup>.

One of the pressing open questions of *QPAL* is whether they have the finite model property (FMP).

---

<sup>1</sup>See a recent survey [6] for an overview.

<sup>2</sup>The dual  $\langle\langle G \rangle\rangle\varphi$  expresses  $\alpha$ -effectivity.

*FMP: A logic has the FMP iff every formula of the logic that is true in some model is also true in a finite model.*

It is a standard result that *EL* has the FMP [14]. As the reader may have already guessed, after having read the title of the paper, we show that *APAL*, *GAL*, and *CAL* do not have the FMP.

The result is important for a couple of reasons. First, it tells us something about the expressivity of *QPAL*'s. In particular, all of *APAL*, *GAL*, and *CAL*, are so expressive that they can force infinite models, i.e. there are formulas of the languages that can only be true on infinite structures.

Second, the lack of FMP sheds light on a connected open problem of finding finitary axiomatisations of *QPAL*'s. It is known [22] that

*Finitary axiomatisation and FMP imply decidability. (\*)*

We also know that *APAL*, *GAL*, and *CAL* are all undecidable [3]. The corresponding proof is quite complex, but ultimately it presents for each logic a formula over a parameterised set of tiles, such that the formula is satisfiable if and only if the set of tiles can tile the plane. The construction of the tiling is not explicit, in the sense that there is no one-to-one correspondence of worlds to unique points on the plane. Rather, given a model that satisfies the formula, a series of finite tilings is created in such a way that it guarantees the existence of an infinite tiling. It is not clear how to extract an argument against FMP from the undecidability proof, or whether it is possible. In any way, our approach presented in Section 3 is simpler and more elegant.

Another reason why we cannot get the lack of the FMP for free from the undecidability is that (\*) requires the axiomatisations at hand to be finitary. To the best of our knowledge, none of *APAL*, *GAL*, and *CAL* have a known finitary axiomatisation, although the first two are recursively axiomatisable, and providing any axiomatisation of *CAL* is an open problem. However, (\*) cannot be relaxed to requiring only recursive axiomatisations instead of finitary ones [22, 16].

On the whole, the relation between the FMP, finitary axiomatisations, and decidability is not trivial, and (\*) can be satisfied in various ways. For example, *EL* has all three properties, while there are modal logics that are decidable and recursively axiomatisable [19, 8], or finitely axiomatisable, decidable, but lack the FMP [9, 10, 7], or not finitely axiomatisable, undecidable, but have the FMP [11, 17, 15], or have none of the three properties [15], and so on.

Due to the undecidability of *QPAL*'s, up until now there was a hope that if the logics have the FMP, then we will be able to conclude they are not finitely axiomatisable. We show that, in fact, the logics do not have the FMP and the problem of the existence of finitary axiomatisations remains.

In what follows, we formally introduce *APAL*, *GAL*, *CAL*, and some technical notions in Section 2. In Section 3 we prove that *APAL* does not have the FMP. The strategy of the proof is such that we, first, present a formula that is true in an infinite model, and then claim that the formula cannot be true in any finite model. The results for *GAL* and *CAL* follow as a corollary. We conclude and discuss further research in Section 4.

## 2 Quantified public announcement logics

Given are a countable (finite or countably infinite) set of *agents*  $A$  and a countably infinite set of *propositional variables*  $P$  (a.k.a. *atoms*, or *variables*). In what follows,  $G \subseteq A$ , and we denote  $A \setminus G$  as  $\overline{G}$ .

### 2.1 Syntax

We start with defining the logical language and some crucial syntactic notions.

**Definition 1 (Language)** *The language of quantified public announcement logic is defined as follows, where  $a \in A$  and  $p \in P$ .*

$$\mathcal{L}(A, P) \ni \varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \psi) \mid K_a\varphi \mid [\varphi]\varphi \mid Q\varphi \quad \dashv$$

Other propositional connectives are defined by abbreviation, and, unless ambiguity results, we often omit parentheses occurring in formulas. We also often omit one or both of the parameters  $A$  and  $P$  in  $\mathcal{L}(A, P)$ , and write  $\mathcal{L}(P)$  or  $\mathcal{L}$ . Formulas are denoted  $\varphi, \psi$ , possibly primed as in  $\varphi', \varphi'', \dots, \psi', \dots$ . Depending on which form the quantifier  $Q$  takes —  $\square$ ,  $[G]$ , or  $[[G]]$ , where  $G \subseteq A$  — we distinguish  $\mathcal{L}_{\text{apal}}$ ,  $\mathcal{L}_{\text{gal}}$ , and  $\mathcal{L}_{\text{cal}}$ , respectively. We also distinguish the language  $\mathcal{L}_{\text{el}}$  of *epistemic logic* (without the constructs  $[\varphi]\varphi$  and  $Q\varphi$ ).

For  $K_a\varphi$  read ‘agent  $a$  knows  $\varphi$ ’. For  $[\varphi]\psi$ , read ‘after public announcement of  $\varphi$ ,  $\psi$ ’. For  $\square\varphi$ , read ‘after any announcement,  $\varphi$  (is true)’. For  $[G]\varphi$ , read ‘after any joint announcement by agents from  $G$ ,  $\varphi$  is true’. And for  $[[G]]\varphi$  read ‘for each announcement by agents from  $G$ , there is a counter-announcement by the remaining agents, such that  $\varphi$  is true after the joint simultaneous announcement’. The dual modalities are defined by abbreviation:  $\widehat{K}_a\varphi := \neg K_a\neg\varphi$ ,  $\langle\varphi\rangle\psi := \neg[\varphi]\neg\psi$ ,  $\diamond\varphi := \neg\square\neg\varphi$ ,  $\langle G\rangle\varphi := \neg[[G]]\neg\varphi$ , and  $[[G]]\varphi := \neg[[G]]\neg\varphi$ .

The set of propositional variables that occur in a given formula  $\varphi$  is denoted  $\text{var}(\varphi)$  (where one that does not occur in  $\varphi$  is called a *fresh variable*), its *modal depth*  $d(\varphi)$  is the maximum nesting of  $K_a$  modalities, and its *quantifier depth*  $D(\varphi)$  is the maximum nesting of  $Q \in \{\square, [G], [[G]]\}$  modalities. These notions are inductively defined as follows.

- $\text{var}(p) = \{p\}$ ,  $\text{var}(\neg\varphi) = \text{var}(K_a\varphi) = \text{var}(Q\varphi) = \text{var}(\varphi)$ ,  $\text{var}(\varphi \wedge \psi) = \text{var}([\varphi]\psi) = \text{var}(\varphi) \cup \text{var}(\psi)$ ;
- $D(p) = 0$ ,  $D(\neg\varphi) = D(K_a\varphi) = D(\varphi)$ ,  $D(\varphi \wedge \psi) = D([\varphi]\psi) = \max\{D(\varphi), D(\psi)\}$ ,  $D(Q\varphi) = D(\varphi) + 1$ ;
- $d(p) = 0$ ,  $d(\neg\varphi) = d(Q\varphi) = d(\varphi)$ ,  $d(\varphi \wedge \psi) = \max\{d(\varphi), d(\psi)\}$ ,  $d([\varphi]\psi) = d(\varphi) + d(\psi)$ ,  $d(K_a\varphi) = d(\varphi) + 1$ .

## 2.2 Structures

We consider the following structures and structural notions in this work.

**Definition 2 (Model)** *An (epistemic) model  $M = (S, \sim, V)$  consists of a non-empty domain  $S$  (or  $\mathcal{D}(M)$ ) of states (or ‘worlds’), an accessibility function  $\sim: A \rightarrow \mathcal{P}(S \times S)$ , where each  $\sim_a$  is an equivalence relation, and a valuation  $V: P \rightarrow \mathcal{P}(S)$ , where each  $V(p)$  represents the set of states where  $p$  is true. For  $s \in S$ , a pair  $(M, s)$ , for which we write  $M_s$ , is a pointed (epistemic) model.  $\dashv$*

We will abuse the language and also call  $M_s$  a model. We will occasionally use the following disambiguating notation: if  $M$  is a model,  $S^M$  is its domain,  $\sim_a^M$  the accessibility relation for an agent  $a$ , and  $V^M$  its valuation.

**Definition 3 (Bisimulation)** *Let  $M = (S, \sim, V)$  and  $M' = (S', \sim', V')$  be epistemic models. A non-empty relation  $\mathfrak{R} \subseteq S \times S'$  is a bisimulation if for every  $(s, s') \in \mathfrak{R}$ ,  $p \in P$ , and  $a \in A$  the conditions **atoms**, **forth** and **back** hold.*

- **atoms:**  $s \in V(p)$  iff  $s' \in V'(p)$ .
- **forth:** for every  $t \sim_a s$  there exists  $t' \sim'_a s'$  such that  $(t, t') \in \mathfrak{R}$ .

- **back**: for every  $t' \sim'_a s'$  there exists  $t \sim_a s$  such that  $(t, t') \in \mathfrak{R}$ .

If there exists a bisimulation  $\mathfrak{R}$  between  $M$  and  $M'$  such that  $(s, s') \in \mathfrak{R}$ , then  $M_s$  and  $M'_s$  are bisimilar, notation  $M_s \Leftrightarrow M'_s$  (or  $\mathfrak{R} : M_s \Leftrightarrow M'_s$ , to be explicit about the bisimulation).

Let  $Q \subseteq P$ . A relation  $\mathfrak{R}$  between  $M$  and  $M'$  satisfying **atoms** for all  $p \in Q$ , and **forth** and **back**, is a  $Q$ -bisimulation (a bisimulation restricted to  $Q$ ). The notation for  $Q$ -restricted bisimilarity is  $\Leftrightarrow_Q$ .  $\dashv$

The notion of  $n$ -bisimulation, for  $n \in \mathbb{N}$ , is given by defining relations  $\mathfrak{R}^0 \supseteq \dots \supseteq \mathfrak{R}^n$ .

**Definition 4 (n-Bisimulation)** Let  $M = (S, \sim, V)$  and  $M' = (S', \sim', V')$  be epistemic models, and let  $n \in \mathbb{N}$ . A non-empty relation  $\mathfrak{R}^0 \subseteq S \times S'$  is a 0-bisimulation if **atoms** holds for pair  $(s, s') \in \mathfrak{R}$ . Then, a non-empty relation  $\mathfrak{R}^{n+1} \subseteq S \times S'$  is a  $(n+1)$ -bisimulation if for all  $p \in P$  and  $a \in A$ :

- $(n+1)$ -**forth**: for every  $t \sim_a s$  there exists  $t' \sim'_a s'$  such that  $(t, t') \in \mathfrak{R}^n$ ;
- $(n+1)$ -**back**: for every  $t' \sim'_a s'$  there exists  $t \sim_a s$  such that  $(t, t') \in \mathfrak{R}^n$ .

Similarly to  $Q$ -bisimulations we define  $Q$ - $n$ -bisimulations, wherein **atoms** is only required for  $p \in Q \subseteq P$ ;  $n$ -bisimilarity is denoted  $M_s \Leftrightarrow^n M'_s$ , and  $Q$ - $n$ -bisimilarity is denoted  $M_s \Leftrightarrow_Q^n M'_s$ .  $\dashv$

### 2.3 Semantics

We continue with the semantics of our logic. Let  $\mathcal{L}_{el}^G := \{\bigwedge_{i \in G} K_i \varphi_i \mid \varphi_i \in \mathcal{L}_{el}\}$  be the set of all announcement by group  $G$ .

**Definition 5 (Semantics)** The interpretation of formulas in  $\mathcal{L}_{apal} \cup \mathcal{L}_{gal} \cup \mathcal{L}_{cal}$  on epistemic models is defined by induction on formulas.

Assume an epistemic model  $M = (S, \sim, V)$ , and  $s \in S$ .

$$\begin{aligned}
M_s \models p & \quad \text{iff } s \in V(p) \\
M_s \models \neg \varphi & \quad \text{iff } M_s \not\models \varphi \\
M_s \models \varphi \wedge \psi & \quad \text{iff } M_s \models \varphi \text{ and } M_s \models \psi \\
M_s \models K_a \varphi & \quad \text{iff for all } t \in S : s \sim_a t \text{ implies } M_t \models \varphi \\
M_s \models [\varphi] \psi & \quad \text{iff } M_s \models \varphi \text{ implies } M_s^\varphi \models \psi \\
M_s \models \Box \psi & \quad \text{iff for all } \varphi \in \mathcal{L}_{el} : M_s \models [\varphi] \psi \\
M_s \models [G] \psi & \quad \text{iff for all } \varphi_G \in \mathcal{L}_{el}^G : M_s \models [\varphi_G] \psi \\
M_s \models \llbracket G \rrbracket \psi & \quad \text{iff for all } \varphi_G \in \mathcal{L}_{el}^G \text{ there is } \chi_{\overline{G}} \in \mathcal{L}_{el}^{\overline{G}} : M_s \models \varphi_G \rightarrow \langle \varphi_G \wedge \chi_{\overline{G}} \rangle \psi
\end{aligned}$$

where  $\llbracket \varphi \rrbracket_M := \{s \in S \mid M_s \models \varphi\}$ ; and where epistemic model  $M^\varphi = (S', \sim', V')$  is such that:  $S' = \llbracket \varphi \rrbracket_M$ ,  $\sim'_a = \sim_a \cap (\llbracket \varphi \rrbracket_M \times \llbracket \varphi \rrbracket_M)$ , and  $V'(p) := V(p) \cap \llbracket \varphi \rrbracket_M$ . For  $(M^\varphi)^\psi$  we may write  $M^{\varphi\psi}$ . Formula  $\varphi$  is valid on model  $M$ , notation  $M \models \varphi$ , if for all  $s \in S$ ,  $M_s \models \varphi$ . Formula  $\varphi$  is valid, notation  $\models \varphi$ , if for all  $M$ ,  $M \models \varphi$ . We call  $\varphi$  satisfiable if there is  $M_s$  such that  $M_s \models \varphi$ .  $\dashv$

Observe that the quantification in the definition of semantics is restricted to the quantifier-free fragment. Moreover, given the eliminability of public announcements from that fragment [21], this amounts to quantifying over formulas of epistemic logic.

For clarity, we also give the semantics of the diamond versions of public and quantified announcements.

$$\begin{aligned}
M_s \models \langle \varphi \rangle \psi & \quad \text{iff } M_s \models \varphi \text{ and } M_s^\varphi \models \psi \\
M_s \models \Diamond \psi & \quad \text{iff there is } \varphi \in \mathcal{L}_{el} : M_s \models \langle \varphi \rangle \psi \\
M_s \models [G] \psi & \quad \text{iff there is } \varphi_G \in \mathcal{L}_{el}^G : M_s \models \langle \varphi_G \rangle \psi \\
M_s \models \langle \llbracket G \rrbracket \rangle \psi & \quad \text{iff there is } \varphi_G \in \mathcal{L}_{el}^G \text{ such that for all } \chi_{\overline{G}} \in \mathcal{L}_{el}^{\overline{G}} : M_s \models \varphi_G \wedge [\varphi_G \wedge \chi_{\overline{G}}] \psi
\end{aligned}$$

**Definition 6 (Modal Equivalence)** Given  $M_s$  and  $M'_s$ , if for all  $\varphi \in \mathcal{L}$ ,  $M_s \models \varphi$  iff  $M'_s \models \varphi$ , we write  $M_s \equiv M'_s$ . Similarly, if this holds for all  $\varphi$  with  $d(\varphi) \leq n$ , we write  $M_s \equiv^n M'_s$ , and if this holds for all  $\varphi$  with  $\text{var}(\varphi) \in Q \subseteq P$ , we write  $M_s \equiv_Q M'_s$ .  $\dashv$

It is a standard standard model-theoretic result for  $\mathcal{L}_{el}$  that bisimulation between models implies their modal equivalence [14]. We can extend the result to  $\mathcal{L}_{apal}$ ,  $\mathcal{L}_{gal}$ , and  $\mathcal{L}_{cal}$ .

**Theorem 7** Let  $M_s$  and  $M'_s$  be models. Then  $M_s \leftrightarrow M'_s$  implies  $M_s \equiv M'_s$ .  $\dashv$

**Proof** By an induction on the structure of a formula. The proof for the case of public announcements can be found, for example, in [7]. The cases of quantified announcements follow by the induction hypothesis in the presence of an appropriate ordering of subformulas such that the  $D$ -depth takes precedence over  $d$ -depth.  $\square$

Another well-known result is that for  $\mathcal{L}_{el}$  (and hence for  $\mathcal{L}_{pal}$  given the translation from  $\mathcal{L}_{pal}$  into  $\mathcal{L}_{el}$  [21]),  $M_s \leftrightarrow^n M'_s$  implies  $M_s \equiv^n M'_s$  [14]. Observe that this is not the case for any of the  $QPAL$ 's because the quantification is over formulas of arbitrary finite modal depth and thus may exceed the given  $n$ . Finally, due to the fact that the quantification in the presented languages is *implicit*, it is also not the case for  $QPAL$ 's that  $M_s \leftrightarrow_Q M'_s$  implies  $M_s \equiv_Q M'_s$ .

**Example 8** In order to highlight the differences between different types of quantification, let us consider models  $M_{s_0}$ ,  $M_{s_0}^\psi$ , and  $M_{s_0}^{\psi_a}$  in Figure 1. There is a formula  $\psi$  that can be announced in  $M_{s_0}$ , and such

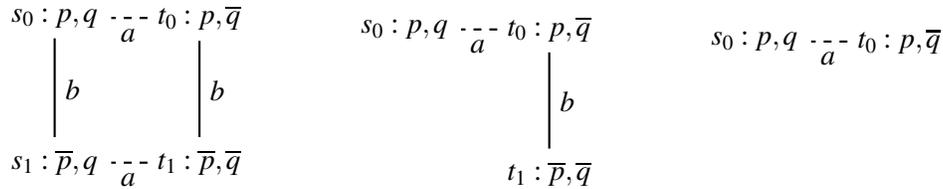


Figure 1: Models, from left to right,  $M_{s_0}$ ,  $M_{s_0}^\psi$ , and  $M_{s_0}^{\psi_a}$ . The names of the worlds indicate which atoms are true there.

that  $\varphi := \widehat{K}_a K_b p \wedge \widehat{K}_a \widehat{K}_b \neg p$  will hold after the announcement. Indeed, let  $\psi := \neg p \rightarrow \neg q$ . The result of updating  $M_{s_0}$  with  $\psi$  is presented in the figure and the reader can verify that  $M_{s_0}^\psi \models \varphi$ . This means that  $M_{s_0} \models \langle \psi \rangle \varphi$ , and hence  $M_{s_0} \models \diamond \varphi$ . Observe that  $M_{s_0} \not\models \langle \{a\} \rangle \varphi$ , since, according to the semantics, each announcement by agent  $a$  should be prefixed with  $K_a$ . This implies that in order to remove  $s_1$ , we also have to remove all  $a$ -reachable states, in particular  $t_1$ . On the other hand,  $M_{s_0} \models \langle \{a\} \rangle (K_b q \wedge \neg K_a q)$ . Indeed, let  $\psi_a := K_a p$ . Such an announcement results in model  $M_{s_0}^{\psi_a}$  in which  $K_b q \wedge \neg K_a q$  is true. Finally,  $M_{s_0} \not\models \langle \{a\} \rangle (K_b q \wedge \neg K_a q)$ , as any announcement by agent  $a$  that results in a model with worlds  $s_0$  and  $t_0$  can be countered by agent  $b$  with a simultaneous announcement  $K_b q$ . Such a joint announcement results in a singleton model with the only world  $s_0$ .  $\dashv$

### 3 APAL, GAL, and CAL do not have the finite model property

In this section, we show that none of the  $QPAL$ 's have the FMP. We do this by proving the result for  $APAL$  first, and then state the corresponding results for  $GAL$  and  $CAL$  as a corollary.

**Definition 9 (Finite Model Property)** A logic has the finite model property if every satisfiable formula is satisfied in a finite model.  $\dashv$

This idea of the proof is to show that there is a formula, **fmp**, that expresses the property that:

There is some subset of worlds in the model that can be partitioned into sets  $X$  and  $Y$ , such that for every element of  $x \in X$ , there is some announcement,  $\psi^x$ , that preserves  $x$  and no states from  $Y$ , but there is no announcement that preserves only the states from  $Y$ , and none of the states from  $X$ .

As the announcements in  $QPAL$ 's are closed under negations and conjunctions, if  $X$  were finite up to bisimulation, then the announcement  $\bigwedge_{x \in X} \neg \psi^x$ , would be adequate to preserve only the states of  $Y$  and none of the states from  $X$ . Therefore, if it can be shown that such a formula is satisfiable, then it would follow that  $QPAL$ 's do not have the finite model property. To show the satisfiability for such a formula, we need a means to identify  $X$  and  $Y$  states. An epistemic formula will not do, as any formula that characterises  $X$ , will have a negation that characterises  $Y$ , and announcing the negation would violate the property we need. However, rather than distinguishing between two partitions using a modal property, we can distinguish them using a second order property that the  $APAL$  quantifier does not range over. We know [3] that  $APAL$  can express such properties; particularly it can specify whether or not two states are  $n$ -bisimilar for all  $n$ . It is this property we use to define the necessary partition of worlds.

The proof will be via construction, where we will present a  $APAL$  formula, show that it has an infinite model, and then show that it is impossible that a finite model exists.

**Theorem 10** *APAL does not have the finite model property.* ←

**Proof** We will give the proof by construction. Consider the following formula **fmp**.

$$\begin{aligned}
 \mathbf{root} &= \Box(\widehat{K}_a(\neg x \wedge K_b \neg x) \rightarrow K_a(\neg x \rightarrow K_b \neg x)) \\
 \mathbf{stem} &= \Diamond(\widehat{K}_a(\neg x \wedge K_b \neg x) \wedge \widehat{K}_a(\neg x \wedge \widehat{K}_b x)) \\
 \mathbf{tier} &= K_b(x \wedge \widehat{K}_a \neg x \wedge K_a(\neg x \rightarrow \widehat{K}_b x)) \\
 \mathbf{fmp} &= \bigwedge \left( \begin{array}{l} \mathbf{tier} \wedge \widehat{K}_b \mathbf{root} \wedge \widehat{K}_b \mathbf{stem} \\ K_b(\mathbf{stem} \rightarrow \Diamond(\mathbf{tier} \wedge K_b \mathbf{stem})) \\ K_b(\mathbf{root} \rightarrow \Box(\mathbf{tier} \rightarrow \widehat{K}_b \mathbf{stem})) \end{array} \right)
 \end{aligned}$$

The formula **fmp** partitions a set of  $b$ -related worlds into two sets: *root* (the worlds where the formula **root** is true); and *stem* (the worlds where the formula **stem** is true). We note that **stem** is equivalent to  $\neg \mathbf{root}$  so this is a partition.

The formula **tier** sets a label  $x$ , where  $x$  is true at all the  $b$ -related worlds (*tier-0*), at every  $b$ -related world there is an  $a$ -related world where  $x$  is false (*tier-1*), and from each of those worlds there is a  $b$ -related world where  $x$  is true (*tier-2*). This creates a consistent labelling used to define **root** and **stem**.

The formula **root** is true at a world if there is only one  $a$ -reachable tier-1 world, up to finite bisimulation, while **stem** is true if there is more than one  $a$ -reachable tier-1 world.

The first line of **fmp** states that **tier** is true, and at least one of the  $b$ -related worlds satisfies **stem**, and at least one of the  $b$ -related worlds satisfies **root**. The second line of **fmp** states that there is some public announcement that removes all of the **root** worlds (and possibly some, but not all, of the **stem** worlds too), and the final line of **fmp** states that there is no public announcement that removes all of the **stem** worlds leaving a **root** world. This line needs a small caveat. Rather than just talking about removing **stem** and **root** worlds, there is the possibility that an announcement could change a **stem** world to a **root** world, or vice-versa. However each of these announcement quantifiers is guarded by the formula **tier**, and we have the property  $\mathbf{root} \rightarrow \Box \mathbf{root}$ . Therefore, we do not need to consider the cases where **root** worlds are transformed into **stem** worlds.

As an example consider a finite model in Figure 2. The model does not satisfy **fmp**, and in particular it does not satisfy the third conjunct. Indeed, let the current world be  $s_0$  that satisfies **root**. We show that  $M_{s_0} \not\models \Box(\mathbf{tier} \rightarrow \widehat{K}_b \mathbf{stem})$ , or, equivalently,  $M_{s_0} \models \Diamond(\mathbf{tier} \wedge K_b \neg \mathbf{stem})$ . By the semantics this amounts to the fact that there is  $\psi \in \mathcal{L}_{el}$  such that  $\mathbf{tier} \wedge K_b \neg \mathbf{stem}$  holds in the updated model. Let, for example,  $\psi := K_a(\neg x \rightarrow K_b(x \rightarrow K_a \neg p_4))$ . It is easy to verify that the resulting updated model, which only consists of states  $s_0, t_0$ , and  $u_0$ , satisfies  $\mathbf{tier} \wedge K_b \neg \mathbf{stem}$ .

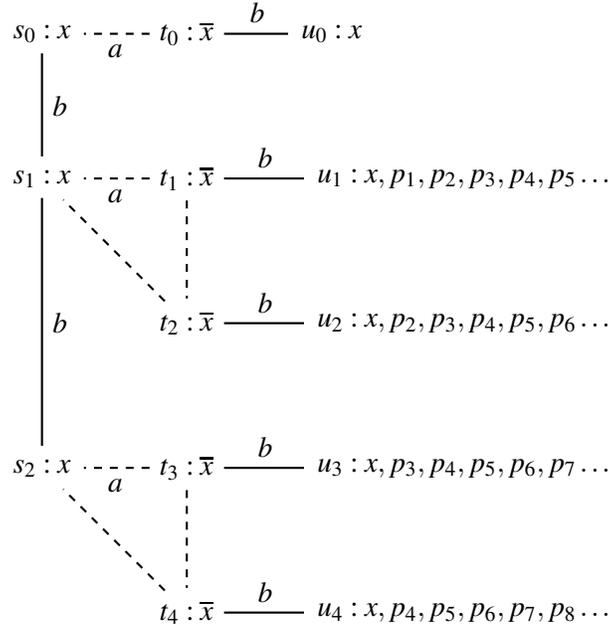


Figure 2: A model that does not satisfy formula **fmp**. The names of the worlds indicate which atoms are true there, e.g.  $x$  is true and  $\bar{x}$  means that  $x$  is false.

Now we show that **fmp** is a satisfiable formula. The model  $M = (S, \sim, V)$ , which satisfies **fmp**, is built as follows:

1.  $S = \{s_0, s_1, \dots\} \cup \{t_0, t_1, \dots\} \cup \{u_0, u_1, \dots\}$
2.  $\forall i \geq 0, s_i \sim_a s_i, s_i \sim_a t_{2i}, t_{2i} \sim_a s_i, t_i \sim_a t_i$  and  $u_i \sim_a u_i$ .
3.  $\forall i > 0, s_i \sim_a t_{2i-1}, t_{2i} \sim_a t_{2i-1}, t_{2i-1} \sim_a s_i$  and  $t_{2i-1} \sim_a t_{2i}$ .
4.  $\forall i, j \geq 0, s_i \sim_b s_j, t_i \sim_b u_i, u_i \sim_b t_i, t_i \sim_b t_i$  and  $u_i \sim_b u_i$ .
5.  $V(x) = \{s_0, s_1, \dots\} \cup \{u_0, u_1, \dots\}$
6.  $\forall i \geq 0, V(p_i) = \{u_k \mid 0 < k \leq i\}$ .

This model is represented in Figure 3.

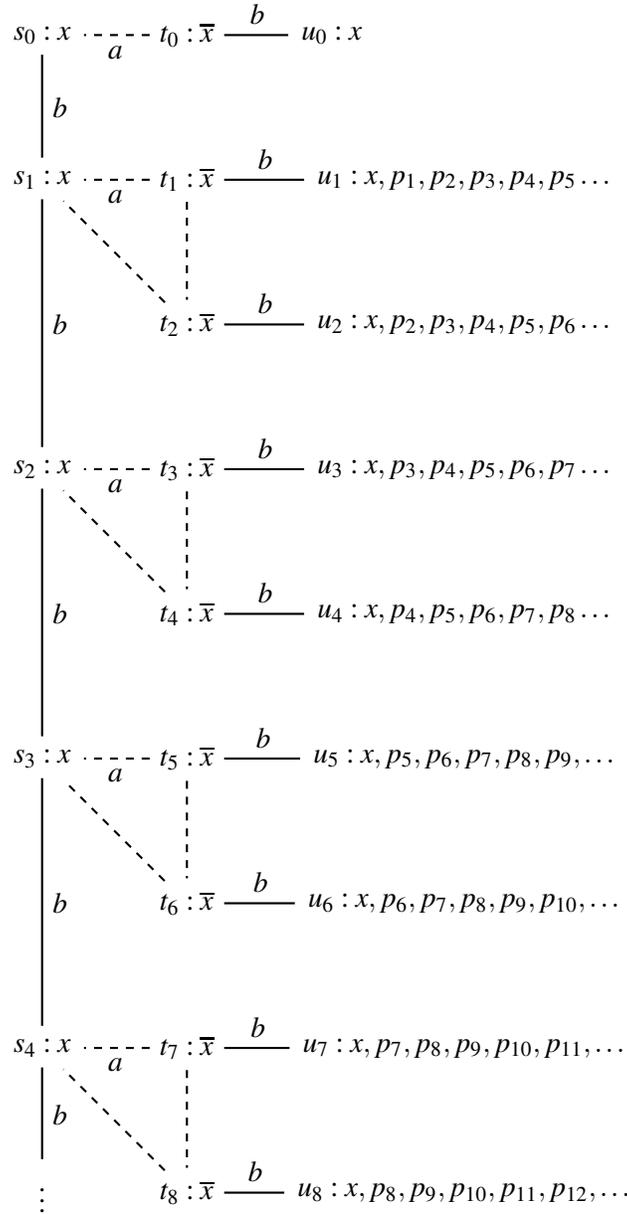


Figure 3: A model that does satisfy formula **fmp**. For each world  $s_i$ , where  $i > 0$ , there is an announcement that preserves  $s_i$ , but not  $s_0$ . However, there is no announcement that preserves only  $s_0$ .

The formula **root** is true only at the state  $s_0$ , and the formula **stem** is true at  $s_i$  for all  $i > 0$ . The states  $s_i$  are tier-0 states, the states  $t_i$  are tier-1 states, and the states  $u_i$  are tier-2 states. Therefore  $M_{s_0} \models \mathbf{tier} \wedge \widehat{K}_b \mathbf{root} \wedge \widehat{K}_b \mathbf{stem}$ . In any state  $s_i$  that satisfies **stem**, we can make an announcement  $\psi_i := \widehat{K}_a \widehat{K}_b p_{2i}$  that will preserve all the states  $s_j, t_j, u_j$  where  $0 < j \leq 2i$ . Therefore  $M_{s_i} \models \langle \psi_i \rangle (\mathbf{tier} \wedge K_b \mathbf{stem})$ , so  $M_{s_0} \models K_b (\mathbf{stem} \rightarrow \diamond (\mathbf{tier} \wedge K_b \mathbf{stem}))$ . Finally, consider any announcement,  $\psi$  that preserves the root state  $s_0$  and keeps **tier** true. Suppose that  $\text{var}(\psi) \subseteq P_n = \{x, p_0, \dots, p_n\}$ . The state  $s_0$  is  $P_n$ -bisimilar to all states  $s_i$  for  $i > n$ , so all these states will be preserved and continue to satisfy **stem**. Therefore  $M_{s_0} \models$

$K_b(\mathbf{root} \rightarrow \Box(\mathbf{tier} \rightarrow \widehat{K}_b\mathbf{stem}))$  as required. Since all three conjuncts are now satisfied,  $M_{s_0} \models \mathbf{fmp}$ .

Finally, we reason that **fmp** cannot have a finite model via a contradiction. Suppose that **fmp** did have some finite model  $M = (S, \sim, V)$ . We know **root** is equivalent to  $\neg\mathbf{stem}$ , and  $\mathbf{root} \rightarrow \Box(\mathbf{tier} \rightarrow \mathbf{root})$ . As  $M_s \models \mathbf{fmp}$  we have that for each  $s_1, s_2, \dots, s_k$ , where  $s \sim_b s_i$  and  $M_{s_i} \models \mathbf{stem}$ , there is some announcement  $\psi_i$  such that  $M_{s_i}^{\psi_i} \models \mathbf{tier} \wedge K_b\mathbf{stem}$ . Therefore, for some  $t \sim_b s$  where  $M_t \models \mathbf{root}$ , we have  $M_t \models \neg\psi_i$  for  $i = 1, \dots, k$ . Announcing  $\varphi := \bigwedge_{i=1}^k (\mathbf{tier} \rightarrow \neg\psi_i)$  at  $t$  therefore preserves  $t$ , and removes every  $s_i \sim s$  where  $M_{s_i} \not\models \mathbf{root}$ . So it follows that  $M_t \models \Diamond(\mathbf{tier} \wedge K_b\mathbf{root})$ , contradicting the fact that  $M_s \models \mathbf{fmp}$ .  $\square$

The proof of Theorem 10 can be extended to the cases of *GAL* and *CAL*. To see this, it is enough to notice that the formula **fmp** forces a model with a root-and-stem structure, and thus arbitrary announcements can be ‘modelled’ by announcements of the grand coalition. In particular, we need to substitute  $\Box$  and  $\Diamond$  in **fmp** with  $[A]$  and  $\langle A \rangle$  for *GAL*, and with  $\langle\langle A \rangle\rangle$  and  $\langle\langle A \rangle\rangle$  for *CAL*. Finally, the model in Figure 3 will also work, since the intersection of *a*- and *b*-relations is the identity, and the set of all possible *APAL* updates then coincides with those of *GAL* and *CAL*.

**Corollary 11** *GAL and CAL do not have the finite model property.*  $\dashv$

## 4 Conclusions and further research

It has been an open question for quite a while whether quantified public announcement logics have the finite model property, and we have answered the question for *APAL*, *GAL*, and *CAL* negatively. Not only this result is interesting in itself, it also clarifies some other properties of *QPAL*’s. In particular, from the expressivity perspective, we presented a formula that forces infinite models. Moreover, we have found the value of one of the unknowns in the expression

*Finitary axiomatisation and FMP imply decidability*

and thus only the problem of finding finitary axiomatisations of *QPAL*’s stands. Finally, the result trivially extends to the logics that are extensions of any of *QPAL*’s (e.g. [13]).

Interestingly, restricting the quantification in *APAL* to just announcements of Boolean formulas still results in a logic with no FMP. This was shown in [7], where the authors used a somewhat simpler partition of worlds distinguishable by a formula of epistemic logic. Since in *APAL* the quantifiers range over all epistemic formulas, we used a more complex second-order property of worlds being *n*-bisimilar for arbitrary *n*. It is unknown whether *APAL* with the quantification restricted only to positive (universal) fragment of epistemic logic [8] has the FMP, but given the aforementioned results, it seems very unlikely.

## Acknowledgments

We thank the anonymous reviewers of the paper. This work was carried out while Hans van Ditmarsch was affiliated to LORIA, CNRS, University of Lorraine, France.

## References

- [1] Thomas Ågotnes, Philippe Balbiani, Hans van Ditmarsch & Pablo Seban (2010): *Group announcement logic*. *Journal of Applied Logic* 8, pp. 62–81, doi:10.1016/j.jal.2008.12.002.
- [2] Thomas Ågotnes & Hans van Ditmarsch (2008): *Coalitions and announcements*. In Lin Padgham, David C. Parkes, Jörg P. Müller & Simon Parsons, editors: *Proceedings of the 7th AAMAS, IFAAMAS*, pp. 673–680.

- [3] Thomas Ågotnes, Hans van Ditmarsch & Tim French (2016): *The Undecidability of Quantified Announcements*. *Studia Logica* 104(4), pp. 597–640, doi:10.1007/s11225-016-9657-0.
- [4] Philippe Balbiani, Alexandru Baltag, Hans van Ditmarsch, Andreas Herzig, Tomohiro Hoshi & Tiago de Lima (2008): ‘Knowable’ as ‘known after an announcement’. *Review of Symbolic Logic* 1(3), pp. 305–334, doi:10.1017/S1755020308080210.
- [5] Thomas Bolander & Mikkel Birkegaard Andersen (2011): *Epistemic planning for single and multi-agent systems*. *Journal of Applied Non-Classical Logics* 21(1), pp. 9–34, doi:10.3166/janc1.21.9-34.
- [6] Hans van Ditmarsch (2020): *Quantifying Notes Revisited*. *CoRR* abs/2004.05802.
- [7] Hans van Ditmarsch & Tim French (2020): *Quantifying over Boolean announcements*. *CoRR* abs/1712.05310.
- [8] Hans van Ditmarsch, Tim French & James Hales (2020): *Positive Announcements*. *Studia Logica*, doi:10.1007/s11225-020-09922-1.
- [9] Dov M. Gabbay (1971): *On decidable, finitely axiomatizable, modal and tense logics without the finite model property. Part I*. *Israel Journal of Mathematics* 10, pp. 478–495, doi:10.1007/BF02771736.
- [10] Dov M. Gabbay (1971): *On decidable, finitely axiomatizable, modal and tense logics without the finite model property. Part II*. *Israel Journal of Mathematics* 10, pp. 496–503, doi:10.1007/BF02771737.
- [11] Dov M. Gabbay & Valentin B. Shehtman (1998): *Products of Modal Logics, Part 1*. *Logic Journal of the IGPL* 6(1), pp. 73–146, doi:10.1093/jigpal/6.1.73.
- [12] Rustam Galimullin (2019): *Coalition Announcements*. Ph.D. thesis, University of Nottingham.
- [13] Rustam Galimullin (2021): *Coalition and Relativised Group Announcement Logic*. *Journal of Logic, Language and Information*, doi:10.1007/s10849-020-09327-2.
- [14] Valentin Goranko & Martin Otto (2007): *Model Theory of Modal Logic*. In Patrick Blackburn, Johan van Benthem & Frank Wolter, editors: *Handbook of Modal Logic, Studies in Logic and Practical Reasoning 3*, Elsevier, pp. 249–329, doi:10.1016/S1570-2464(07)80008-5.
- [15] Robin Hirsch, Ian M. Hodkinson & Ágnes Kurucz (2002): *On Modal Logics Between  $K \times K \times K$  and  $S5 \times S5 \times S5$* . *Journal of Symbolic Logic* 67(1), pp. 221–234, doi:10.2178/jsl/1190150040.
- [16] Marcus Kracht (1991): *A solution to a problem of Urquhart*. *Journal of Philosophical Logic* 20(3), pp. 285–286, doi:10.1007/BF00250541.
- [17] Ágnes Kurucz (2000): *On Axiomatizing Products of Kripke Frames*. *Journal of Symbolic Logic* 65(2), pp. 923–945, doi:10.2307/2586578.
- [18] John-Jules Ch. Meyer & Wiebe van der Hoek (1995): *Epistemic Logic for AI and Computer Science*. Cambridge Tracts in Theoretical Computer Science, CUP, doi:10.1017/cbo9780511569852.
- [19] Zoran Ognjanović (2006): *Discrete Linear-time Probabilistic Logics: Completeness, Decidability and Complexity*. *Journal of Logic and Computation* 16(2), pp. 257–285, doi:10.1093/logcom/exi077.
- [20] Marc Pauly (2002): *A Modal Logic for Coalitional Power in Games*. *Journal of Logic and Computation* 12(1), pp. 149–166, doi:10.1093/logcom/12.1.149.
- [21] Jan Plaza (2007): *Logics of public communications*. *Synthese* 158(2), pp. 165–179, doi:10.1007/s11229-007-9168-7.
- [22] Alasdair Urquhart (1981): *Decidability and the finite model property*. *Journal of Philosophical Logic* 10(3), pp. 367–370, doi:10.1007/BF00293428.

# Fire!\*

Krisztina Fruzsa<sup>†</sup>

Roman Kuznets

Ulrich Schmid

TU Wien

{kfruzsa,rkuznets,s}@ecs.tuwien.ac.at

In this paper, we provide an epistemic analysis of a simple variant of the fundamental consistent broadcasting primitive for byzantine fault-tolerant asynchronous distributed systems. Our Firing Rebels with Relay (FRR) primitive enables agents with a local preference for acting/not acting to trigger an action (FIRE) at all correct agents, in an all-or-nothing fashion. By using the epistemic reasoning framework for byzantine multi-agent systems introduced in our TARK'19 paper, we develop the necessary and sufficient state of knowledge that needs to be acquired by the agents in order to FIRE. It involves eventual common hope (a modality related to belief), which we show to be attained already by achieving eventual mutual hope in the case of FRR. We also identify subtle variations of the necessary and sufficient state of knowledge for FRR for different assumptions on the local preferences.

## 1 Motivation and Background

In their PODC'18 paper “Silence” [13], Goren and Moses introduced and epistemically analyzed *silent choirs* as a fundamental primitive for message-optimal protocols in synchronous fault-tolerant distributed systems where computing nodes (agents<sup>1</sup>) can crash. In synchronous systems, where one can time-out messages, it is well-known [20] that an agent can convey information also by *not* sending some message. In a system where the sender may also crash, however, the receiver cannot infer this information from not receiving the message. Still, if only up to  $f$  of the  $n > f$  agents in a system may crash, a silent choir of  $f + 1$  agents that convey identical information suffices: at least one agent in the choir must be correct, so its silence can be relied on.

Whereas silent choirs also work in systems where the faulty agents may behave arbitrarily (byzantine [21]), the problem of not conveying information faithfully now also plagues messages that *are* sent, as they could originate from a faulty sender or forwarding agent. In this paper, we will introduce and epistemically analyze a fundamental primitive *Firing Rebels with Relay* (FRR), which nicely captures exactly these issues. It is a simplified version of the *consistent broadcasting* primitive introduced by Srikanth and Toueg in [27], which has been used as a pivotal building block in distributed algorithms for byzantine fault-tolerant clock synchronization [6, 11, 26, 27, 30] and synchronous consensus [28], for example.

Informally, FRR requires that *every* correct agent perform an action called FIRE, in an all-or-none fashion (though not necessarily simultaneously), and only if at least one correct agent locally observed a trigger event called START. Note that we have replaced the need to broadcast explicit information by just triggering an action, which makes FRR essentially a non-synchronous variant of the Firing Squad problem [4], hence its name. In crash-prone systems, FRR is trivial to solve, even for large  $f$ : Indeed, every agent who observes START or receives a notification message (for the first time) just invokes FIRE and sends a notification message to everyone. This guarantees that if a single correct agent observes START,

---

\*Funded by the Austrian Science Fund (FWF) project ByzDEL P33600.

<sup>†</sup>PhD student in the FWF doctoral program LogiCS (W1255).

<sup>1</sup>Since distributed systems are just one instance of multi-agent systems, we will use the term “agent” instead of “process.”

every correct agent will invoke FIRE (agents that crash during the run may or may not issue FIRE here). Observe that this solution involves a trivial silent choir, namely, when no agent observes START.

In the presence of byzantine agents, however, this solution does not work, as faulty agents may send a notification without having observed anything. A correct solution for FRR must, hence, prevent the faulty agents from triggering FIRE at any correct agent. In this paper, we will establish the necessary and sufficient state of knowledge for correctly solving FRR in our epistemic reasoning framework for byzantine multi-agent systems [17, 18, 19]. At least since the ground-breaking work by Halpern and Moses [14], the knowledge-based approach [8] has been known as a powerful tool for analyzing distributed systems. In a nutshell, it uses epistemic logic [16] to reason about knowledge and belief in distributed systems. As agents take actions (e.g., FIRE) based on the accumulated local knowledge, reasoning about the latter is useful both for protocol design and impossibility proofs.

In the *runs-and-systems* framework for reasoning about multi-agent systems [8, 14], the set of all possible runs  $r$  (executions) of a system  $I$  determines the Kripke model, formed by pairs  $(r, t)$  of a run  $r \in I$  and time  $t \in \mathbb{N}$  representing global states  $r(t)$ . Note that time is modeled as discrete for simplicity, without necessarily being available to the agents. Two pairs  $(r, t)$  and  $(r', t')$  are indistinguishable for agent  $i$  iff  $i$  has the same local state in both global states represented by those points, formally, if  $r_i(t) = r'_i(t')$ . A modal *knowledge operator*  $K_i$  is used to capture that agent  $i$  knows some fact  $\varphi$  in run  $r \in I$  at time  $t \in \mathbb{N}$ . Formally,  $(I, r, t) \models K_i \varphi$  iff for every  $r' \in I$  and for every  $t'$  with  $r_i(t) = r'_i(t')$  it holds that  $(I, r', t') \models \varphi$ . Note that  $\varphi$  can be a formula containing arbitrary atomic propositions like  $\overline{\text{occurred}}(e)$  (event  $e$  occurred) or  $\text{correct}_i$  ( $i$  did not fail yet), as well as other knowledge operators and temporal modalities like  $\diamond$  (eventually) and  $\square$  (always), combined by standard logical operators  $\neg, \wedge, \vee$ , and  $\rightarrow$ . For example,  $(I, r, t) \models \diamond K_i \overline{\text{occurred}}(e)$  states that there is some time  $t' \geq t$  when  $i$  knows that event  $e$  occurred. Important additional modalities for a group  $G$  of agents are *mutual knowledge*  $E_G \varphi := \bigwedge_{i \in G} K_i \varphi$  and *common knowledge*  $C_G \varphi$  that can be informally expressed as an infinite conjunction  $C_G \varphi \equiv E_G \varphi \wedge E_G(E_G \varphi) \wedge \dots$ ; in other words, this means that every agent in  $G$  knows  $\varphi$ , and every agent in  $G$  knows that every agent in  $G$  knows  $\varphi$ , and so on.

Actions performed by the agents when executing a protocol take place when they have accumulated some specific epistemic knowledge. According to the pivotal *Knowledge of Preconditions Principle* [22], it is universally true that if  $\varphi$  is a necessary condition for agent  $i$  to take a certain action then  $i$  may act only if  $K_i \varphi$  is true. For example, in order for agent  $i$  to decide on 0 in a binary consensus algorithm,  $i$  must know that some agent  $j$  has started with initial value  $x_j = 0$ , i.e.,  $K_i(\exists j : x_j = 0)$  must hold true. Showing that agents act without having attained the respective necessary knowledge is, hence, a very effective way for proving incorrectness of protocols. Conversely, optimal distributed algorithms can be designed by letting agents act as soon as all respective necessary knowledge has been established. Prominent examples are the protocols based on silent choirs analyzed in [13] and the *unbeatable* consensus protocols introduced in [5], which are not just worst-case optimal but also not strictly dominated w.r.t. termination time by any other protocol in *any* execution.

**Related work:** The knowledge-based approach has been used for studying several distributed computing problems in systems with uncertainty but no failures. In [3], Ben-Zvi and Moses considered the simple *ordered response* problem in distributed systems, where the agents had to respond to an external START event by executing a special one-shot action FIRE in a given order  $i_1, i_2, \dots$ . The authors showed that, in every correct solution, agent  $i_k$  has to establish nested knowledge  $K_{i_k} K_{i_{k-1}} \dots K_{i_1} \overline{\text{occurred}}(\text{START})$  before it can issue FIRE and that this nested knowledge is also sufficient. In the conference version [1] of [3], the authors also considered the *simultaneous response* problem where all agents had to issue FIRE at the same time. It requires the group  $G$  of firing agents to establish common knowl-

edge  $C_G \overline{\text{occurred}}(\text{START})$  [14]. This work was later extended to responses that are not simultaneous but tightly coordinated in time [2, 12].

The knowledge-based approach has also been successfully applied to fault-tolerant synchronous distributed systems. Agents suffering from crash or omission failures have been studied in [24, 25], primarily in the context of agreement problems [7, 15], which require some form of common knowledge. Important ingredients here are the indexical set of correct agents and a related belief operator  $B_i\varphi := K_i(\text{correct}_i \rightarrow \varphi)$  [23], which states that agent  $i$  knows  $\varphi$  to be true in all runs where  $i$  is correct. This notion of “defeasible knowledge” also underlies a variant of common knowledge that has been used successfully for characterizing simultaneous distributed agreement [24, 25]. Closer related to our FRR problem is eventual distributed agreement studied in [15], where the stronger notion of continual common knowledge proved its value. The latter needs to hold throughout a run, i.e., from the beginning, which makes sense here since it is only applied to conditions on the initial state. Continual common knowledge does not seem readily applicable to FRR, however, as START can occur at any time in a run. More recent results are the already mentioned unbeatable consensus algorithms in synchronous systems with crash failures [5] and the silent-choir based message-optimal protocols [13].

**Detailed contributions:** We rigorously define the FRR problem and its weaker variant FR, without the all-or-nothing requirement (agreement), in epistemic terms and identify the necessary and sufficient state of knowledge that must be established by a correct agent in order to issue FIRE in every correct solution for FRR. Since FRR involves distributed agreement, the required state of knowledge involves some form of (eventual) common knowledge of  $\overline{\text{occurred}}(\text{START})$ . Interestingly, it turned out that establishing the respective eventual mutual knowledge (namely, “eventual mutual hope” where the hope modality is defined as  $H_i\varphi := \text{correct}_i \rightarrow B_i\varphi$ ) already implies the required common knowledge (namely, “eventual common hope”). We also identify subtle variations of the necessary and sufficient state of knowledge for FRR for different assumptions on the occurrence of START.

Whereas identifying the necessary and sufficient state of knowledge for the agents to FIRE does not immediately lead to efficient practical protocols, it is an important first step towards this goal. Indeed, as for the ordered response problem in [3], for example, we expect this knowledge to lead to necessary and sufficient *communication structures*, which must be present in every run of any correct protocol solving FRR. Knowing the latter would not only enable us to decide right away whether the communication guarantees provided by some distributed system allow to solve FRR, but also facilitate the design of efficient protocols.

**Paper organization:** In Section 2, we introduce some minimal basic notation from our modeling framework [19]. In Section 3, we provide the detailed definition and an epistemic analysis of the FRR problem. Some conclusions and directions of future work in Section 4 complete our paper.

## 2 Preliminaries

In this section, we outline the basic concepts and facts that are employed in our epistemic analysis of Firing Rebels with Relay (FRR). Our analysis was actually performed within the rigorous framework (that is based on the standard runs-and-systems framework) first developed in [19], which incorporates agents’ ability to arbitrarily deviate from normative behavior. This framework enables one to formally express the epistemic limitations (of the agents) that the presence of possibly fully byzantine agents in the system imposes.

In particular, we proved in [18] that asynchronous agents in a system with fully byzantine agents can never *know* that a particular event took place or even that an agent performed a particular action (since

the local state of a malfunctioning agent may have been corrupted, the above statement applies even to the agent's knowledge of its own actions). This rather disappointing result stems from the inability of even correct agents to exclude the possibility of the so-called *brain-in-a-vat scenario*. In other words, an agent can never be sure that the events and actions recorded in its local history truly happened as recorded rather than being figments of its own malfunction. To make matters worse, agents can also never *know* that they are correct either. Thus, unable to rely on knowledge or their own correctness, our agents are forced to rely instead on *belief*  $B_i\varphi := K_i(\text{correct}_i \rightarrow \varphi)$  (cf. [23]). In this paper, we show that even belief is not always appropriate and needs to be replaced with a modality we called *hope* defined as  $H_i\varphi := \text{correct}_i \rightarrow B_i\varphi$  (for a detailed explanation see Remark 11).

Since the epistemic analysis presented in this paper is protocol-independent and does not rely on the artefacts of our modeling, we omit all details irrelevant to the task at hand and present our findings in an epistemic language with temporal modalities that is interpreted in Kripke models generated by runs in our framework. The purpose of this section is to provide all the necessary ingredients (referring the reader to [17, 18, 19] for full details of the said framework).

We fix a finite set  $\mathcal{A} = \{1, \dots, n\}$  of *asynchronous agents* with *perfect recall*. Each agent  $i \in \mathcal{A}$  can perform *actions* (according to its protocol), e.g., send *messages*. One of the actions any agent can perform is FIRE. Agents also witness *events* (triggered by the *environment*) such as message delivery. One of the events that can be observed by any agent is START. We use a discrete time model governed by a global clock with domain  $\mathbb{T} = \mathbb{N}$ . All events taking place after clock time  $t \in \mathbb{T}$  and no later than  $t + 1$  are grouped into a *round* denoted  $t + \frac{1}{2}$  and are treated as happening simultaneously. Apart from actions, everything in the system is governed by the *environment*. Unlike the environment, agents only have limited local information, in particular, being asynchronous, do not have access to the global clock. This is achieved by allowing them not to perform actions in some rounds and allowing them, in the absence of either actions or events, to stay in the same local state for several rounds in a row. The agents have perfect recall in the sense that, once recorded in their local history, actions and events are never forgotten.

No assumptions apart from liveness are made about the communication. Messages can be lost, arbitrarily delayed, and/or delivered in the wrong order. In addition, the environment may cause at most  $f$  agents to become *byzantine* faulty. A byzantine faulty agent can perform any action irrespective of its protocol and “observe” events that did not happen. It can also have false memories about actions it has performed. At the same time, like the global clock, such malfunctions are not directly visible to an agent.

Throughout the paper, horizontal bars signify phenomena that are correct. Note that the absence of this bar should not be equated to faultiness but rather means the absence of a claim of correctness.

Agent  $i$ 's local view of the system immediately after round  $t + \frac{1}{2}$ , referred to as (*process-time* or *agent-time*) *node*  $(i, t + 1)$ , is recorded in  $i$ 's *local state*  $r_i(t + 1)$ , also called  $i$ 's *local history*. A *run*  $r$  is a sequence of *global states*  $r(t) = (r_\varepsilon(t), r_1(t), \dots, r_n(t))$  of the whole system consisting of the *state*  $r_\varepsilon(t)$  of the *environment* and local states  $r_i(t)$  of every agent. Unlike local states, the global state of the system necessarily updates every round to include all actions and events that happened (even the empty set thereof is faithfully recorded and modifies the global state). The set of all global states is denoted  $\mathcal{G}$ .

What happens in each round is determined by nondeterministic protocols  $P_i$  of the agents, the non-deterministic protocol  $P_\varepsilon$  of the environment, and chance, the latter implemented as the *adversary* part of the environment (the exact technical details are not important for this paper).

In our epistemic analysis, we consider pairs  $(r, t)$  of a run  $r$  and time  $t$ . A *valuation function*  $\pi$  determines whether an atomic proposition from *Prop* is true in run  $r$  at time  $t$ . The determination is arbitrary except for a small set of *designated atomic propositions* whose truth value at  $(r, t)$  is fully determined by the state of the system. More specifically, for  $i \in \mathcal{A}$  and  $t \in \mathbb{T}$ ,

$\text{correct}_i$  is true at  $(r, t)$  iff no faulty event happened to  $i$  by time  $t$ ;

$\overline{\text{occurred}}_i(o)$  is true at  $(r, t)$  iff  $i$ 's local history  $r_i(t)$  contains an *accurate* record of action/event  $o$  occurring (for example, in this paper we use action  $o = \text{FIRE}$  and event  $o = \text{START}$ );

$$\overline{\text{occurred}}(o) := \bigvee_{i \in \mathcal{A}} \overline{\text{occurred}}_i(o).$$

An *interpreted system* is a pair  $I = (R, \pi)$  where  $R$  is the set of considered runs. The language is  $\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_i\varphi \mid \diamond\varphi \mid Y\varphi$  where  $p \in \text{Prop}$  and  $i \in \mathcal{A}$ ; derived Boolean connectives are defined in the usual way;  $\Box\varphi := \neg\diamond\neg\varphi$ . Truth for these *formulas* is defined in the standard way, in particular, for a run  $r \in R$ , time  $t \in \mathbb{T}$ , atomic proposition  $p \in \text{Prop}$ , agent  $i \in \mathcal{A}$ , and formula  $\varphi$ , we have  $(I, r, t) \models p$  iff  $(r, t) \in \pi(p)$ , and  $(I, r, t) \models K_i\varphi$  iff  $(I, r', t') \models \varphi$  for any  $r' \in R$  and  $t' \in \mathbb{T}$  such that  $r_i(t) = r'_i(t')$ , and  $(I, r, t) \models \diamond\varphi$  iff  $(I, r, t') \models \varphi$  for some  $t' \geq t$ , and  $(I, r, t) \models Y\varphi$  iff  $t > 0$  and  $(I, r, t-1) \models \varphi$ . A formula  $\varphi$  is valid in  $I$ , written  $I \models \varphi$ , iff  $(I, r, t) \models \varphi$  for all  $r \in R$  and  $t \in \mathbb{T}$ .

### 3 The Firing Rebels Problem

In [27], Srikanth and Toueg introduced the consistent broadcasting primitive, which proved its value in several different contexts, ranging from low-level byzantine fault-tolerant tick generation in various system models [6, 11, 26, 30] to classic clock synchronization [27] to Byzantine Agreement [28]. As argued already in Section 1, it gave rise to our *Firing Rebels with Relay* problem FRR [9], which can be seen as a natural generalization of a *silent choir* in byzantine fault-tolerant systems [13]. As an important building block for designing byzantine fault-tolerant systems, it is therefore a natural target for a detailed epistemic analysis in our framework [19].

Our Firing Rebels problems assume that every agent  $i \in \mathcal{A}$  may observe an event START and may generate an action FIRE according to the following specification:

**Definition 1** (Firing Rebels with and without Relay). *A system is consistent with Firing Rebels (FR) for  $f \geq 0$  when all runs satisfy:*

- (C) *Correctness: If at least  $2f + 1$  agents learn that START occurred at a correct agent, all correct agents perform FIRE eventually.*
- (U) *Unforgeability: If a correct agent performs FIRE, then START occurred at a correct agent.*

*Moreover, the system is consistent with Firing Rebels with Relay (FRR) if every run also satisfies:*

- (R) *Relay: If a correct agent performs FIRE, all correct agents perform FIRE eventually.*

**Remark 2** (Variants of Correctness). *A different specification for Correctness can sometimes be found in literature: “If at least  $f + 1$  reliable agents locally observed START, then some reliable agent fires eventually” (see, e.g., [4]). Here, a reliable agent is one that will always follow its protocol, which corresponds to a forever correct agent in our terminology. In the case of FRR, by invoking (R), this specification implies “If at least  $f + 1$  reliable agents locally observed START, then all reliable agents fire eventually.” Relying on such a specification in asynchronous settings is problematic, however, because reliability depends on the future behavior of the system. Even complete knowledge of the global state, at a given time in a run, does not allow to identify the reliable agents whose observations of START could be relied upon. Thus, we require  $2f + 1$  arbitrary (correct or faulty) agents instead. Of course, given the limit of  $f$  faulty agents per run, at least  $f + 1$  (not necessarily the same) of these agents will remain reliable in every run. Moreover, we relax the condition of the  $2f + 1$  agents locally observing START to each of them learning that START happened to some correct agent. This is preferable, because direct observation is only one possible way of ascertaining that START occurred. For instance, if an agent has*

already determined<sup>2</sup> who the  $f$  faulty agents are, e.g., due to their erratic behavior in the past, then a confirmation of START from just one other agent would be sufficient.

We use the following abbreviations:

$$\begin{aligned} B_i\varphi &:= K_i(\text{correct}_i \rightarrow \varphi) & H_i\varphi &:= \text{correct}_i \rightarrow B_i\varphi = \text{correct}_i \rightarrow K_i(\text{correct}_i \rightarrow \varphi) \\ E^B\varphi &:= \bigwedge_{j \in \mathcal{A}} B_j\varphi & E^H\varphi &:= \bigwedge_{j \in \mathcal{A}} H_j\varphi \\ E^{\diamond B}\varphi &:= \bigwedge_{j \in \mathcal{A}} \diamond B_j\varphi & E^{\diamond H}\varphi &:= \bigwedge_{j \in \mathcal{A}} \diamond H_j\varphi \end{aligned}$$

It has been shown in [10] that hope is a normal modality, in particular,  $\models H_i(\varphi \wedge \psi) \rightarrow H_i\varphi \wedge H_i\psi$ . We define eventual common hope  $C^{\diamond H}\varphi$  as the greatest fixed point of the equation  $\chi \leftrightarrow E^{\diamond H}(\varphi \wedge \chi)$  in the standard way (using the Knaster–Tarski theorem [29]) and use the following properties (the general versions of which can be found in Lemma 11.5.7 in [8]): for any interpreted system  $I$ ,

$$I \models C^{\diamond H}\varphi \leftrightarrow E^{\diamond H}(\varphi \wedge C^{\diamond H}\varphi); \quad (1)$$

$$\text{if } I \models \psi \rightarrow E^{\diamond H}(\varphi \wedge \psi), \quad \text{then } I \models \psi \rightarrow C^{\diamond H}\varphi. \quad (2)$$

### 3.1 Modeling

**Definition 3.** For an agent  $i \in \mathcal{A}$ , we define:

$$\begin{aligned} \overline{\text{start}}_i &:= Y\overline{\text{occurred}}_i(\text{START}) \wedge \text{correct}_i & \overline{\text{fire}}_i &:= \overline{\text{occurred}}_i(\text{FIRE}) \wedge \text{correct}_i \\ \overline{\text{start}} &:= \bigvee_{j \in \mathcal{A}} \overline{\text{start}}_j & \overline{\text{fire}} &:= \bigvee_{j \in \mathcal{A}} \overline{\text{fire}}_j \end{aligned}$$

Note that for one of these formulas to be true, it is necessary for (one of) the involved agent(s) to be correct not only at the time the event/action in question occurred but also at the time of the evaluation. Using the yesterday modality  $Y$  in  $\overline{\text{start}}_i$  accounts for the fact that agents cannot act on a precondition in the same round it is established.

Using Def. 3, we can translate the specification of FRR (stated in Def. 1) as follows:

**Definition 4** (Modeling Firing Rebels). *An interpreted system  $I$  is consistent with Firing Rebels with Relay for  $f \geq 0$  if the following conditions Correctness (C), Unforgeability (U), and Relay (R) hold:*

$$\begin{aligned} \text{(C)} \quad I &\models \bigvee_{\substack{G \subseteq \mathcal{A} \\ |G|=2f+1}} \bigwedge_{j \in G} K_j(\text{correct}_j \rightarrow \overline{\text{start}}) \rightarrow \bigwedge_{i \in \mathcal{A}} \diamond(\text{correct}_i \rightarrow \overline{\text{fire}}_i) \\ \text{(U)} \quad I &\models \overline{\text{fire}} \rightarrow \overline{\text{start}} \\ \text{(R)} \quad I &\models \overline{\text{fire}} \rightarrow \bigwedge_{i \in \mathcal{A}} \diamond(\text{correct}_i \rightarrow \overline{\text{fire}}_i) \end{aligned}$$

**Remark 5** (Variants of eventuality). *The phrase all correct agents fulfill  $\varphi_i$  eventually in Def. 1 can be formalized in two different ways:*

- $\bigwedge_{i \in \mathcal{A}} \diamond(\text{correct}_i \rightarrow \varphi_i)$  states that each agent will either become faulty at some point in the future or will fulfill its respective  $\varphi_i$  at some point in the future.

<sup>2</sup>Strictly speaking, the agent in this situation *does not know* that the  $f$  agents are faulty, but rather that they are faulty if it itself is not. By the same token, whenever we say “learned,” “determined,” or “ascertained” above, what we mean is reasoning under the assumption of its own correctness, i.e., the belief modality  $B_i$  rather than the knowledge modality  $K_i$ .

- $\diamond \bigwedge_{i \in \mathcal{A}} (\text{correct}_i \rightarrow \varphi_i)$  states that there is one moment in the future by which every agent still correct fulfills its respective  $\varphi_i$ .

The second statement is a strengthening of the first by demanding agents to have one common moment by which all correct agents fulfill their respective  $\varphi_i$ 's. On the other hand, the first variant is a *more intuitive* reading that is more widely applicable. Fortunately, for our model of FRR with  $\varphi_i = \overline{\text{fire}}_i$ , the two formulations are equivalent because, due to agents having perfect recall,  $\text{correct}_i \rightarrow \overline{\text{fire}}_i$  is a stable fact.

### 3.2 Necessary and Sufficient Level of Knowledge

The goal of this subsection is to

- strengthen the given necessary conditions on a single agent's firing — namely, that  $\overline{\text{start}}$  must hold by Unforgeability (U) and  $\bigwedge_{i \in \mathcal{A}} \diamond (\text{correct}_i \rightarrow \overline{\text{fire}}_i)$  must hold by Relay (R) — to statements that describe the state of knowledge necessary for the agent to achieve before firing;
- show that firing upon reaching this state of knowledge is sufficient for satisfying the conditions Unforgeability (U) and Relay (R) on all correct agents eventually firing;
- show how Correctness (C) helps simplify these necessary and sufficient conditions in the presence of sufficiently many agents.

Thus, protocols prescribing an agent to fire as soon as this state of knowledge is achieved are correct and optimal in the sense that firing earlier would violate the necessary conditions whereas firing prescribed by this state of knowledge is guaranteed to fulfill all requirements of FRR.

Note that the case when insufficiently many agents learn that START occurred at a correct agent trivially satisfies condition (C). In this case, FRR reduces to (U)+(R), a problem with a trivial solution of all correct agents not firing. It is the combination of all all three conditions that makes FRR a problem worth the analysis.

The first lemma formalizes the fact that, since agents have perfect recall of their past perceptions, reasoning under the assumption of their own correctness leads them to *believe* that these perceptions were accurate. For instance, an agent who recalls observing START believes that, unless it is faulty, a correct agent (namely, itself) observed START. (A formal proof can be found in the Appendix on p. 151.)

**Lemma 6.** *For any interpreted system  $I$  and any agent  $i \in \mathcal{A}$ :*

$$I \models \overline{\text{fire}}_i \rightarrow B_i \overline{\text{fire}}_i \quad (3)$$

$$I \models \overline{\text{fire}}_i \rightarrow B_i \overline{\text{fire}} \quad (4)$$

$$I \models \overline{\text{start}}_i \rightarrow B_i \overline{\text{start}}_i \quad (5)$$

$$I \models \overline{\text{start}}_i \rightarrow B_i \overline{\text{start}} \quad (6)$$

Unforgeability (U) states that  $\overline{\text{start}}$  is a necessary condition for a correct agent firing. It follows from the Knowledge of Preconditions Principle that any correct agent must ascertain  $\overline{\text{start}}$  (modulo its own correctness) before firing. We formalize this argument and provide an independent proof:

**Lemma 7** (State of knowledge necessary for firing in presence of Unforgeability (U)). *Let  $I$  be an interpreted system consistent with Unforgeability (U). For any agent  $i \in \mathcal{A}$ ,*

$$I \models \overline{\text{fire}}_i \rightarrow B_i \overline{\text{start}}. \quad (7)$$

*Proof.* Immediately follows from (4), (U), and monotonicity/normality of  $B_i$ . □

**Corollary 8.** *For any interpreted system consistent with FR, (7) is satisfied for all agents.*

Similarly, lifting the Relay condition (R) to the level of agent's knowledge yields the requirement that, in order to fire, an agent must believe that all correct agents eventually will have fired.

**Lemma 9** (State of knowledge necessary for firing in presence of Relay (R)). *Let  $I$  be an interpreted system consistent with Relay (R). For any agent  $i \in \mathcal{A}$ ,*

$$I \models \overline{\text{fire}}_i \rightarrow B_i \bigwedge_{j \in \mathcal{A}} \diamond(\text{correct}_j \rightarrow \overline{\text{fire}}_j). \quad (8)$$

*Proof.* Immediately follows from (4), (R), and monotonicity of  $B_i$ .  $\square$

Combining the conditions necessary for (U) and (R), we establish the following level of knowledge necessary for firing in FRR (a proof can be found in the Appendix on p. 151):

**Theorem 10** (State of knowledge necessary for firing in presence of both (U) and (R)). *Let  $I$  be an interpreted system consistent with (U) and (R). For any agent  $i \in \mathcal{A}$ ,*

$$I \models \overline{\text{fire}}_i \rightarrow B_i \left( \overline{\text{start}} \wedge E^{\diamond H} \overline{\text{start}} \right).$$

**Remark 11** (Emergence of hope). *Note that  $I \models \overline{\text{fire}}_i \rightarrow B_i E^{\diamond B} \overline{\text{start}}$  does not generally hold. We cannot strengthen the necessary condition by replacing eventual mutual hope with eventual mutual belief, i.e., by omitting  $\text{correct}_j$  therein. In other words, the use of hope for deeper iterations of knowledge modalities is crucial for the correct formulation. Indeed, in the case of our notion of belief, agent  $i$  can rarely have unconditional beliefs about another agent  $j$ 's beliefs. The problematic situation is when agent  $j$ 's perception is compromised. In that case, agent  $i$  has no way of ascertaining what  $j$ 's erroneous input data might be and, hence, cannot determine what a correct agent would have inferred from these incorrect inputs. According to our notion of belief, whether agent  $i$  itself is correct or not, it reasons assuming that its own perceptions are the objective reality. The  $\text{correct}_j$  assumption is, therefore, necessary to anchor  $j$  to the same (allegedly) objective reality contemplated by  $i$ , even though  $j$ 's access to the facts of this objective reality is generally different from  $i$ 's. Note also that  $j$ 's reasoning is generally happening in the future relative to  $i$ 's current reasoning, meaning that we also implicitly assume reality to be stable.*

**Remark 12** (Relation to indexical sets). *Another approach to describing beliefs of fault-prone agents is via so-called indexical sets [8, 25], which are variable (non-rigid) sets that can be used to represent the set of all correct agents at every point in the system. While our results could be reformulated in terms of indexical sets, there were several reasons for us to choose another language. Besides the ability to reason about all agents, whether correct or faulty, in a uniform way, we tried to stay as close as possible to the standard language of epistemic modal logic. Perhaps more importantly, however, was the moral lesson of the already mentioned Knowledge of Preconditions Principle [22], which reveals how important it is for an agent to know all ingredients affecting its behavior, correctness of itself and other agents being one of them. Thus, we believe that the transparent and explicit use of correctness in our language is advantageous. An immediate example is the distinction between belief and hope discussed in Remark 11, which would have remained somewhat obscured in the indexical set notation.*

**Remark 13** (Eventual mutual hope is not sufficient). *While using  $B_i (\overline{\text{start}} \wedge E^{\diamond H} \overline{\text{start}})$  as a trigger for agent  $i$  firing will ensure Unforgeability (U), it is too weak to guarantee Relay (R). Indeed, consider a system with 3 agents ( $n = 3$ ), at most one of which can become faulty ( $f = 1$ ). In such a system, receiving the same information from two independent sources is sufficient to believe in its validity, while information from only one source without observing it first hand is not. Suppose that the protocol forces*

a correct agent to notify all other agents whenever it observed START. Consider a run where agent  $b$  is byzantine from the beginning, whereas agents  $c_1$  and  $c_2$  remain correct. Let  $c_1$  and  $c_2$  each observe START and, hence, notify all agents about it. Meanwhile  $b$  falsely notifies  $c_2$  that it too observed START but will never duplicate this message to  $c_1$ . Thus,

- correct  $c_2$  observed START and eventually received 2 confirmations of START from  $c_1$  and  $b$ ;
- correct  $c_1$  observed START and eventually received 1 confirmation of START from  $c_2$ ;
- faulty  $b$  did not observe START but was eventually notified of START by both  $c_1$  and  $c_2$ .

In this situation, all agents eventually believe that START was correctly observed ( $c_1$  and  $c_2$  saw it themselves, whereas  $b$  has 2 independent confirmations). Moreover,  $c_2$  has a reason to believe in the eventual mutual hope of START. Indeed, hope would be trivially satisfied for a faulty agent, whereas any correct agent would eventually receive at least 2 confirmations out of 3 that  $c_2$  itself possesses. Thus, according to the proposed knowledge threshold,  $c_2$  should fire. On the other hand,  $c_1$  will never fire because it cannot be sure that  $b$  will eventually hope that START occurred. In  $c_1$ 's mind, if  $b$  were correct and  $c_2$  were faulty and did not send a confirmation to  $b$ , then  $b$  would only ever receive 1 confirmation, which is not sufficient to make it trust START truly occurred. Hence,  $c_1$  would never fire, and Relay (R) would be violated.

The issue here is that  $B_i E^{\diamond H} \overline{\text{start}}$  for one correct agent  $i$  does not generally imply that eventually  $B_j E^{\diamond H} \overline{\text{start}}$  for all other correct agents  $j$ .

Thus, although  $B_i E^{\diamond H} \overline{\text{start}}$  is necessary before  $i$  can fire and is in principle actionable, acting on it may be premature. The necessary state of knowledge must be further strengthened. Since FRR involves an agreement property (one correct agent fires only if all other correct agents also fire eventually), it is not very surprising that, in fact, some form of common knowledge, specifically *eventual common hope*, plays a role. We have shown (see a proof in the Appendix on p. 152) that Unforgeability and Relay together imply that, in order to fire an agent must ascertain (modulo its own correctness) both that START was observed by some correct agent and the eventual common hope of the same fact:

**Theorem 14** (State of knowledge necessary for firing in presence of both (U) and (R)). *Let  $I$  be an interpreted system consistent with (U) and (R). For any agent  $i \in \mathcal{A}$ ,*

$$I \models \overline{\text{fire}}_i \rightarrow B_i \left( \overline{\text{start}} \wedge C^{\diamond H} \overline{\text{start}} \right). \quad (9)$$

**Corollary 15.** *For any interpreted system consistent with FRR, (9) is satisfied for all agents.*

We now show (see a proof in the Appendix on p. 152) that, unlike belief in eventual mutual hope (see Remark 13), belief in eventual *common hope* is sufficient to fulfill Unforgeability and Relay, i.e., that firing as soon as the necessary state of knowledge from Theorem 14 is achieved does guarantee that both (U) and (R) are fulfilled:

**Theorem 16** (Sufficient conditions for (U) and (R)). *For any interpreted system  $I$ :*

1. (U) is fulfilled if  $I \models \bigwedge_{i \in \mathcal{A}} (\neg B_i \overline{\text{start}} \rightarrow \neg \overline{\text{fire}}_i)$ .
2. Both (U) and (R) are fulfilled if

$$I \models \bigwedge_{i \in \mathcal{A}} \left( \left( \neg B_i \left( \overline{\text{start}} \wedge C^{\diamond H} \overline{\text{start}} \right) \rightarrow \neg \overline{\text{fire}}_i \right) \wedge \left( B_i \left( \overline{\text{start}} \wedge C^{\diamond H} \overline{\text{start}} \right) \rightarrow \diamond (\text{correct}_i \rightarrow \overline{\text{fire}}_i) \right) \right). \quad (10)$$

**Remark 17** (Belief in  $\overline{\text{start}}$  is not redundant). *Since common knowledge is the strongest type of knowledge and knowledge is supposed to be factive, it might be tempting to think that the conjunct  $\overline{\text{start}}$  is redundant in the formulations of Theorems 10, 14, and 16.2. The difference in our setting is that the relevant epistemic state is eventual, meaning that it need not be factual at present. Still, one might question how it could be possible to achieve even an eventual knowledge/belief/hope without the event actually happening. Indeed, if there is no reason for agents to expect START to necessarily occur, their predictions about START occurring can only rely on it already having occurred. This observation is formalized in Lemma 23 and the immediately following corollary.*

**Definition 18** (Potentially persistent formulas). *A formula  $\varphi$  is called potentially persistent in an interpreted system  $I = (R, \pi)$  if, for any run  $r \in R$  and any time  $t \in \mathbb{T}$  such that  $(I, r, t) \models \varphi$ , there exists a run  $r' \in R$  such that  $r'(t) = r(t)$  — i.e.,  $r'$  is an alternative continuation of the global state  $r(t)$  — and such that  $(I, r', t) \models \Box\varphi$ . In other words, a true potentially persistent formula can stay true forever.*

The following 3 lemmas follow from definitions (proofs of the last two are in the Appendix on p. 152).

**Lemma 19.**  $I \models \neg \text{correct}_i \rightarrow \Box \neg \text{correct}_i$  for any  $i \in \mathcal{A}$  and interpreted system  $I$ .

**Lemma 20.**  $I \models K_i \Diamond \neg \varphi \rightarrow K_i \neg \varphi$  for any  $i \in \mathcal{A}$  and  $\varphi$  potentially persistent in an interpreted system  $I$ .

**Lemma 21.**  $I \models B_i \Diamond (\text{correct}_i \rightarrow \varphi) \leftrightarrow K_i \Diamond (\text{correct}_i \rightarrow \varphi)$  for any  $i \in \mathcal{A}$ , formula  $\varphi$ , and interpreted system  $I$ , i.e., believing something eventually happens modulo one's own correctness is as strong as knowing it eventually happens modulo one's own correctness.

**Corollary 22.**  $I \models B_i \Diamond H_i \varphi \leftrightarrow K_i \Diamond H_i \varphi$  for any  $i \in \mathcal{A}$ , formula  $\varphi$ , and interpreted system  $I$ .

**Lemma 23** (Early local belief). *If  $\text{correct}_i \wedge \overline{\text{start}}$  is potentially persistent in an interpreted system  $I$ ,*

$$I \models B_i \Diamond H_i \overline{\text{start}} \rightarrow B_i \overline{\text{start}}.$$

*Proof.* A proof can be found in the Appendix on p. 153. □

Noting that  $I \models C^{\Diamond H} \overline{\text{start}} \rightarrow E^{\Diamond H} \overline{\text{start}}$  because of the normality of the hope modality [10], we can derive:

**Corollary 24.** *If  $\text{correct}_i \wedge \overline{\text{start}}$  is potentially persistent in an interpreted system  $I$ , then*

$$I \models B_i E^{\Diamond H} \overline{\text{start}} \rightarrow B_i \overline{\text{start}}, \quad (11)$$

$$I \models B_i C^{\Diamond H} \overline{\text{start}} \rightarrow B_i \overline{\text{start}}. \quad (12)$$

**Remark 25** (Conditions on dropping  $B_i \overline{\text{start}}$ ). *If, contrary to the conditions of the early local belief lemma,  $\overline{\text{start}}$  is inevitable, agents may be able to predict the eventual arrival of START before the fact. For instance, if START eventually happens to every agent, i.e., if  $I \models \Diamond \text{occurred}_i(\text{START})$ , then  $I \models \Diamond B_i \overline{\text{start}}$  for every agent  $i \in \mathcal{A}$ . It follows that  $I \models CC^{\Diamond H} \overline{\text{start}}$ , i.e., it is common knowledge, from the very beginning, that there is eventual common hope of  $\overline{\text{start}}$ . Thus, this state of knowledge is achieved independently of START happening, and triggering FIRE risks violating Unforgeability (U).*

*On the other hand, even if START is assured, it may not always be possible to predict it in advance. While sufficient for dropping the conjunct  $\overline{\text{start}}$  from the conditions triggering FIRE in Theorem 16.2, the potential persistency of  $\text{correct}_i \wedge \overline{\text{start}}$  is not necessary. Indeed, (12) can hold even when START is always guaranteed to happen. For instance, in an interpreted system where START happens exactly once per run, no agent ever becomes faulty, and, in addition, agents never communicate,  $I \models \neg B_i E^{\Diamond H} \overline{\text{start}}$*

because only the agent who observed START can learn that it already occurred. All the others can only be sure that START will occur eventually. By (1), also  $I \models \neg B_i C^{\diamond H} \overline{start}$ . Thus, both implications (11) and (12) are vacuously true, allowing to drop  $\overline{start}$  without affecting the behavior of agents, though admittedly in such interpreted systems agents should never fire anyways.

The following ‘‘Lifting Lemma’’ shows that Correctness (C) lifts eventual mutual hope to eventual common hope. This way, the arbitrarily deep nested hope implied by the latter effectively collapses, a phenomenon that has also been reported for other problems [1]. A proof of the lemma can be found in the Appendix on p. 153.

**Lemma 26** (Lifting Lemma). *Let  $I$  be an interpreted system consistent with (C) and let  $|\mathcal{A}| \geq 3f + 1$ , where  $f \geq 0$  is the maximum number of byzantine faulty agents in a run. Furthermore, assume that*

$$I \models \overline{fire}_i \rightarrow B_i \left( \overline{start} \wedge E^{\diamond H} \overline{start} \right) \quad (13)$$

holds. Then,

$$I \models E^{\diamond H} \overline{start} \rightarrow C^{\diamond H} \overline{start}. \quad (14)$$

**Corollary 27.** *Let  $I$  be an interpreted system with at least  $3f + 1$  agents. If  $I$  is consistent with FRR, then (14) holds.*

*Proof.* In interpreted systems consistent with (U) and (R), property (13) follows from Theorem 10.  $\square$

## 4 Conclusions and Future Work

We introduced a problem called Firing Rebels with Relay (FRR) and its weaker variant called Firing Rebels (FR), which capture the essentials of a well-known building block for byzantine fault-tolerant distributed algorithms. The main purpose of our paper was to determine the state of knowledge correct agents must achieve in order to act (FIRE) according to the specification of the problem at hand. Through a detailed epistemic analysis, we established that the necessary and sufficient levels of knowledge required for acting rely on the novel notion of eventual common hope. We also found the conditions under which a single level of eventual mutual hope can guarantee infinitely many levels of eventual common hope and explored the surprisingly non-trivial relationship of the eventual common hope of START with the actual appearance of START.

Regarding future work, our next step is to complete the characterization of (eventual) common hope. More precisely, what remains to be done is developing an independent axiomatization of the (eventual) common hope modality based on our existing axiomatization of the hope modality (which does not depend on the knowledge modality). In addition, we are working on identifying necessary and sufficient communication structures and optimal protocols for FRR.

## References

- [1] Ido Ben-Zvi & Yoram Moses (2010): *Beyond Lamport’s Happened-Before: On the Role of Time Bounds in Synchronous Systems*. In Nancy A. Lynch & Alexander A. Shvartsman, editors: *DISC 2010: Distributed Computing*, LNCS 6343, Springer, pp. 421–436, doi:10.1007/978-3-642-15763-9\_42.
- [2] Ido Ben-Zvi & Yoram Moses (2013): *Agent-Time Epistemics and Coordination*. In Kamal Lodaya, editor: *ICLA 2013: Logic and Its Applications*, LNCS 7750, Springer, pp. 97–108, doi:10.1007/978-3-642-36039-8\_9.

- [3] Ido Ben-Zvi & Yoram Moses (2014): *Beyond Lamport's Happened-before: On Time Bounds and the Ordering of Events in Distributed Systems*. *Journal of the ACM* 61(2:13), doi:10.1145/2542181.
- [4] James E. Burns & Nancy A. Lynch (1987): *The Byzantine Firing Squad Problem*. In Franco P. Preparata, editor: *Parallel and Distributed Computing, Advances in Computing Research: A research annual* 4, JAI Press, pp. 147–161. Available at <https://apps.dtic.mil/docs/citations/ADA154770>.
- [5] Armando Castañeda, Yannai A. Gonczarowski & Yoram Moses (2014): *Unbeatable Consensus*. In Fabian Kuhn, editor: *DISC 2014: Distributed Computing, LNCS 8784*, Springer, pp. 91–106, doi:10.1007/978-3-662-45174-8\_7.
- [6] Danny Dolev, Matthias Függer, Markus Posch, Ulrich Schmid, Andreas Steininger & Christoph Lenzen (2014): *Rigorously modeling self-stabilizing fault-tolerant circuits: An ultra-robust clocking scheme for systems-on-chip*. *Journal of Computer and System Sciences* 80(2), pp. 860–900, doi:10.1016/j.jcss.2014.01.001.
- [7] Cynthia Dwork & Yoram Moses (1990): *Knowledge and Common Knowledge in a Byzantine Environment: Crash Failures*. *Information and Computation* 88(2), pp. 156–186, doi:10.1016/0890-5401(90)90014-9.
- [8] Ronald Fagin, Joseph Y. Halpern, Yoram Moses & Moshe Y. Vardi (1995): *Reasoning About Knowledge*. MIT Press.
- [9] Patrik Fimml (2018): *Temporal-Epistemic Logic in Byzantine Message-Passing Contexts*. Master's thesis, Technische Universität Wien, Institut für Computer Engineering. Available at [http://publik.tuwien.ac.at/files/publik\\_273448.pdf](http://publik.tuwien.ac.at/files/publik_273448.pdf).
- [10] Krisztina Fruzsa (2019): *Hope for Epistemic Reasoning with Faulty Agents!* In: *ESSLLI 2019 Student Session, FOLLI*. Available at [http://esslli2019.folli.info/wp-content/uploads/2019/08/tentative\\_proceedings.pdf](http://esslli2019.folli.info/wp-content/uploads/2019/08/tentative_proceedings.pdf).
- [11] Matthias Függer & Ulrich Schmid (2012): *Reconciling fault-tolerant distributed computing and systems-on-chip*. *Distributed Computing* 24(6), pp. 323–355, doi:10.1007/s00446-011-0151-7.
- [12] Yannai A. Gonczarowski & Yoram Moses (2013): *Timely Common Knowledge: Characterising Asymmetric Distributed Coordination via Vectorial Fixed Points*. In Burkhard C. Schipper, editor: *TARK 2013: Theoretical Aspects of Rationality and Knowledge*, pp. 79–93. Available at [http://www.tark.org/proceedings/tark\\_jan7\\_13/p79-gonczarowski.pdf](http://www.tark.org/proceedings/tark_jan7_13/p79-gonczarowski.pdf).
- [13] Guy Goren & Yoram Moses (2018): *Silence*. In: *PODC 2018: Principles of Distributed Computing, ACM*, pp. 285–294, doi:10.1145/3212734.3212768.
- [14] Joseph Y. Halpern & Yoram Moses (1990): *Knowledge and Common Knowledge in a Distributed Environment*. *Journal of the ACM* 37(3), pp. 549–587, doi:10.1145/79147.79161.
- [15] Joseph Y. Halpern, Yoram Moses & Orli Waarts (2001): *A characterization of eventual Byzantine agreement*. *SIAM Journal on Computing* 31(3), pp. 838–865, doi:10.1137/S0097539798340217.
- [16] Jaakko Hintikka (1962): *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press.
- [17] Roman Kuznets, Laurent Proserpi, Ulrich Schmid & Krisztina Fruzsa (2019): *Causality and Epistemic Reasoning in Byzantine Multi-Agent Systems*. In Lawrence S. Moss, editor: *TARK 2019: Theoretical Aspects of Rationality and Knowledge, EPTCS 297*, Open Publishing Association, pp. 293–312, doi:10.4204/EPTCS.297.19.
- [18] Roman Kuznets, Laurent Proserpi, Ulrich Schmid & Krisztina Fruzsa (2019): *Epistemic Reasoning with Byzantine-Faulty Agents*. In Andreas Herzig & Andrei Popescu, editors: *FroCoS 2019: Frontiers of Combining Systems, LNCS 11715*, Springer, pp. 259–276, doi:10.1007/978-3-030-29007-8\_15.
- [19] Roman Kuznets, Laurent Proserpi, Ulrich Schmid, Krisztina Fruzsa & Lucas Gréaux (2019): *Knowledge in Byzantine Message-Passing Systems I: Framework and the Causal Cone*. Technical Report TUW-260549, TU Wien. Available at [https://publik.tuwien.ac.at/files/publik\\_260549.pdf](https://publik.tuwien.ac.at/files/publik_260549.pdf).

- [20] Leslie Lamport (1978): *Time, Clocks, and the Ordering of Events in a Distributed System*. *Communications of the ACM* 21(7), pp. 558–565, doi:10.1145/359545.359563.
- [21] Leslie Lamport, Robert Shostak & Marshall Pease (1982): *The Byzantine Generals Problem*. *ACM Transactions on Programming Languages and Systems* 4(3), pp. 382–401, doi:10.1145/357172.357176.
- [22] Yoram Moses (2015): *Relating Knowledge and Coordinated Action: The Knowledge of Preconditions Principle*. In R. Ramanujam, editor: *TARK 2015: Theoretical Aspects of Rationality and Knowledge*, EPTCS 215, Open Publishing Association, pp. 231–245, doi:10.4204/EPTCS.215.17.
- [23] Yoram Moses & Yoav Shoham (1993): *Belief as defeasible knowledge*. *Artificial Intelligence* 64(2), pp. 299–321, doi:10.1016/0004-3702(93)90107-M.
- [24] Yoram Moses & Mark R. Tuttle (1986): *Programming Simultaneous Actions Using Common Knowledge: Preliminary Version*. In: *27th Annual Symposium on Foundations of Computer Science*, IEEE, pp. 208–221, doi:10.1109/SFCS.1986.46.
- [25] Yoram Moses & Mark R. Tuttle (1988): *Programming Simultaneous Actions Using Common Knowledge*. *Algorithmica* 3, pp. 121–169, doi:10.1007/BF01762112.
- [26] Peter Robinson & Ulrich Schmid (2011): *The Asynchronous Bounded-Cycle model*. *Theoretical Computer Science* 412(40), pp. 5580–5601, doi:10.1016/j.tcs.2010.08.001.
- [27] T. K. Srikanth & Sam Toueg (1987): *Optimal Clock Synchronization*. *Journal of the ACM* 34(3), pp. 626–645, doi:10.1145/28869.28876.
- [28] T. K. Srikanth & Sam Toueg (1987): *Simulating authenticated broadcasts to derive simple fault-tolerant algorithms*. *Distributed Computing* 2(2), pp. 80–94, doi:10.1007/BF01667080.
- [29] Alfred Tarski (1955): *A lattice-theoretical fixpoint theorem and its applications*. *Pacific Journal of Mathematics* 5(2), pp. 285–309, doi:10.2140/pjm.1955.5.285.
- [30] Josef Widder & Ulrich Schmid (2009): *The Theta-Model: achieving synchrony without clocks*. *Distributed Computing* 22(1), pp. 29–47, doi:10.1007/s00446-009-0080-x.

## Appendix

*Proof of Lemma 6.* The argument is the same for FIRE and START. We only provide it for the former. Let  $I = (R, \pi)$ . Consider a run  $r \in R$  and a node  $(i, t) \in \mathcal{A} \times \mathbb{T}$ . Assume  $(I, r, t) \models \overline{fire}_i$ . Since  $i$  has perfect recall and this was a correct FIRE action, it was recorded and still remains in  $i$ 's local history  $r(t)$ . Consider any  $r' \in R$  and  $t' \in \mathbb{N}$  such that  $r_i(t) = r'_i(t')$ . Then  $r'_i(t')$  also contains a record of FIRE. If  $(I, r', t') \models correct_i$ , this record must correspond to a correct action and, consequently,  $(I, r', t') \models \overline{fire}_i$ . Since  $(I, r', t') \models correct_i \rightarrow \overline{fire}_i$  whenever  $r_i(t) = r'_i(t')$ , we have  $(I, r, t) \models K_i(correct_i \rightarrow \overline{fire}_i)$ , i.e.,  $(I, r, t) \models B_i \overline{fire}_i$ . The other statement about FIRE follows from  $\models \overline{fire}_i \rightarrow \overline{fire}$  and the monotonicity/normality of  $B_i$ .  $\square$

*Proof of Theorem 10.* Since the system is consistent with (U), (7) holds according to Lemma 7. Thus, given that  $B_i$  is a normal modality, it only remains to show that

$$I \models \overline{fire}_i \rightarrow B_i E^{\diamond H} \overline{start}. \quad (15)$$

Since the system is consistent with (R), (8) holds according to Lemma 9. Using the replacement property for positive subformulas and the already discussed validity  $I \models \overline{fire}_j \rightarrow B_j \overline{start}$  from (7), we obtain  $I \models \overline{fire}_i \rightarrow B_i \bigwedge_{j \in \mathcal{A}} \diamond (correct_j \rightarrow B_j \overline{start})$ , in other words, (15).  $\square$

*Proof of Theorem 14.* Since (7) holds by Lemma 7, it is sufficient to demonstrate  $I \models \overline{\text{fire}}_i \rightarrow B_i C^{\diamond H} \overline{\text{start}}$ . Combining (R) with (4) by applying the replacement property for positive subformulas, we obtain  $I \models \overline{\text{fire}} \rightarrow E^{\diamond H} \overline{\text{fire}}$ . Thus, using (2) with  $\varphi = \psi = \overline{\text{fire}}$ , we conclude  $I \models \overline{\text{fire}} \rightarrow C^{\diamond H} \overline{\text{fire}}$ . Since the greatest fixed point of a monotone operator is itself monotone, it follows from (U) that  $I \models \overline{\text{fire}} \rightarrow C^{\diamond H} \overline{\text{start}}$ . It remains to use (4) and monotonicity of  $B_i$  to obtain  $I \models \overline{\text{fire}}_i \rightarrow B_i C^{\diamond H} \overline{\text{start}}$ .  $\square$

*Proof of Theorem 16.* For either assumption,  $I \models \overline{\text{fire}}_i \rightarrow B_i \overline{\text{start}}$ . Since

$$I \models \overline{\text{fire}}_i \rightarrow \text{correct}_i \quad \text{and} \quad I \models \text{correct}_i \rightarrow (B_i \varphi \rightarrow \varphi) \quad \text{for any formula } \varphi, \quad (16)$$

we have  $I \models \overline{\text{fire}}_i \rightarrow \overline{\text{start}}$  for each  $i \in \mathcal{A}$ . Since  $\overline{\text{fire}}$  is  $\bigvee_{i \in \mathcal{A}} \overline{\text{fire}}_i$ , (U) holds by propositional reasoning.

It remains to show that Relay (R) holds under the assumption of (10). Once again, it is sufficient to demonstrate that, for each  $i \in \mathcal{A}$ ,

$$I \models \overline{\text{fire}}_i \rightarrow \bigwedge_{j \in \mathcal{A}} \diamond(\text{correct}_j \rightarrow \overline{\text{fire}}_j). \quad (17)$$

It follows from the first conjunct of (10) that  $I \models \overline{\text{fire}}_i \rightarrow B_i C^{\diamond H} \overline{\text{start}}$ . Using (16) again, we conclude that  $I \models \overline{\text{fire}}_i \rightarrow C^{\diamond H} \overline{\text{start}}$ . Since  $I \models C^{\diamond H} \varphi \rightarrow \bigwedge_{j \in \mathcal{A}} \diamond H_j(\varphi \wedge C^{\diamond H} \varphi)$  for any formula  $\varphi$  according to (1),

$$I \models \overline{\text{fire}}_i \rightarrow \bigwedge_{j \in \mathcal{A}} \diamond(\text{correct}_j \rightarrow B_j(\overline{\text{start}} \wedge C^{\diamond H} \overline{\text{start}})). \quad (18)$$

Using the second conjunct of (10) and monotonicity of  $B_j$  and  $\diamond$  in (18), we obtain

$$I \models \overline{\text{fire}}_i \rightarrow \bigwedge_{j \in \mathcal{A}} \diamond(\text{correct}_j \rightarrow \diamond(\text{correct}_j \rightarrow \overline{\text{fire}}_j)).$$

To get (17), it remains to note that  $I \models \diamond(\varphi \rightarrow \diamond(\varphi \rightarrow \psi)) \rightarrow \diamond(\varphi \rightarrow \psi)$  for all formulas  $\varphi$  and  $\psi$ .  $\square$

*Proof of Lemma 20.* Let  $I = (R, \pi)$ . Assume that  $(I, r, t) \not\models K_i \neg \varphi$  for some  $r \in R$  and  $t \in \mathbb{T}$ . Then there exists another run  $r' \in R$  and time  $t' \in \mathbb{T}$  such that  $r_i(t) = r'_i(t')$  and  $(I, r', t') \models \varphi$ . By the potential persistence of  $\varphi$ , there exists an alternative continuation  $r'' \in R$  of the prefix  $r'(t')$  such that  $r''(t') = r'(t')$  and  $(I, r'', t') \models \square \varphi$ . Thus,  $(I, r'', t') \not\models \diamond \neg \varphi$ . It remains to note that  $r''_i(t') = r'_i(t') = r_i(t)$ . Hence,  $(I, r, t) \not\models K_i \diamond \neg \varphi$ .  $\square$

*Proof of Lemma 21.* The right-to-left direction is trivial. Hence, we prove the implication from left to right. Firstly,  $\neg \text{correct}_i \rightarrow (\text{correct}_i \rightarrow \varphi)$  is a propositional tautology. Hence,

$$I \models \square \neg \text{correct}_i \rightarrow \square(\text{correct}_i \rightarrow \varphi).$$

Using Lemma 19, the fact that  $I \models \square \psi \rightarrow \diamond \psi$  by seriality of temporal modalities, and knowledge necessitation, we obtain

$$I \models K_i(\neg \text{correct}_i \rightarrow \diamond(\text{correct}_i \rightarrow \varphi)).$$

By epistemically internalized propositional reasoning,

$$I \models K_i(\text{correct}_i \rightarrow \diamond(\text{correct}_i \rightarrow \varphi)) \wedge K_i(\neg \text{correct}_i \rightarrow \diamond(\text{correct}_i \rightarrow \varphi)) \rightarrow K_i \diamond(\text{correct}_i \rightarrow \varphi).$$

Since we have just shown the second conjunct above to be valid, we obtain the desired

$$I \models K_i(\text{correct}_i \rightarrow \diamond(\text{correct}_i \rightarrow \varphi)) \rightarrow K_i \diamond(\text{correct}_i \rightarrow \varphi). \quad \square$$

*Proof of Lemma 23.* By Corollary 22,  $I \models B_i \diamond H_i \overline{start} \rightarrow K_i \diamond H_i \overline{start}$ . Applying factivity of knowledge and propositional reasoning to the expanded version of  $K_i \diamond H_i \overline{start}$  yields

$$I \quad \models \quad K_i \diamond (correct_i \rightarrow K_i (correct_i \rightarrow \overline{start})) \rightarrow K_i \diamond (correct_i \rightarrow \overline{start}).$$

Since  $correct_i \wedge \neg \overline{start}$  is potentially persistent, and its negation is equivalent to  $correct_i \rightarrow \overline{start}$ , we have by Lemma 20 that

$$I \quad \models \quad K_i \diamond (correct_i \rightarrow \overline{start}) \rightarrow K_i (correct_i \rightarrow \overline{start})$$

Combining all implications together, we conclude that  $I \models B_i \diamond H_i \overline{start} \rightarrow B_i \overline{start}$ .  $\square$

*Proof of Lemma 26.* Let  $I = (R, \pi)$ . Assume  $(I, r, t) \models E^{\diamond H} \overline{start}$  for some  $r \in R$  and time  $t \in \mathbb{T}$ . This means that, for every agent  $j \in \mathcal{A}$ , there is some  $t'_j \geq t$  such that  $(I, r, t'_j) \models H_j \overline{start}$ . Since  $|\mathcal{A}| \geq 3f + 1$ , it follows that there exists a group  $G$  of  $2f + 1$  correct agents such that  $(I, r, t'_j) \models H_j \overline{start}$  for all  $j \in G$ . Since these agents are correct,<sup>3</sup> we have  $(I, r, t'_j) \models B_j \overline{start}$ , i.e.,  $(I, r, t'_j) \models K_j (correct_j \rightarrow \overline{start})$  for all  $j \in G$ . Let  $t' := \max\{t'_j \mid j \in G\}$ . We claim that

$$(I, r, t') \quad \models \quad \bigwedge_{j \in G} K_j (correct_j \rightarrow \overline{start}). \quad (19)$$

Indeed, for any agent  $j \in G$  consider any alternative run  $\bar{r} \in R$  and time  $\bar{t}' \in \mathbb{T}$  such that  $\bar{r}_j(\bar{t}') = r_j(t')$ . Given that  $t' \geq t'_j$  and our agents have perfect recall, there must exist some time  $\bar{t}'_j \leq \bar{t}'$  such that  $\bar{r}_j(\bar{t}'_j) = r_j(t'_j)$ . Thus,  $(I, \bar{r}, \bar{t}'_j) \models correct_j \rightarrow \overline{start}$ . Since the latter formula is stable, it remains true in  $\bar{r}$  by the time  $\bar{t}'$ . We showed that  $(I, \bar{r}, \bar{t}') \models correct_j \rightarrow \overline{start}$  whenever  $\bar{r}_j(\bar{t}') = r_j(t')$ , meaning  $(I, r, t') \models K_j (correct_j \rightarrow \overline{start})$ . This argument applies to every  $j \in G$ , hence, (19) is demonstrated for the group  $G$  of  $2f + 1$  correct agents.

Correctness (C) applied to  $G$  at time  $t'$  ensures  $(I, r, t') \models \bigwedge_{i \in \mathcal{A}} \diamond (correct_i \rightarrow \overline{fire}_i)$ , and, since  $t \leq t'$ , we also have

$$(I, r, t) \quad \models \quad \bigwedge_{i \in \mathcal{A}} \diamond (correct_i \rightarrow \overline{fire}_i).$$

Given that  $r$  and  $t$  were chosen arbitrarily, we have proved

$$I \quad \models \quad E^{\diamond H} \overline{start} \rightarrow \bigwedge_{i \in \mathcal{A}} \diamond (correct_i \rightarrow \overline{fire}_i).$$

Using (13), we can conclude

$$I \quad \models \quad E^{\diamond H} \overline{start} \rightarrow \bigwedge_{i \in \mathcal{A}} \diamond (correct_i \rightarrow B_i (\overline{start} \wedge E^{\diamond H} \overline{start})),$$

i.e.,

$$I \quad \models \quad E^{\diamond H} \overline{start} \rightarrow \bigwedge_{i \in \mathcal{A}} \diamond H_i (\overline{start} \wedge E^{\diamond H} \overline{start}).$$

In other words, we have demonstrated

$$I \quad \models \quad E^{\diamond H} \overline{start} \rightarrow E^{\diamond H} (\overline{start} \wedge E^{\diamond H} \overline{start}).$$

Using (2) with  $\psi = E^{\diamond H} \overline{start}$  and  $\varphi = \overline{start}$ , we conclude

$$I \quad \models \quad E^{\diamond H} \overline{start} \rightarrow C^{\diamond H} \overline{start}. \quad \square$$

<sup>3</sup>While we only use the fact that agent  $j \in G$  is correct at  $t'_j$ , these agents will necessarily remain correct throughout run  $r$ .



# Are the Players in an Interactive Belief Model Meta-certain of the Model Itself? (Extended Abstract)

Satoshi Fukuda

Department of Decision Sciences and IGIER, Bocconi University\*  
Milan, Italy

satoshi.fukuda@unibocconi.it

In an interactive belief model, are the players “commonly meta-certain” of the model itself? This paper formalizes such implicit “common meta-certainty” assumption. To that end, the paper expands the objects of players’ beliefs from events to functions defined on the underlying states. Then, the paper defines a player’s belief-generating map: it associates, with each state, whether a player believes each event at that state. The paper formalizes what it means by: “a player is (meta-)certain of her own belief-generating map” or “the players are (meta-)certain of the profile of belief-generating maps (i.e., the model).” The paper shows: a player is (meta-)certain of her own belief-generating map if and only if her beliefs are introspective. The players are commonly (meta-)certain of the model if and only if, for any event which some player  $i$  believes at some state, it is common belief at the state that player  $i$  believes the event. This paper then asks whether the “common meta-certainty” assumption is needed for an epistemic characterization of game-theoretic solution concepts. The paper shows: if each player is logical and (meta-)certain of her own strategy and belief-generating map, then each player correctly believes her own rationality. Consequently, common belief in rationality alone leads to actions that survive iterated elimination of strictly dominated actions.

## 1 Introduction

In an economic or game-theoretic model in which the players make their interactive reasoning about their strategies or rationality, the analysts “from outside of the model” assume that the players “commonly know” the model itself. Since the pioneering work of [1, 2, 3], how to model such assumption and what consequences such assumption has have been puzzling economic and game theorists.<sup>1</sup>

This paper has two objectives. The first aims at formalizing the “common knowledge” assumption of a model within the model itself. An interactive belief/knowledge model represents players’ beliefs/knowledge about its ingredients, that is, events. The model itself does not tell whether the players (commonly) believe/know the model itself, although the analysts assume that the players (commonly) believe/know the model in a meta-sense. I refer to the knowledge/belief of the model as the “meta-knowledge/meta-belief” of the model.<sup>2</sup>

The second objective is to examine the role that “meta-knowledge” of a model plays in game-theoretic analyses such as epistemic characterizations of solution concepts. For a given epistemic characterization of a game-theoretic solution concept such as iterated elimination of strictly dominated actions, do the outside analysts need to formally assume that the players “meta-know” an epistemic model of a game (that describes their interactive beliefs about their strategies and rationality)?

---

\*I would like to thank three anonymous referees of TARK 2021 for their helpful comments. This is an abbreviated technical summary of part of the full paper, which is available at the author’s website (<https://websfukuda.com/research/>).

<sup>1</sup>For this question, see also [4, 5], [6], [10, 11], [12], [13], [15], [16], [21], [25], [26], [30], [32], [33], [31], and [34].

<sup>2</sup>Since different epistemic models may feature different notions of qualitative or probabilistic beliefs or knowledge, I use the word the “(meta-)certainty” of a model to refer generically to the meta-knowledge or meta-belief of the model.

The first main result, presented in Section 4, characterizes the “common meta-certainty” assumption as follows. According to the formal test to be discussed, the players are commonly (meta-)certain of a model if and only if, for any event which some player  $i$  believes at some state, it is common belief that player  $i$  believes the event at that state.

In Section 2, I start with introducing a (belief) model. The model consists of the following three ingredients. The first is a measurable space of states of the world  $(\Omega, \mathcal{D})$ . Each state  $\omega \in \Omega$  is a list of possible specifications of the world, and the collection  $\mathcal{D}$  of events (i.e., subsets of  $\Omega$ ) are the objects of the players’ beliefs. The second is the players’ monotone belief operators  $(B_i)_{i \in I}$ . Player  $i$ ’s belief operator  $B_i$  associates, with each event  $E$ , the event that she believes  $E$ . Monotonicity means that if player  $i$  believes  $E$  at a state and if  $E$  implies (i.e., is included in)  $F$ , then she believes  $F$  at that state. The third is a common belief operator  $C$ , which associates, with each event  $E$ , the event that the players commonly believe  $E$ . Under certain assumptions on the players’ beliefs, an event  $E$  is common belief if and only if everybody believes  $E$ , everybody believes that everybody believes  $E$ , and so on *ad infinitum*.

The framework nests the following two standard models of belief or knowledge (or combinations thereof). The first is a possibility correspondence model of qualitative belief or knowledge (e.g., [1, 3], [15], [20], and [24]). The possibility correspondence associates, with each state, the set of states that she considers possible. The player believes an event  $E$  at a state when the possibility set at  $\omega$  is included in the event  $E$ . The second is a type space ([22]), where each player’s probabilistic beliefs are induced by her type mapping. The type mapping  $\tau_i$  associates, with each state  $\omega$ , her probability measure  $\tau_i(\omega)$  on the underlying states at that state. The type mapping  $\tau_i$  of player  $i$  induces her  $p$ -belief operator ([23]): it associates, with each event  $E$ , the event that player  $i$  believes  $E$  with probability at least  $p$ .

With the framework in mind, I formalize the (meta-)certainty of a model in two steps. In the first step, Section 3.1 expands the objects of the players’ beliefs from events to functions defined on the underlying states. Examples of such functions are random variables, strategies, and type mappings. Any such function  $x$  has to be defined on the state space  $\Omega$ , but the co-domain  $X$  can be any set such as the set  $\mathbb{R}$  of real numbers (a random variable), a set  $A_i$  of player  $i$ ’s actions (her strategy), and the set  $\Delta(\Omega)$  of probability measures on  $(\Omega, \mathcal{D})$  (a type mapping). I call the function  $x : \Omega \rightarrow X$  a signal if its co-domain  $X$  has “observational” contents  $\mathcal{X}$  (where “observation” is broadly construed as being an object of reasoning): it is a collection of subsets of  $X$  such that each  $F \in \mathcal{X}$  is deemed an event  $x^{-1}(F)$ . Formally, a *signal (mapping)* is a function  $x : (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  such that each observation  $F \in \mathcal{X}$  is considered to be an event  $x^{-1}(F) \in \mathcal{D}$ . Player  $i$  is *certain* of the value of the signal  $x$  at a state  $\omega$  if, for any observational content  $F$  that holds at  $\omega$  (i.e.,  $\omega \in x^{-1}(F)$ ), player  $i$  believes the event  $x^{-1}(F)$  at  $\omega$  (i.e.,  $\omega \in B_i(x^{-1}(F))$ ). Player  $i$  is *certain* of  $x$  if she is certain of the value of  $x$  at every state. For example, let  $x : \Omega \rightarrow X$  be the strategy of player  $i$  and let every singleton action  $a \in X$  be observable to her; then, player  $i$  is certain of her own strategy if, wherever she takes an action  $a = x(\omega)$  at a state  $\omega$ , she believes at  $\omega$  that she takes action  $a$ . Having defined individual players’ (meta-)certainty, the players are *commonly certain* of the value of the signal  $x$  at a state  $\omega$  if, for any observational content  $F$  that holds at  $\omega$ , the event  $x^{-1}(F)$  is common belief at  $\omega$  (i.e.,  $\omega \in C(x^{-1}(F))$ ). The players are *commonly certain* of the signal  $x$  if they are commonly certain of its value at every state.

In the second step, Section 3.2 formulates a players’ “belief-generating map” as a signal that associates, with each state, her beliefs at that state. By the second step, I can apply the formalization of certainty and common certainty in the first step to the ingredients of a given model (i.e., players’ belief-generating maps). To that end, take player  $i$ ’s belief operator  $B_i$  from the model. I define a qualitative-type mapping  $t_{B_i}$ : it associates, with each state, whether player  $i$  believes each event or not at that state (formally, a binary set function from the collection of events to the binary values  $\{0, 1\}$  where 1 indicates the belief of an event). The qualitative-type mapping is a binary “type” mapping analogous to a type map-

ping  $\tau_i$  that represents player  $i$ 's probabilistic beliefs at each state in the context of probabilistic beliefs. Thus, the qualitative-type mapping  $t_{B_i}$  associates, with each state  $\omega$ , her qualitative belief  $t_{B_i}(\omega) \in \{0, 1\}$  (where  $t_{B_i}(\omega)(E) = 1$  if and only if  $\omega \in B_i(E)$ ) on  $(\Omega, \mathcal{D})$  at  $\omega$ . The qualitative-type mapping  $t_{B_i}$  is player  $i$ 's belief-generating mapping. Since the belief operator  $B_i$  and the qualitative-type mapping  $t_{B_i}$  are equivalent means of representing player  $i$ 's beliefs, a model means the profile of qualitative-type mappings. Thus, the formal test for whether the players are commonly certain of a given belief model is whether the players are certain of the profile of their qualitative-type mappings.

Before asking when a player is certain of all the players' qualitative-type mappings (i.e., the model), Section 3.3 characterizes when a player is certain of her own qualitative type-mapping in terms of her introspective properties of beliefs. Roughly, Proposition 1 shows that each player is certain of her own qualitative-type mapping if and only if her belief is introspective.

To the best of my knowledge, this is the first paper which systematically formalizes the statement that the players are (commonly) (meta-)certain of any given belief model within the model itself. The main result on this question, nevertheless, is related to [21]. He constructs a particular syntactic model in which the statement that the model is common knowledge is incorporated within itself. He formulates the sense in which the model is commonly known from Positive Introspection of common knowledge: if a statement is common knowledge then it is commonly known that the statement is common knowledge. In Theorem 1, in contrast, the players are commonly certain of a given model if and only if, at each state and for any event which some player believes at that state, it is common belief that the player believes the event at that state. Thus, in this paper, the key criteria is the positive introspective property of common belief with respect to each player's beliefs. Whenever some individual player believes some event, it is common belief that she believes it. [4, 5], in the context of partitional possibility correspondence models, formalizes the event that a player has an information partition by regarding it as a function.

Having characterized the common-certainty of the model, Section 5 examines the role that the "common meta-certainty" assumption plays in the epistemic characterization of iterated elimination of strictly dominated actions (IESDA) in a strategic game. Informally, it states: if the players are "logical," if they are commonly meta-certain of a game, and if they commonly believe their rationality, then the resulting actions survive any process (i.e., any order) of IESDA. Formally, it states: if the players commonly believe their rationality and if their common belief in their rationality is correct, then the resulting actions survive any process of IESDA.<sup>3</sup> Theorem 2 connects these two statements. If the players' beliefs are monotone (they believe any logical implication of their beliefs), consistent (i.e., they do not simultaneously believe an event and its negation), and finitely conjunctive (if they believe  $E$  and  $F$  then they believe its conjunction  $E \cap F$ ), and if each player is certain of her own strategy and the part of her own belief-generating process in the model (each player is not necessarily certain of how the opponents' beliefs are generated in the model), then each player correctly believes her own rationality, and hence they have correct common belief in their rationality. Thus, if the players are "logical" and each of them is meta-certain of the part of the model that governs her own beliefs, then common belief in rationality leads to actions that survive any process of IESDA.

On the one hand, Theorem 1 states that the players may not always be commonly certain of a belief model in the sense that each player is commonly certain of how the other players' beliefs are represented within the model. On the other hand, Theorem 2 asserts that common belief in rationality leads to IESDA even if the players may not be commonly certain of the belief model.

The paper is organized as follows. Section 2 defines a belief model. Section 3 characterizes the sense

---

<sup>3</sup>The formal statement is taken from [18, Theorem 3], which holds irrespective of the nature of beliefs. For seminal papers on implications of common belief in rationality, see, for example, [9], [29], and [30].

in which each player is certain of how her belief is generated in a model. Section 4 examines the sense in which the players are commonly certain of a model itself (i.e., how the players' beliefs are generated in the model). Section 5 studies how the assumption that the players are commonly certain of a model itself can make game-theoretic analyses coherent.

## 2 Framework

Throughout the paper, let  $I$  denote a non-empty finite set of *players*. The framework represents players' interactive beliefs by belief operators on a state space. Section 2.1 defines a belief model. Section 2.2 defines properties of beliefs.

### 2.1 A Belief Model

A *belief model* is a tuple  $\vec{\Omega} := \langle (\Omega, \mathcal{D}), (B_i)_{i \in I}, C \rangle$ , where: (i)  $(\Omega, \mathcal{D})$  is a non-empty measurable space of states of the world (call  $\Omega$  the *state space*); (ii)  $B_i : \mathcal{D} \rightarrow \mathcal{D}$  is player  $i$ 's (monotone) *belief operator*; and (iii)  $C : \mathcal{D} \rightarrow \mathcal{D}$  is a (monotone) *common belief operator* to be defined below.

While  $\Omega$  constitutes a non-empty set of *states* of the world, each element  $E$  of  $\mathcal{D}$  is an *event* about which the players reason. For each event  $E$ , the set  $B_i(E)$  denotes the event that (i.e., the set of states at which) a player  $i$  believes  $E$ . Thus, the player  $i \in I$  believes an event  $E \in \mathcal{D}$  at a state  $\omega \in \Omega$  if  $\omega \in B_i(E)$ . I assume that each player's belief operator satisfies *Monotonicity*:  $E \subseteq F$  implies  $B_i(E) \subseteq B_i(F)$ . It means that if player  $i$  believes some event then she believes any of its logical consequences.<sup>4</sup>

Since the players' beliefs are monotone, I introduce the common belief operator  $C : \mathcal{D} \rightarrow \mathcal{D}$  following [23]. Call an event  $E$  *publicly evident* if  $E \subseteq B_I(E) := \bigcap_{i \in I} B_i(E)$ . That is, everybody believes  $E$  whenever  $E$  is true. Denote by  $\mathcal{J}_{B_I}$  the collection of publicly-evident events. An event  $E$  is *common belief* at a state  $\omega$  if there is a publicly-evident event that is true at  $\omega$  and that implies the mutual belief in  $E$ : that is,  $\omega \in F \subseteq B_I(E)$  for some  $F \in \mathcal{J}_{B_I}$ . Now,  $C$  is assumed to satisfy that the set of states at which  $E$  is common belief is an event for each  $E \in \mathcal{D}$ :  $C(E) := \{\omega \in \Omega \mid \text{there is } F \in \mathcal{J}_{B_I} \text{ with } \omega \in F \subseteq B_I(E)\} \in \mathcal{D}$ .

Since players' beliefs are monotone and since  $\mathcal{D}$  is closed under countable intersection, if  $E$  is common belief, then everybody believes  $E$ , everybody believes that everybody believes  $E$ , and so forth *ad infinitum*:  $C(E) \subseteq \bigcap_{n \in \mathbb{N}} B_I^n(E)$ . The converse (set inclusion) holds, for example, when the mutual belief operator  $B_I$  (or every  $B_i$ ) satisfies *Countable Conjunction*:  $\bigcap_{n \in \mathbb{N}} B_I(E_n) \subseteq B_I(\bigcap_{n \in \mathbb{N}} E_n)$ , meaning that everybody believes the countable conjunction of events whenever everybody believes each of them.

I represent the players' beliefs on a measurable space  $(\Omega, \mathcal{D})$  so that I can analyze players' qualitative and probabilistic beliefs under the same framework. The full paper introduces the players' probabilistic beliefs on a measurable space  $(\Omega, \mathcal{D})$  by  $p$ -belief operators [23]. For each  $p \in [0, 1]$ , player  $i$ 's  $p$ -belief operator  $B_i^p$  associates, with each event  $E$ , the event that player  $i$  believes  $E$  with probability at least  $p$ . The full paper also introduces the common  $p$ -belief operator  $C^p$ .

### 2.2 Properties of Beliefs

Next, I introduce additional eight properties of beliefs. Various possibility correspondence models of qualitative beliefs and knowledge are represented as belief models that satisfy certain properties specified below. Fix a player  $i$ . I first introduce the following five logical properties of beliefs.

<sup>4</sup>The full paper dispenses with the monotonicity assumption, which enables one to define the meta-certainty of any belief model.

1. *Necessitation*:  $B_i(\Omega) = \Omega$ .
2. *Countable Conjunction*:  $\bigcap_{n \in \mathbb{N}} B_i(E_n) \subseteq B_i(\bigcap_{n \in \mathbb{N}} E_n)$  (for any events  $(E_n)_{n \in \mathbb{N}}$ ).
3. *Finite Conjunction*:  $B_i(E) \cap B_i(F) \subseteq B_i(E \cap F)$ .
4. The *Kripke property*:  $B_i(E) = \{\omega \in \Omega \mid b_{B_i}(\omega) \subseteq E\}$ , where  $b_{B_i}(\omega) := \bigcap \{E \in \mathcal{D} \mid \omega \in B_i(E)\}$  is the set of states player  $i$  considers *possible* at  $\omega$ .
5. *Consistency*:  $B_i(E) \subseteq (\neg B_i)(E^c)$ .

First, Necessitation means that the player believes a tautology such as  $E \cup E^c$ . Second, as discussed, Countable Conjunction means that if the player believes each of a countable collection of events, then she believes its conjunction. Third, Finite Conjunction is weaker than Countable Conjunction: if the player believes  $E$  and  $F$  then she believes its conjunction  $E \cap F$ . Fourth, to discuss the Kripke property, the player considers  $\omega'$  possible at  $\omega$  if, for any event  $E$  which she believes at  $\omega$ ,  $E$  is true at  $\omega'$ . The Kripke property provides the condition under which  $i$ 's belief is induced by her *possibility correspondence*  $b_{B_i} : \Omega \rightarrow \mathcal{P}(\Omega)$ : she believes  $E$  at  $\omega$  if and only if (hereafter, iff) her possibility set  $b_{B_i}(\omega)$  at  $\omega$  implies  $E$ . The Kripke property implies the previous three properties as well as Monotonicity. Fifth, Consistency means that the player cannot simultaneously believe an event  $E$  and its negation  $E^c$ .

Next, I move on to truth and introspective properties.

6. *Truth Axiom*:  $B_i(E) \subseteq E$  (for all  $E \in \mathcal{D}$ ).
7. *Positive Introspection*:  $B_i(\cdot) \subseteq B_i B_i(\cdot)$  (i.e.,  $B_i(E) \subseteq B_i B_i(E)$  for all  $E \in \mathcal{D}$ ).
8. *Negative Introspection*:  $(\neg B_i)(\cdot) \subseteq B_i(\neg B_i)(\cdot)$ .

Sixth, Truth Axiom turns belief into knowledge in that knowledge has to be true while belief can be false. Truth Axiom implies Consistency. While knowledge satisfies Truth Axiom, qualitative and probabilistic beliefs are often assumed to satisfy Consistency. Seventh, Positive Introspection states that if the player believes some event then she believes that she believes it. Eighth, Negative Introspection states that if the player does not believe some event then she believes that she does not believe it. Truth Axiom and Negative Introspection yield Positive Introspection (e.g., [3]).

Three remarks are in order. First, the introspective properties will play important roles in whether a player is (meta-)certain of a belief model. Intuitively, Positive Introspection provides the sense in which the player believes her own belief (at least at face value) while Negative Introspection yields the sense in which the player believes the lack of her own belief. To see these points, an event  $E$  is *self-evident* to  $i$  if  $E \subseteq B_i(E)$ . That is,  $i$  believes  $E$  whenever  $E$  is true. Positive Introspection means that  $i$ 's belief in  $E$  is self-evident to  $i$ , and Negative Introspection means that  $i$ 's lack of belief in  $E$  is self-evident to  $i$ . Denote by  $\mathcal{J}_{B_i}$  the collection of self-evident events to  $i$ .

Second, the last four properties are restated in terms of  $b_{B_i}$  under the Kripke property:  $B_i$  satisfies Consistency iff  $b_{B_i}$  is serial (i.e.,  $b_{B_i}(\cdot) \neq \emptyset$ );  $B_i$  satisfies Truth Axiom iff  $b_{B_i}$  is reflexive (i.e.,  $\omega \in b_{B_i}(\omega)$  for all  $\omega \in \Omega$ );  $B_i$  satisfies Positive Introspection iff  $b_{B_i}$  is transitive (i.e.,  $\omega' \in b_{B_i}(\omega)$  implies  $b_{B_i}(\omega') \subseteq b_{B_i}(\omega)$ ); and  $B_i$  satisfies Negative Introspection iff  $b_{B_i}$  is Euclidean (i.e.,  $\omega' \in b_{B_i}(\omega)$  implies  $b_{B_i}(\omega) \subseteq b_{B_i}(\omega')$ ).

Third, various models of probabilistic and qualitative beliefs and knowledge take different sets of axioms. The framework accommodates possibility correspondence models of qualitative beliefs and knowledge when  $B_i$  satisfies the Kripke property. A partitional model of knowledge corresponds to the case when  $B_i$  satisfies Truth Axiom, Positive Introspection, and Negative Introspection.<sup>5</sup> A reflexive

<sup>5</sup>In fact, Truth Axiom, Negative Introspection, and the Kripke property yield all the other properties defined in this section.

and transitive (non-partitional) possibility correspondence model is characterized by Truth Axiom and Positive Introspection.<sup>6</sup> When it comes to fully-introspective qualitative beliefs,  $b_{B_i}$  is serial, transitive, and Euclidean iff  $B_i$  satisfies Consistency, Positive Introspection, and Negative Introspection.

### 3 When Is a Player Certain of Her Belief-Generating Mapping?

Section 3.1 extends an object of beliefs in a model from an event to a function (“signal”) defined on the state space. That is, the subsection formulates the statement that a player is certain of a function defined on the state space. Section 3.2 represents a player’s “belief-generating mapping” as a signal which associates, with each state, whether she believes each event or not. Section 3.3 asks the sense in which she is certain of her own belief-generating mapping in terms of the introspective properties.

#### 3.1 Functions as Objects of Players’ Beliefs

I start with defining a notion of a signal mapping. A signal mapping is any function  $x$  defined on the state space  $\Omega$  with “observational” contents. A signal is interpreted as a mapping from the underlying state space  $\Omega$  into a space of “observation”  $X$  endowed with “observational” contents  $\mathcal{X} \subseteq \mathcal{P}(X)$ . By observation, it means that each  $F \in \mathcal{X}$  is deemed an object of reasoning. That is, we call a mapping  $x : \Omega \rightarrow X$  a signal mapping if each “observational” content  $F \in \mathcal{X}$  can be regarded as an event  $x^{-1}(F) \in \mathcal{D}$  through inverting the mapping.

Formally, for a non-empty set  $X$  and a non-empty subset  $\mathcal{X}$  of  $\mathcal{P}(X)$ , call  $x : (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  a *signal* (mapping) if  $x^{-1}(\mathcal{X}) \subseteq \mathcal{D}$ . Mathematically,  $x : (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  is a signal if  $x : (\Omega, \mathcal{D}) \rightarrow (X, \sigma(\mathcal{X}))$  is measurable. Examples include strategies, random variables, and so on.

The main purpose of this subsection is to define the statement that a player is certain of a signal. A player  $i$  is *certain of the value of* a signal  $x : (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  at  $\omega$ , if she believes any observational content  $F$  (i.e., believes  $x^{-1}(F)$ ) at  $\omega$  whenever it is true:  $x(\omega) \in F$ . She is *certain of* the signal  $x : (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  if she is certain of its value at every  $\omega$ . Likewise, the players are *commonly certain of the value of* the signal  $x : (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  at  $\omega$ , if the players commonly believe any observational content  $F$  at  $\omega$  whenever it is true. The players are *commonly certain of* the signal  $x : (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  if they are certain of its value at every  $\omega$ . Formally:

**Definition 1.** Let  $\vec{\Omega}$  be a belief model, and let  $x : (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  be a signal mapping.

1. (a) Player  $i$  is *certain of the value of the signal*  $x : (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  at  $\omega$  if  $\omega \in B_i(x^{-1}(F))$  for any  $F \in \mathcal{X}$  with  $x(\omega) \in F$ .  
 (b) Player  $i$  is *certain of the signal*  $x : (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  if she is certain of the value of the signal  $x$  at any state.
2. (a) The players are *commonly certain of the value of the signal*  $x : (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  at  $\omega$  if  $\omega \in C(x^{-1}(F))$  for any  $F \in \mathcal{X}$  with  $x(\omega) \in F$ .  
 (b) The players are *commonly certain of* the signal  $x : (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  if they are commonly certain of the value of the signal  $x$  is at every state.

For example, suppose that  $x : \Omega \rightarrow X$  is a decision function of a player which associates, with each state, the action taken at that state. Suppose the set of actions  $X$  is endowed with the collection of singleton actions  $\mathcal{X} = \{\{a\} \mid a \in X\}$ . Each action  $a$  corresponds to an observational content to the

<sup>6</sup>See, for example, [4], [6], [15], [20], [25], [27], and [28].

player, and  $x$  is a signal mapping if the set of states at which the player takes action  $a$  is an event:  $x^{-1}(\{a\}) = \{\omega \in \Omega \mid x(\omega) = a\} \in \mathcal{D}$  for each  $a \in X$ .

More specifically, let  $\Omega = \{\omega_1, \omega_2, \omega_3\}$  and  $X = \{a, b\}$ . For each  $i \in I = \{1, 2\}$ , let  $B_i$  be given by (i)  $B_i(E) = E \setminus \{\omega_3\}$  for each  $E \neq \Omega$ ; and (ii)  $B_i(\Omega) = \Omega$ . Suppose player 1's decision function  $x : (\Omega, \mathcal{P}(\Omega)) \rightarrow (X, \{\{a\}, \{b\}\})$  is given by  $(x(\omega))_{\omega \in \Omega} = (a, a, a)$ . Since  $B_1(\Omega) = \Omega$  and  $B_1(\emptyset) = \emptyset$ , whenever player 1 takes a certain action, she believes that she takes that action. Thus, player 1 is certain of  $x$ . If, instead, her decision function  $x : (\Omega, \mathcal{P}(\Omega)) \rightarrow (X, \{\{a\}, \{b\}\})$  is given by  $(x(\omega))_{\omega \in \Omega} = (a, b, a)$ , then at  $\omega_3$  at which she takes action  $a$ , she does not believe that she takes action  $a$ , because  $B_1(\{\omega_1, \omega_3\}) = \{\omega_1\}$ . Thus, player 1 is not certain of the value of  $x$  at  $x_3$ . Since  $C = B_1$ , the same arguments hold for the common certainty of  $x$ .

For ease of terminology, player  $i$  is *certain of (the value of) the signal  $x : \Omega \rightarrow X$  (at  $\omega$ ) with respect to  $\mathcal{X}$*  if she is certain of (the value of) the signal  $x : (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  (at  $\omega$ ). Likewise, the players are *commonly certain of (the value of) the signal  $x : \Omega \rightarrow X$  (at  $\omega$ ) with respect to  $\mathcal{X}$*  if they are commonly certain of (the value of) the signal  $x : (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  (at  $\omega$ ).

Four remarks on Definition 1 are in order. First, I restate the fact that a player is certain of a signal in terms of self-evidence. Namely, player  $i$  is certain of a signal  $x : (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  iff any observational content  $F \in \mathcal{X}$  (i.e., any event  $x^{-1}(F) \in \mathcal{D}$ ) is self-evident to  $i$ . Likewise, the players are commonly certain of the signal  $x : (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  iff any observational content  $F \in \mathcal{X}$  is publicly-evident. Consequently, the players are commonly certain of a signal iff every player is certain of it.<sup>7</sup>

Second, when  $x : (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  is a player's strategy, Definition 1 formalizes the statement that the player is certain of the strategy (e.g., [12] and [20]). To see this, assume that  $\mathcal{X}$  contains a singleton  $\{x(\omega)\}$  to reason about the action taken at  $\omega$ . That is, the set of states  $[x(\omega)] := x^{-1}(\{x(\omega)\}) = \{\omega' \in \Omega \mid x(\omega') = x(\omega)\}$  at which player  $i$  takes the same action as she does at  $\omega$  is an event. Then, player  $i$  is certain of her strategy iff  $[x(\omega)]$  is self-evident at every  $\omega \in \Omega$ . In words, player  $i$  is certain of her strategy  $x$  iff, whenever she takes action  $a = x(\omega)$  at  $\omega$ , she believes at  $\omega$  that she takes action  $a = x(\omega)$ .

Definition 1 also subsumes the formulation of the certainty of the strategy by [2] in the (countable) partitional state space model of knowledge. Let  $(b_{B_i}(\omega))_{\omega \in \Omega}$  be a countable partition on  $\Omega$ . In [2], the player "knows" her own strategy  $x$  iff the strategy  $x$  is measurable with respect to the partition (which turns out to be equivalent to  $b_{B_i}(\cdot) \subseteq [x(\cdot)]$ ). Since the partition is countable, the  $\sigma$ -algebra generated by the partition is equal to the self-evident collection:  $\mathcal{J}_{B_i} = \sigma(\{b_{B_i}(\omega) \in \mathcal{D} \mid \omega \in \Omega\})$ . Hence, player  $i$  is certain of her strategy  $x : (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  iff  $x : (\Omega, \mathcal{J}_{B_i}) \rightarrow (X, \sigma(\mathcal{X}))$  is measurable.

Third, player  $i$  satisfies Necessitation iff she is certain of any constant signal. Likewise, the common belief operator  $C$  satisfies Necessitation (equivalently, every  $B_i$  satisfies Necessitation) iff the players are commonly certain of any constant signal.

Necessitation allows the players to be certain of any constant "random" variable that does not depend on the realization of a state. For example, consider whether player  $i$  is certain that an event  $B_j(E)$  is equal to an event  $F$  in a belief model. The outside analysts determine whether player  $i$  believes that player  $j$  believes an event  $E$  at a state  $\omega$  by examining whether  $\omega \in B_i B_j(E)$  since player  $j$ 's belief  $B_j(E)$  itself is an event. The (implicit) assumption in any (semantic) belief model is that  $E = F$  implies  $B_i(E) = B_i(F)$ . Thus, if two events are extensionally the same, then each player's belief in the two events are the same.<sup>8</sup> To assess player  $i$ 's belief about player  $j$ 's belief about  $E$ , how can the outside

<sup>7</sup>In contrast, it is not necessarily the case that each player is certain of a signal at a state  $\omega$  iff the players are commonly certain of the signal at  $\omega$ .

<sup>8</sup>Although such identification of events are implicitly assumed for any (semantic) belief model, one can construct a canonical ("universal") semantic model from a syntactic language which maximally distinguishes the denotations of events. In the canonical model, such identification of events can be minimized in a way such that two events are equated only when they are

analysts justify the fact that player  $i$  is able to equate  $B_j(E)$  with another event (say,  $F$ )? Since either  $B_j(E) = F$  or  $B_j(E) \neq F$ , player  $i$  is *certain that  $B_j(E)$  is an event  $F$*  if player  $i$  is certain of the indicator function  $\mathbb{I}_{B_j \leftrightarrow F}$ , where  $(B_j(E) \leftrightarrow F) := ((\neg B_j)(E) \cup F) \cap ((\neg F) \cup B_j(E))$ . If player  $i$ 's belief operator  $B_i$  satisfies Necessitation and if  $B_j(E) = F$ , then player  $i$  is certain of the constant indicator function  $\mathbb{I}_{B_j \leftrightarrow F}$ . Thus, under Necessitation, player  $i$  is certain that  $B_j(E) = F$  if it is indeed the case. This argument justifies that, under Necessitation, the outside analysts can say that the players are certain of equating two extensionally equivalent events (say,  $B_j(E)$  and  $F$ ) if they are indeed extensionally equivalent.

Fourth, it can be formally shown that player  $i$  is certain of a profile of signals (e.g., a strategy profile) iff she is certain of each of them. Thus, the players are commonly certain of a profile of signals iff every player is certain of every signal.

### 3.2 A Qualitative-Type Mapping that Represents a Player's Beliefs

In order to formulate a test under which the outside analysts can examine whether the players are commonly certain of a belief model, I define the “belief-generating map,” which I call the qualitative-type mapping ([17]), of a player. Given the belief operator of the player, the qualitative-type mapping associates, with each state, a binary value indicating whether the player believes each event in an analogous manner to the type mapping in the type-space literature.

To that end, recall that a (probabilistic-)type mapping associates, with each state  $\omega$ , the player's probabilistic beliefs  $\tau_i(\omega) \in \Delta(\Omega)$  at that state. With this in mind, let  $M(\Omega)$  be the set of binary set functions  $\mu : \mathcal{D} \rightarrow \{0, 1\}$  (i.e.,  $M(\Omega) \subseteq \{0, 1\}^{\mathcal{D}}$ ) that satisfy a given set of logical properties of beliefs defined in Section 2.2 (these properties will be shortly expressed in terms of  $\mu$ ). Call each  $\mu \in M(\Omega)$  a *qualitative-type*. Interpret  $\mu(E) = 1$  as the belief in an event  $E \in \mathcal{D}$ . Once  $M(\Omega) \subseteq \{0, 1\}^{\mathcal{D}}$  is defined as the set of qualitative-types that satisfy the given set of logical properties of beliefs, I represent player  $i$ 's beliefs by a *qualitative-type mapping*  $t_i : \Omega \rightarrow M(\Omega)$  satisfying a certain measurability condition specified below. It is a measurable mapping which associates, with each state  $\omega \in \Omega$ , player  $i$ 's qualitative-type  $t_i(\omega) \in M(\Omega)$  at  $\omega$ . Thus, player  $i$  believes an event  $E$  at  $\omega$  if  $t_i(\omega)(E) = 1$ .

Now, I define the logical properties of  $\mu$  in an analogous way to the corresponding logical properties of belief operators. Fix  $\mu \in \{0, 1\}^{\mathcal{D}}$ .

0. *Monotonicity*:  $E \subseteq F$  implies  $\mu(E) \leq \mu(F)$ .
1. *Necessitation*:  $\mu(\Omega) = 1$ .
2. *Countable Conjunction*:  $\min_{n \in \mathbb{N}} \mu(E_n) \leq \mu(\bigcap_{n \in \mathbb{N}} E_n)$ .
3. *Finite Conjunction*:  $\min(\mu(E), \mu(F)) \leq \mu(E \cap F)$ .
4. *The Kripke property*:  $\mu(E) = 1$  iff  $\bigcap \{F \in \mathcal{D} \mid \mu(F) = 1\} \subseteq E$ .
5. *Consistency*:  $\mu(E) \leq 1 - \mu(E^c)$ .

The interpretations of the above properties are similar to those in Section 2.2. Whether all of these properties are assumed or not depend on the model that the outside analysts study. For example, if the outside analysts examine a partitioned possibility correspondence model, then  $M(\Omega)$  is the set of qualitative-types that satisfy all the logical properties.

I formally define the measurability condition of a qualitative-type mapping. A *qualitative-type mapping* is a measurable mapping  $t_i : (\Omega, \mathcal{D}) \rightarrow (M(\Omega), \mathcal{D}_M)$  which satisfies given (logical and) introspective properties of beliefs, where  $\mathcal{D}_M$  is the  $\sigma$ -algebra generated by the sets of the form  $\beta_E := \{\mu \in M(\Omega) \mid$

---

explicitly assumed to be equivalent by the outside analysts (see [19] for a formal assertion).

$\mu(E) = 1\}$  for all  $E \in \mathcal{D}$ . The set  $\beta_E$  is the set of types under which  $E$  is believed. Thus,  $\beta_E$  is an informational content indicating that event  $E$  is believed. Note that  $t_i : \Omega \rightarrow M(\Omega)$ , by construction, satisfies given logical properties because any element in  $M(\Omega)$  satisfies them. For example, if every  $\mu \in M(\Omega)$  satisfies the Kripke property, then every  $t_i(\omega)$  satisfies it. Denote  $b_i(\omega) := \bigcap \{E \in \mathcal{D} \mid t_i(\omega)(E) = 1\}$  for each  $\omega \in \Omega$ .

The measurability condition of  $t_i$  requires each  $t_i^{-1}(\beta_E) = \{\omega \in \Omega \mid t_i(\omega)(E) = 1\}$  to be the event that player  $i$  believes  $E$ . Next, I define Truth Axiom and the introspective properties of  $t_i$ .

6. *Truth Axiom:*  $t_i(\omega)(E) = 1$  implies  $\omega \in E$ .
7. *Positive Introspection:*  $t_i(\omega)(E) = 1$  implies  $t_i(\omega)(t_i^{-1}(\beta_E)) = 1$ .
8. *Negative Introspection:*  $t_i(\omega)(E) = 0$  implies  $t_i(\omega)(\neg t_i^{-1}(\beta_E)) = 1$ .

### 3.3 Certainty of Own Type Mapping

I apply the certainty of a signal to a qualitative-type mapping. Proposition 1 below roughly states that a player is certain of her own qualitative-type mapping iff her beliefs are introspective.

**Proposition 1.** *Let  $\vec{\Omega}$  be a belief model, and let  $t_{B_i} : \Omega \rightarrow M(\Omega)$  be player  $i$ 's qualitative-type mapping.*

1. (a) *Player  $i$  is certain of  $t_{B_i}$  with respect to  $\{\beta_E \mid E \in \mathcal{D}\}$  iff  $B_i$  satisfies Positive Introspection.*  
 (b) *Player  $i$  is certain of  $t_{B_i}$  with respect to  $\{\neg\beta_E \mid E \in \mathcal{D}\}$  iff  $B_i$  satisfies Negative Introspection.*  
 (c) *If player  $i$  is certain of  $t_{B_i} : (\Omega, \mathcal{D}) \rightarrow (M(\Omega), \mathcal{D}_M)$ , then  $B_i$  satisfies Positive Introspection and Negative Introspection.*
2. (a) *Let  $B_i$  satisfy Truth Axiom. Player  $i$  is certain of  $t_{B_i} : (\Omega, \mathcal{D}) \rightarrow (M(\Omega), \mathcal{D}_M)$  iff  $B_i$  satisfies (Positive Introspection and) Negative Introspection.*  
 (b) *Let  $B_i$  satisfy Consistency and Countable Conjunction. Player  $i$  is certain of  $t_{B_i} : (\Omega, \mathcal{D}) \rightarrow (M(\Omega), \mathcal{D}_M)$  iff  $B_i$  satisfies Positive Introspection and Negative Introspection.*

Part (1) characterizes the certainty of the qualitative-type mapping  $t_{B_i}$  with respect to the possession or lack of beliefs. In contrast, Parts (2a) and (2b), respectively, examine the sense in which player  $i$  is certain of her qualitative-type mapping  $t_{B_i} : (\Omega, \mathcal{D}) \rightarrow (M(\Omega), \mathcal{D}_M)$  in a model of knowledge and belief.

Part (1a) states that player  $i$  is certain of her qualitative-type mapping  $t_{B_i}$  with respect to the possession of beliefs iff her belief operator  $B_i$  satisfies Positive Introspection. Parts (1a) and (1b) jointly state that  $B_i$  satisfies Positive Introspection and Negative Introspection iff player  $i$  is certain of her qualitative-type mapping  $t_{B_i}$  with respect to  $\{\beta_E \mid E \in \mathcal{D}\} \cup \{\neg\beta_E \mid E \in \mathcal{D}\}$ .

I discuss two implications of Proposition 1. First, it sheds light on the literature of non-partitional knowledge models without Negative Introspection (see footnote 6). Part (1) implies that, without imposing Negative Introspection, player  $i$  is not certain of her own qualitative-type mapping with respect to  $\mathcal{D}_M$  (or  $\{\beta_E \mid E \in \mathcal{D}\} \cup \{\neg\beta_E \mid E \in \mathcal{D}\}$ ). Rather, she takes her own information at face value in the sense that she is only certain of her qualitative-type mapping with respect to her own beliefs  $\{\beta_E \mid E \in \mathcal{D}\}$ . Proposition 1 formalizes the sense in which “she takes her own information at face value.”

In contrast, Proposition 1 (2a) shows that, in a partitional possibility correspondence model of knowledge, a player is fully certain of her possibility correspondence when Truth Axiom, (Positive Introspection) and Negative Introspection hold. While the proposition does not necessarily require  $B_i$  to satisfy the Kripke property, consider a model of knowledge in which  $B_i$  satisfies Truth Axiom and the Kripke property, i.e.,  $B_i$  is induced by the reflexive possibility correspondence  $b_{B_i}$ . Then, player  $i$  is certain of her “knowledge-generating” mapping iff  $B_i$  satisfies (Positive Introspection and) Negative Introspection.

Proposition 1 (2b) shows that, for a serial possibility correspondence, a player is fully certain of her possibility correspondence when her beliefs satisfy Positive Introspection and Negative Introspection.

Second, suppose a player has qualitative belief and knowledge. Consider a model  $\langle (\Omega, \mathcal{D}), (K_i)_{i \in I}, C \rangle$  where  $K_i : \mathcal{D} \rightarrow \mathcal{D}$  is player  $i$ 's (monotone) knowledge operator. Now, for each player  $i$ , let  $B_i : \mathcal{D} \rightarrow \mathcal{D}$  be her (monotone) qualitative-belief operator. Let  $t_{B_i}$  be player  $i$ 's qualitative-type mapping that represents  $B_i$ , and ask whether player  $i$  is certain of her qualitative-type mapping  $t_{B_i} : (\Omega, \mathcal{D}) \rightarrow (M(\Omega), \mathcal{D}_M)$ . Proposition 1 (2a) implies that player  $i$  is certain of  $t_{B_i} : (\Omega, \mathcal{D}) \rightarrow (M(\Omega), \mathcal{D}_M)$  iff  $K_i$  satisfies *Positive Certainty* (with respect to  $B_i$ ):  $B_i(\cdot) \subseteq K_i B_i(\cdot)$  and *Negative Certainty* (with respect to  $B_i$ ):  $(\neg B_i)(\cdot) \subseteq K_i(\neg B_i)(\cdot)$ . Whenever player  $i$  believes an event, she knows that she believes it. Whenever player  $i$  does not believe an event, she knows that she does not believe it. In fact, these two properties are often assumed in a model of belief and knowledge. Proposition 1 (2a) justifies the assumptions in terms of the certainty of one's knowledge about her own beliefs.

## 4 When are the Players Commonly Certain of a Belief Model?

I formalize the sense in which the players are commonly certain of a belief model itself: the players are commonly certain of the profile of their qualitative-type mappings. As discussed, it is sufficient to ask when every player  $i$  is certain of each player  $j$ 's qualitative-type mapping.

To that end, observe that Proposition 1 applies to the case in which player  $i$  is certain of player  $j$ 's qualitative-type mapping. For example, if player  $i$  is certain of player  $j$ 's qualitative-type mapping  $t_j : (\Omega, \mathcal{D}) \rightarrow (M(\Omega), \mathcal{D}_M)$ , then  $B_{t_j}(\cdot) \subseteq B_i B_{t_j}(\cdot)$  and  $(\neg B_{t_j})(\cdot) \subseteq B_i(\neg B_{t_j})(\cdot)$  hold. Proposition 1 implies:

**Remark 1.** Let  $\vec{\Omega}$  be a belief model, and let  $t_{B_j} : \Omega \rightarrow M(\Omega)$  be player  $j$ 's qualitative-type mapping.

1. (a) Player  $i$  is certain of  $t_{B_j}$  with respect to  $\{\beta_E \mid E \in \mathcal{D}\}$  iff  $B_j(\cdot) \subseteq B_i B_j(\cdot)$ .  
 (b) Player  $i$  is certain of  $t_{B_j}$  with respect to  $\{\neg\beta_E \mid E \in \mathcal{D}\}$  iff  $(\neg B_j)(\cdot) \subseteq B_i(\neg B_j)(\cdot)$ .  
 (c) If player  $i$  is certain of  $t_{B_j} : (\Omega, \mathcal{D}) \rightarrow (M(\Omega), \mathcal{D}_M)$ , then  $B_j(\cdot) \subseteq B_i B_j(\cdot)$  and  $(\neg B_j)(\cdot) \subseteq B_i(\neg B_j)(\cdot)$ .
2. (a) Let  $B_i$  satisfy Truth Axiom. Player  $i$  is certain of  $t_{B_j} : (\Omega, \mathcal{D}) \rightarrow (M(\Omega), \mathcal{D}_M)$  iff  $(B_j(\cdot) \subseteq B_i B_j(\cdot))$  and  $(\neg B_j)(\cdot) \subseteq B_i(\neg B_j)(\cdot)$ .  
 (b) Let  $B_i$  satisfy Consistency and Countable Conjunction. Player  $i$  is certain of  $t_{B_j} : (\Omega, \mathcal{D}) \rightarrow (M(\Omega), \mathcal{D}_M)$  iff  $B_j(\cdot) \subseteq B_i B_j(\cdot)$  and  $(\neg B_j)(\cdot) \subseteq B_i(\neg B_j)(\cdot)$ .

Roughly, Remark 1 states that player  $i$  is certain of player  $j$ 's qualitative-type mapping  $t_{B_j}$  if and only if (i) whenever player  $j$  believes an event  $E$  at  $\omega$ , player  $i$  believes player  $j$  believes  $E$  at  $\omega$ ; and (ii) whenever player  $j$  does not believe an event  $E$  at  $\omega$ , player  $i$  believes player  $j$  does not believe  $E$  at  $\omega$ .

Now, I ask when the players are commonly certain of the qualitative-type mappings in a belief model.

**Theorem 1.** Let  $\vec{\Omega}$  be a belief model, and let  $t_{B_i} : \Omega \rightarrow M(\Omega)$  be player  $i$ 's qualitative-type mapping.

1. Assume Truth Axiom for every  $B_i$ . The players are commonly certain of the profile of qualitative-type mappings  $t_{B_i} : (\Omega, \mathcal{D}) \rightarrow (M(\Omega), \mathcal{D}_M)$  iff  $B_i = B_j$  for every  $i, j \in I$ , (Positive Introspection), and Negative Introspection. In particular,  $B_i = C$  for each  $i \in I$ .
2. Assume Consistency and Countable Conjunction for every  $B_i$ . The players are commonly certain of the profile of qualitative-type mappings  $t_{B_i} : (\Omega, \mathcal{D}) \rightarrow (M(\Omega), \mathcal{D}_M)$  iff  $B_i(\cdot) \subseteq C B_i(\cdot)$  and  $(\neg B_i)(\cdot) \subseteq C(\neg B_i)(\cdot)$  for every  $i \in I$ . In particular,  $C = B_I$ .

While Part (1) studies a knowledge model, Part (2) does a belief model. I start with discussing implications of Part (2). This part states that the players are commonly certain of their qualitative-type mappings iff (i) for any event  $E$  which some player  $i$  believes at some state  $\omega$ , it is commonly believed that player  $i$  believes  $E$  at  $\omega$ ; and (ii) for any event  $E$  which some player  $i$  does not believe at some state  $\omega$ , it is commonly believed that player  $i$  does not believe  $E$  at  $\omega$ .

This part imposes a strong requirement that, under Consistency and Countable Conjunction, the mutual belief and common belief operators coincide if the players are commonly certain of their qualitative-type mappings.<sup>9</sup> For any event  $E$  which everybody believes at some state  $\omega$ , it is commonly believed that everybody believes  $E$  at  $\omega$ :  $B_I(\cdot) \subseteq CB_I(\cdot)$ . Intuitively, in a model of which the players are commonly certain, if everybody believes an event  $E$  then it is common belief that everybody believes  $E$ . Thus, if everybody believes  $E$  then everybody believes that everybody believes  $E$ . Hence, the first-order mutual belief itself implies any higher-order mutual beliefs, and thus the mutual and common beliefs coincide.

Next, Part (1) provides a contrast between knowledge and belief. In a knowledge model with Truth Axiom, for the players to be commonly certain of the model, it is *necessary* that their knowledge coincides with each other. In contrast, in a belief model without Truth Axiom, there exists a model in which the players' beliefs are different but they are commonly certain of their qualitative-type mappings.

Yet, Theorem 1 is an impossibility result in the following sense. In Part (1), every player's knowledge operator coincides. In Part (2), the mutual and common belief operators coincide. In this regard, informally, Theorem 1 has some similarity with the impossibility of agreeing-to-disagree [1]: (under a common prior) if two players have common knowledge of their posteriors then the posteriors coincide. Here, if players are commonly certain of their knowledge operators, then their knowledge operators coincide.

Finally, as an implication of Theorem 1, suppose that the players are commonly certain of a belief model. If player  $i$  is certain of a signal  $x$ , then is player  $j$  certain of the signal  $x$ , too? While the players' beliefs may not be homogeneous, the proposition below shows that this is the case.

**Proposition 4.** *Let  $\vec{\Omega}$  be a belief model such that each  $B_i$  satisfies Consistency. Let  $x: (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  be a signal such that, for any  $F \in \mathcal{X}$ , there exists a sub-collection  $(F_\lambda)_{\lambda \in \Lambda}$  of  $\mathcal{X}$  with  $F^c = \bigcup_{\lambda \in \Lambda} F_\lambda$ .*

1. *If player  $i$  is certain of  $x: (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  and if player  $j$  is certain of player  $i$ 's qualitative-type mapping  $t_{B_i}: (\Omega, \mathcal{D}) \rightarrow (M(\Omega), \mathcal{D}_M)$ , then player  $j$  is certain of  $x: (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$ .*
2. *Suppose that the players are commonly certain of the profile of their qualitative-type mappings  $t_{B_i}: (\Omega, \mathcal{D}) \rightarrow (M(\Omega), \mathcal{D}_M)$ . Then, player  $i$  is certain of  $x: (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$  iff player  $j$  is certain of  $x: (\Omega, \mathcal{D}) \rightarrow (X, \mathcal{X})$ .*

The meta-common-certainty assumption states that if player  $i$  is certain of her own strategy and if player  $j$  is certain of player  $i$ 's type mapping then player  $j$  is certain of player  $i$ 's strategy. In particular, if the players are commonly certain of the profile of their type mappings and if each player is certain of her own strategy, then it follows that the players are commonly certain of the strategy profile. The next section examines the role of such meta-certainty assumptions on game-theoretic solution concepts.

---

<sup>9</sup>The converse does not hold, i.e.,  $C = B_I$  does not necessarily imply that the players are certain of the profile of their qualitative-type mappings.

## 5 What Role Does the “Meta-Certainty” of a Model Play in Game-theoretic Analyses?

This section studies the role that the “meta-certainty” assumption plays in game-theoretic analyses of solution concepts. Specifically, it considers the solution concept of iterated elimination of strictly dominated actions (IESDA) in a strategic game. Informally, an epistemic characterization of IESDA states that, in a strategic game, if the (i) “logical” players are (ii) “commonly (meta-)certain of the game” and if they (iii) commonly believe their rationality, then their resulting actions survive IESDA. Formally, in the context of the framework of this paper, [18] shows that if the players commonly believe each player’s rationality and if each of them correctly believes their own rationality, then their resulting actions survive IESDA, without assuming any property on individual players’ beliefs. This paper connects these two statements as follows: first, suppose that the players are logical in that their beliefs satisfy Consistency and Finite Conjunction in addition to Monotonicity. Second, suppose that each of them is certain of their own qualitative-type mapping and strategy. Third, suppose that the players commonly believe their rationality. Then, their resulting actions survive IESDA.

Here I show that the certainty (of her own strategy and type mapping) allows her to correctly believe her own rationality. In other words, if a player is able to reason about informativeness of her own beliefs, she is able to correctly believe her own rationality.

### 5.1 A Strategic Game, a Model of a Game, and Rationality

To define the notion of rationality in a game, define a (*strategic*) *game* as a tuple  $\Gamma = \langle (A_i)_{i \in I}, (\succsim_i)_{i \in I} \rangle$ :  $A_i$  is a non-empty at-most-countable set of player  $i$ ’s actions, and  $\succsim_i$  is  $i$ ’s (complete and transitive) preference relation on  $A := \times_{i \in I} A_i$ .<sup>10</sup> Denote by  $\sim_i$  and  $\succ_i$  the indifference and strict relations, respectively.

A (belief) *model* of the game  $\Gamma$  is a tuple  $\langle (\Omega, \mathcal{D}), (B_i)_{i \in I}, C, (\sigma_i)_{i \in I} \rangle$  (abusing the notation, denote it by  $\vec{\Omega}$ ) with the following two properties. First,  $\langle (\Omega, \mathcal{D}), (B_i)_{i \in I}, C \rangle$  is a belief model. Second,  $\sigma_i : \Omega \rightarrow A_i$  is a *strategy* of player  $i$  satisfying the measurability condition that  $\sigma_i^{-1}(\{a_i\}) \in \mathcal{D}$  for all  $a_i \in A_i$ . Denote  $[\sigma_i(\omega)] := \sigma_i^{-1}(\{\sigma_i(\omega)\})$  for each  $\omega \in \Omega$ .

Denote by  $[a'_i \succsim_i a_i] := \{\omega' \in \Omega \mid (a'_i, \sigma_{-i}(\omega')) \succsim_i (a_i, \sigma_{-i}(\omega'))\} \in \mathcal{D}$  for any  $a_i, a'_i \in A_i$ . In words,  $[a'_i \succsim_i a_i]$  is the event that player  $i$  prefers taking action  $a'_i$  to  $a_i$  given the opponents’ strategies  $\sigma_{-i}$ . The set  $[a'_i \succsim_i a_i]$  is an event because  $[a'_i \succsim_i a_i] = \sigma_{-i}^{-1}(\{a_{-i} \in A_{-i} \mid (a'_i, a_{-i}) \succsim_i (a_i, a_{-i})\}) \in \mathcal{D}$ . Define  $[a'_i \succ_i a_i]$  and  $[a'_i \sim_i a_i]$  analogously.

Denote by  $\text{RAT}_i$  the event that player  $i$  is *rational* (see, e.g., [7, 8, 14]):

$$\text{RAT}_i := \{\omega \in \Omega \mid \omega \in B_i([a'_i \succ_i \sigma_i(\omega)]) \text{ for no } a'_i \in A_i\}.$$

It can be seen that  $\text{RAT}_i$  is indeed an event. Let  $\text{RAT}_I := \bigcap_{i \in I} \text{RAT}_i$ . Player  $i$  is *rational* at  $\omega \in \Omega$  if there is no action  $a'_i \in A_i$  such that she believes that playing  $a'_i$  is strictly better than playing  $\sigma_i(\omega)$  given the opponents’ strategies  $\sigma_{-i}$ . In other words, player  $i$  is rational at  $\omega$  if, for any action  $a'_i$ , she always considers it possible that playing  $\sigma_i(\omega)$  is at least as good as playing  $a'_i$  given the opponents’ strategies  $\sigma_{-i}$ :  $\omega \in (\neg B_i)(\neg[\sigma_i(\omega) \succsim_i a'_i])$  for any  $a'_i \in A_i$ .

Now, the epistemic characterization of IESDA is stated as follows. Suppose that each player  $i$  correctly believes her own rationality:  $B_i(\text{RAT}_i) \subseteq \text{RAT}_i$ . If every player’s rationality is common belief at

<sup>10</sup>The assumption on the cardinality of each action set  $A_i$  is to simplify the analysis. It guarantees that each player is able to reason about any subset of action profiles and that the rationality of each player is an event.

$\omega$ , i.e.,  $\omega \in \bigcap_{i \in I} C(\text{RAT}_i)$ , then the resulting actions  $(\sigma_i(\omega))_{i \in I} \in A$  survive any process of IESDA.<sup>11</sup>

Finally, player  $i$  is *certain of her own strategy*  $\sigma_i$  if she is certain of  $\sigma_i : (\Omega, \mathcal{D}) \rightarrow (A_i, \{\{a_i\} \mid a_i \in A_i\})$ , equivalently,  $[\sigma_i(\cdot)] \subseteq B_i([\sigma_i(\cdot)])$ . Note that, under Consistency in addition to Monotonicity, if player  $i$  is certain of her own strategy then  $B_i([\sigma_i(\cdot)]) = [\sigma_i(\cdot)]$ ,  $[\sigma_i(\cdot)]^c = B_i([\sigma_i(\cdot)]^c)$ , and  $B_i(\Omega) = \Omega$ .<sup>12</sup>

## 5.2 The Role of Meta-certainty in Correctly Believing One's Own Rationality

I ask under what conditions player  $i$  correctly believes her own rationality:  $B_i(\text{RAT}_i) \subseteq \text{RAT}_i$ . For qualitative belief, the standard assumptions on qualitative belief (i.e., Consistency, Positive Introspection, Negative Introspection, and the Kripke property) guarantee that  $B_i(\text{RAT}_i) = \text{RAT}_i$  (e.g., [7, 8]).<sup>13</sup> Here, I provide a compatibility condition on belief with informativeness, under which a player correctly believes her own rationality. The compatibility condition does not hinge on a particular form of belief, i.e., whether it is qualitative or probabilistic.

To that end, a state  $\omega$  is *at least as informative as* another state  $\omega'$  to  $i$  (precisely, according to  $t_{B_i}$ ) iff  $t_{B_i}(\omega')(\cdot) \leq t_{B_i}(\omega)(\cdot)$ . Fix  $\omega \in \Omega$ , and let  $(\uparrow t_{B_i}(\omega)) := \{\omega' \in \Omega \mid t_{B_i}(\omega)(\cdot) \leq t_{B_i}(\omega')(\cdot)\}$  be the set of states that are at least as informative to  $i$  as  $\omega$ . Under the Kripke property,  $\omega' \in (\uparrow t_{B_i}(\omega))$  iff  $b_{B_i}(\omega') \subseteq b_{B_i}(\omega)$ . The full paper extensively studies the notion of informativeness. Now:

**Definition 2.** Player  $i$ 's belief (operator  $B_i$ ) is *compatible with informativeness* if  $(\uparrow t_{B_i}(\omega)) \cap E \neq \emptyset$  for any  $E \in \mathcal{D}$  with  $\omega \in B_i(E)$ .

In words, player  $i$ 's beliefs are compatible with informativeness if, for any event  $E$  which player  $i$  believes at some  $\omega$ , there exists a state  $\omega'$  in  $E$  which is at least as informative as  $\omega$ . In the context of qualitative beliefs, if player  $i$ 's belief operator  $B_i$  satisfies the Kripke property, Consistency, and Positive Introspection, then  $B_i$  is compatible with informativeness. The compatibility with informativeness does not necessarily imply the Kripke property (and vice versa). If player  $i$ 's belief operator  $B_i$  is compatible with informativeness, then it satisfies  $B_i(\emptyset) = \emptyset$ . Thus, under Finite Conjunction, if  $B_i$  is compatible with informativeness, then it satisfies Consistency.

The following proposition states that the compatibility of beliefs with informativeness is implied by the certainty of a type mapping.

**Proposition 5.** Let  $\vec{\Omega}$  be a belief model. Assume: (i)  $(\uparrow t_{B_i}(\cdot)) \in \mathcal{D}$ ; (ii)  $B_i$  satisfies Consistency and Finite Conjunction; and that (iii) player  $i$  is certain of  $t_{B_i} : \Omega \rightarrow M(\Omega)$  with respect to  $\{\{\mu \in M(\Omega) \mid \mu(\cdot) \geq t_{B_i}(\omega)(\cdot)\} \mid \omega \in \Omega\}$ . Then,  $B_i$  is compatible with informativeness.

The proposition states that, under the regularity condition (i), if player  $i$  is logical (in that her belief operator satisfies Consistency and Finite Conjunction) and if she is certain of her qualitative-type mapping, then her beliefs are compatible with informativeness. Theorem 2 below establishes that if player  $i$ 's beliefs are compatible with informativeness then she correctly believes her rationality, which is a part of the preconditions of the epistemic characterization of IESDA.

<sup>11</sup>Since each player's belief operator  $B_i$  satisfies Monotonicity,  $B_i(\text{RAT}_I) \subseteq \bigcap_{i \in I} B_i(\text{RAT}_i)$  and  $C(\text{RAT}_I) \subseteq \bigcap_{i \in I} C(\text{RAT}_i)$ . Thus, if every player  $i$  correctly believes the rationality of the players, then each player correctly believes her own rationality. Likewise, if it is common belief that the players are rational, then, for every  $i \in I$ , it is common belief that player  $i$  is rational. Hence, I examine the weaker condition that each player  $i$  correctly believes her own rationality.

<sup>12</sup>Thus, under Consistency and Monotonicity of  $B_i$ , the certainty of own strategy implies that if player  $i$  is rational at  $\omega$ , then she never takes a strictly dominated action at  $\omega$  (if she takes a strictly dominated action, then her belief violates Necessitation).

<sup>13</sup>It can be seen that Consistency, Positive Introspection, and the Kripke property in addition to the certainty of  $i$ 's own strategy yield  $B_i(\text{RAT}_i) \subseteq \text{RAT}_i$ . Likewise, Negative Introspection and the Kripke property in addition to the certainty of  $i$ 's own strategy yield  $\text{RAT}_i \subseteq B_i(\text{RAT}_i)$ .

Now, the main result of this section is as follows: a player correctly believes her own rationality if: (i) she is certain of her own strategy; (ii) her belief is compatible with the informativeness; and if (iii) her belief is (finitely) conjunctive so that she can simultaneously reason about her own strategy and her own rationality.

**Theorem 2.** *Suppose that player  $i$  is certain of her own strategy (i.e.,  $[\sigma_i(\cdot)] \subseteq B_i([\sigma_i(\cdot)])$ ). Also, let  $B_i$  be compatible with informativeness and satisfy Finite Conjunction. Then, player  $i$  correctly believes her own rationality:  $B_i(\text{RAT}_i) \subseteq \text{RAT}_i$ .*

Proposition 5 and Theorem 2 imply that player  $i$  correctly believes her own rationality if she is logical in that her belief operator satisfies Consistency and Finite Conjunction and if she is certain of her own type mapping and strategy. Theorem 2 states that, for the role of the meta-certainty assumption of a belief model on IESDA, it is not necessary that each player is certain of the profile of type mappings but it is sufficient that each player is certain of her own type mapping. One can incorporate the assumptions that each player is certain of her own qualitative type-mapping and strategy into the condition that she is certain of the part of the model of a game  $\langle (\Omega, \mathcal{D}), (t_{B_i}, \sigma_i) \rangle$  that dictates her beliefs and strategy.

## References

- [1] Robert J. Aumann (1976): *Agreeing to Disagree*. *The Annals of Statistics* 4(6), pp. 1236–1239, doi:10.1214/aos/1176343654.
- [2] Robert J. Aumann (1987): *Correlated Equilibrium as an Expression of Bayesian Rationality*. *Econometrica* 55(1), pp. 1–18, doi:10.2307/1911154.
- [3] Robert J. Aumann (1999): *Interactive Epistemology I, II*. *International Journal of Game Theory* 28(3), pp. 263–300, 301–314. doi:10.1007/s001820050111, doi:10.1007/s001820050112.
- [4] Michael Bacharach (1985): *Some Extensions of a Claim of Aumann in an Axiomatic Model of Knowledge*. *Journal of Economic Theory* 37(1), pp. 167–190, doi:10.1016/0022-0531(85)90035-3.
- [5] Michael Bacharach (1990): *When Do We Have Information Partitions?* In M. Bacharach & M. Dempster, editors: *Mathematical Models in Economics*, Oxford University Press, pp. 1–23.
- [6] K. Binmore & A. Brandenburger (1990): *Common Knowledge and Game Theory*. In K. Binmore, editor: *Essays on the Foundations of Game Theory*, Basil Blackwell, pp. 105–150.
- [7] Giacomo Bonanno (2008): *A Syntactic Approach to Rationality in Games with Ordinal Payoffs*. In Giacomo Bonanno, Wiebe van der Hoek & Michael Wooldridge, editors: *Logic and the Foundations of Game and Decision Theory (LOFT 7)*, Amsterdam University Press, pp. 59–86, doi:10.5117/9789089640260.
- [8] Giacomo Bonanno (2015): *Epistemic Foundations of Game Theory*. In Hans van Ditmarsch, Joseph Y. Halpern, Wiebe van der Hoek & Barteld Pieter Kooi, editors: *Handbook of Epistemic Logic*, College Publications, pp. 443–487.
- [9] Adam Brandenburger & Eddie Dekel (1987): *Rationalizability and Correlated Equilibria*. *Econometrica* 55(6), pp. 1391–1402, doi:10.2307/1913562.
- [10] Adam Brandenburger & Eddie Dekel (1989): *The Role of Common Knowledge Assumptions in Game Theory*. In Frank Hahn, editor: *The Economics of Missing Markets, Information, and Games*, Oxford University Press, pp. 46–61.
- [11] Adam Brandenburger & Eddie Dekel (1993): *Hierarchies of Beliefs and Common Knowledge*. *Journal of Economic Theory* 59(1), pp. 189–198, doi:10.1006/jeth.1993.1012.
- [12] Adam Brandenburger, Eddie Dekel & John Geanakoplos (1992): *Correlated Equilibrium with Generalized Information Structures*. *Games and Economic Behavior* 4(2), pp. 182–201, doi:10.1016/0899-8256(92)90014-J.

- [13] Adam Brandenburger & H. Jerome Keisler (2006): *An Impossibility Theorem on Beliefs in Games*. *Studia Logica* 84(2), pp. 211–240, doi:10.1007/s11225-006-9011-z.
- [14] Yi-Chun Chen, Ngo Van Long & Xiao Luo (2007): *Iterated Strict Dominance in General Games*. *Games and Economic Behavior* 61(2), pp. 299–315, doi:10.1016/j.geb.2007.02.002.
- [15] Eddie Dekel & Faruk Gul (1997): *Rationality and Knowledge in Game Theory*. In David M. Kreps & Kenneth F. Wallis, editors: *Advances in Economics and Econometrics: Theory and Applications, Seventh World Congress*, 1, Cambridge University Press, pp. 87–172, doi:10.1017/CCOL521580110.005.
- [16] Ronald Fagin, John Geanakoplos, Joseph Y. Halpern & Moshe Y. Vardi (1999): *The Hierarchical Approach to Modeling Knowledge and Common Knowledge*. *International Journal of Game Theory* 28(3), pp. 331–365, doi:10.1007/s001820050114.
- [17] Satoshi Fukuda (2017): *The Existence of Universal Knowledge Spaces*. In: *Essays in the Economics of Information and Epistemology*, Ph.D. Dissertation, the University of California at Berkeley, pp. 1–113.
- [18] Satoshi Fukuda (2020): *Formalizing Common Belief with No Underlying Assumption on Individual Beliefs*. *Games and Economic Behavior* 121, pp. 169–189, doi:10.1016/j.geb.2020.02.007.
- [19] Satoshi Fukuda (2021): *The Existence of Universal Qualitative Belief Spaces*.
- [20] John Geanakoplos (1989): *Game Theory without Partitions, and Applications to Speculation and Consensus*. Cowles Foundation Discussion Paper No. 914, Yale University.
- [21] Itzhak Gilboa (1988): *Information and Meta-Information*. In: *Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge*, Morgan Kaufmann Publishers Inc., pp. 227–243.
- [22] John C. Harsanyi (1967-1968): *Games with Incomplete Information Played by “Bayesian” Players, I-III*. *Management Science* 14, pp. 159–182, 320–334, 486–502. doi:10.1287/mnsc.14.3.159, doi:10.1287/mnsc.14.5.320, doi:10.1287/mnsc.14.7.486.
- [23] Dov Monderer & Dov Samet (1989): *Approximating Common Knowledge with Common Beliefs*. *Games and Economic Behavior* 1(2), pp. 170–190, doi:10.1016/0899-8256(89)90017-1.
- [24] Stephen Morris (1996): *The Logic of Belief and Belief Change: A Decision Theoretic Approach*. *Journal of Economic Theory* 69(1), pp. 1–23, doi:10.1006/jeth.1996.0035.
- [25] Cesaltina Pacheco Pires (1994): *Do I know  $\Omega$ ? An Axiomatic Model of Awareness and Knowledge*.
- [26] Olivier Roy & Eric Pacuit (2013): *Substantive assumptions in interaction: a logical perspective*. *Synthese* 190(5), pp. 891–908, doi:10.1007/s11229-012-0191-y.
- [27] Dov Samet (1990): *Ignoring Ignorance and Agreeing to Disagree*. *Journal of Economic Theory* 52(1), pp. 190–207, doi:10.1016/0022-0531(90)90074-T.
- [28] Hyun Song Shin (1993): *Logical Structure of Common Knowledge*. *Journal of Economic Theory* 60(1), pp. 1–13, doi:10.1006/jeth.1993.1032.
- [29] Robert Stalnaker (1994): *On the Evaluation of Solution Concepts*. *Theory and Decision* 37(1), pp. 49–73, doi:10.1007/BF01079205.
- [30] Tommy Chin-Chiu Tan & Sérgio Ribeiro da Costa Werlang (1988): *The Bayesian Foundations of Solution Concepts of Games*. *Journal of Economic Theory* 45(2), pp. 370–391, doi:10.1016/0022-0531(88)90276-1.
- [31] Tommy Chin-Chiu Tan & Sérgio Ribeiro da Costa Werlang (1992): *On Aumann’s Notion of Common Knowledge: an Alternative Approach*. *Revista Brasileira de Economia - RBE* 46(2), pp. 151–166.
- [32] Spyros Vassilakis & Shmuel Zamir (1993): *Common Belief and Common Knowledge*. *Journal of Mathematical Economics* 22(5), pp. 495–505, doi:10.1016/0304-4068(93)90039-N.
- [33] Sérgio Ribeiro da Costa Werlang (1987): *Common Knowledge*. In John Eatwell, Murray Milgate & Peter Newman K., editors: *The New Palgrave: A Dictionary of Economics*, 2nd edition, Macmillan, pp. 74–85, doi:10.1007/978-1-349-20181-5\_5.

- [34] Robert Wilson (1987): *Game-Theoretic Analyses of Trading Processes*. In Truman Fassett Bewley, editor: *Advances in Economic Theory: Fifth World Congress*, Cambridge University Press, pp. 33–70, doi:10.1017/CCOL0521340446.002.

# Knowledge from Probability

Jeremy Goodman

School of Philosophy  
University of Southern California, USA  
jeremy.goodman@usc.edu

Bernhard Salow

Faculty of Philosophy  
University of Oxford, UK  
bernhard.salow@philosophy.ox.ac.uk

We give a probabilistic analysis of inductive knowledge and belief and explore its predictions concerning knowledge about the future, about laws of nature, and about the values of inexactly measured quantities. The analysis combines a theory of knowledge and belief formulated in terms of relations of comparative normality with a probabilistic reduction of those relations. It predicts that only highly probable propositions are believed, and that many widely held principles of belief-revision fail.

How can we have knowledge that goes beyond what we have observed – knowledge about the future, or about lawful regularities, or about the distal causes of the readings of our scientific instruments? Many philosophers think we can't. Nelson Goodman, for example, disparagingly writes that “obviously the genuine problem [of induction] cannot be one of attaining unattainable knowledge or of accounting for knowledge that we do not in fact have” [20, p. 62]. Such philosophers typically hold that the best we can do when it comes to inductive hypotheses is to assign them high probabilities. Here we argue that such pessimism is misplaced. We give a purely probabilistic analysis of inductive knowledge and (rational) belief, and motivate it by drawing out its attractive predictions about a range of cases.

Our analysis builds on two recent strands of research. The first strand is the idea that knowledge and (rational) belief can be analyzed in terms of a notion of *normality*: among the possibilities that are compatible with an agent's evidence, their knowledge rules out those that are sufficiently less normal than their actual circumstances, and their beliefs rule out those that are sufficiently less normal than some other evidential possibilities.<sup>1</sup> The second strand is that what a person knows or believes is always relative to a contextually supplied question.<sup>2</sup> Our guiding observation is this: *there is a natural way to define the comparative normality of evidential possibilities in terms of the probabilities of the answers to a question*. This fact allows us to give an analysis of knowledge and belief in terms of probability and evidence, two notions that even skeptics about inductive knowledge typically accept.

Here is our plan. We begin by presenting a version of the theory of knowledge and belief in terms of comparative normality that we have defended elsewhere [19, 18]. We next explain how comparative normality can be reduced to evidential probability in a question-relative way. We then use this framework to model knowledge and belief about ongoing chancy processes (section 3), lawful regularities (section 4), multiple independent subject matters (section 5), and the values of quantities measured using instruments that are subject to random noise (section 6).<sup>3</sup> As these case studies will illustrate, the framework is conservative in its synchronic predictions but revisionist in its diachronic ones: what an agent knows and believes is closed under entailment and always has a high probability of being true, but getting new evidence can lead to changes in what one believes that violate widely endorsed principles about belief-revision. Three appendices explain ways in which the framework can be extended in order to model a wider range of cases of inductive knowledge.

<sup>1</sup> See [44, 45, 46, 17, 21, 11, 19, 18, 6, 33, 4, 7, 34, 16] for related ideas about knowledge, [39, 40, 41, 42, 43] for related ideas about justified belief, and [27, 35] and references therein for related ideas about non-monotonic reasoning.

<sup>2</sup> See [38, 55, 25] for precedents in the case of knowledge, and [29, 56, 24, 5] for precedents in the case of belief.

<sup>3</sup> See [10, 2, 3, 41, 42, 19, 18] for recent discussion of the kind of cases in §§3-5, and [50, 51, 53, 54, 8, 17, 47, 45, 6, 12, 7, 18] for recent discussion of the kind of cases in §6.

## 1 The Normality Framework

Our models of knowledge and belief will be a version of Hintikka semantics for a single agent.<sup>4</sup> Both knowledge and belief are given by accessibility relations, in the sense that an agent knows/believes  $p$  in a world  $w$  iff  $p$  is true in all worlds epistemically/doxastically accessible from  $w$ . What is distinctive of the framework is how these accessibility relations are defined in terms of other relations between worlds, encoding the agent's evidence and worlds' comparative normality.

Let a *normality structure* be a tuple  $\langle S, \mathcal{E}, W, \succ, \gg \rangle$  such that:

1.  $S$  is a non-empty set (of *states*),
2.  $\mathcal{E} \subseteq \mathcal{P}(S) \setminus \{\emptyset\}$  (the *possible bodies of evidence*)
3.  $W = \{\langle s, E \rangle : s \in E \in \mathcal{E}\}$  (the set of (*centered*) *worlds*),
4.  $\succ$  is a preorder on  $W$  (read ' $w \succ v$ ' as ' $w$  is *at least as normal as*  $v$ '),
5.  $\gg$  is a well-founded relation on  $W$  (read ' $w \gg v$ ' as ' $w$  is *sufficiently more normal than*  $v$ '), such that, for any worlds  $w_1, w_2, w_3, w_4$ :
  - (a) If  $w_1 \gg w_2$ , then  $w_1 \succ w_2$ ;
  - (b) If  $w_1 \succ w_2 \gg w_3 \succ w_4$ , then  $w_1 \gg w_4$ .

The intuitive idea behind modeling worlds as state/set-of-state pairs is that we are only modelling the agent's knowledge and beliefs about a certain subject matter – the state of the world – and for this purpose we may idealize and treat worlds as individuated by the state of the world together with the agent's evidence about the state of the world, modelled as the set of states compatible with their evidence. As we will understand it, a person's evidence is a subset of their knowledge, and hence is true; this is why in any world the actual state is a member of the set of states compatible with the agent's evidence. We use  $R_e$  to denote the function mapping each world to the set of worlds that are *evidentially accessible* from it, in the sense of being compatible with the agent's evidence:

$$R_e(\langle s, E \rangle) := \{\langle s', E \rangle : s' \in E\}$$

Next, we define a function  $R_b$  for *doxastic accessibility*, characterizing the set of worlds compatible with what the agent believes in any given world. The idea is that what the agent believes goes beyond what is entailed by their evidence: the doxastic possibilities are those evidential possibilities that are not sufficiently less normal than any other. Formally,

$$R_b(w) := \{v \in R_e(w) : \neg(\exists u \in R_e(w) : u \gg v)\}$$

Finally, we define a function  $R_k$  for *epistemic accessibility*, characterizing the set of worlds compatible with what the agent knows in any given world. There are two natural definitions here: one for those who follow Stalnaker [44, 45, 46] in thinking that epistemic accessibility is a transitive relation (in which case knowing  $p$  entails knowing that you know  $p$ ), and another for those who follow Williamson [48, 50, 52, 53] in thinking that epistemic accessibility cannot be transitive because knowledge requires a margin for error. The difference concerns whether worlds that are doxastically inaccessible should be epistemically accessible when they are less normal but not sufficiently less normal than actuality: the Stalnakerian answers “no” (the agent knows those worlds don't obtain), while the Williamsonian answers “yes” (for all the agent knows, those worlds obtain). Formally, these answers correspond to the following respective definitions:

<sup>4</sup>We can model the knowledge/beliefs of  $n$  agents by generalizing clause 2 of the definition of a normality structure so  $\mathcal{E} \subseteq \mathcal{P}(S)^n$  (giving the possible patterns of bodies of evidence among the agents) and modify the remaining definitions accordingly.

$$\begin{aligned}
R_k(w) &:= R_b(w) \cup \{v \in R_e(w) : v \succ w\} && \text{(Stalnakerian)} \\
R_k(w) &:= R_b(w) \cup \{v \in R_e(w) : v \succ w\} \cup \overbrace{\{v \in R_e(w) : w \succ v \wedge \neg(w \succ v)\}}^{\text{the margin for error}} && \text{(Williamsonian)}
\end{aligned}$$

The results we will be exploring in this paper don't depend on which definition of knowledge we adopt.<sup>5</sup> Although we will remain neutral on which definition is preferable by focusing mainly on belief, we believe that the normality framework is recommended in large part by its ability to integrate an anti-skeptical theory of knowledge with a non-trivial theory of inductive belief; we make this case at greater length in [18].

The framework also allows us to model the dynamics of knowledge and belief *about the state of the world* in response to new evidence about the state of the world. To make this idea precise, we first introduce the projection functions  $\pi_i$  such that  $\pi_i(\langle x_1, \dots, x_n \rangle) = x_i$ . We then define which *states* are accessible from a world  $w$  as  $\mathcal{R}_*(w) = \{\pi_1(v) : v \in R_*(w)\}$  where  $*$   $\in \{e, b, k\}$ . (So, e.g.,  $\mathcal{R}_e(w) = \pi_2(w)$ .) For any set of states  $p$  and pair of worlds  $w$  and  $v$ , we say that  $v$  is the *result of discovering  $p$  in  $w$*  iff  $\pi_1(v) = \pi_1(w)$  and  $\pi_2(v) = p \cap \pi_2(w)$ . Although  $R_b(w) \cap R_b(v) = \emptyset$  whenever  $v$  is the result of (non-trivially) discovering  $p$  in  $w$  (non-trivially in the sense that  $w \neq v$ ), the dynamics relating  $\mathcal{R}_b(w)$  and  $\mathcal{R}_b(v)$  (and  $\mathcal{R}_k(w)$  and  $\mathcal{R}_k(v)$ ) are more interesting, as we will explore below. Note that, unlike standard models of belief-revision, there may be no  $v$  that is the result of discovering  $p$  in  $w$  – for example, if  $p$  is incompatible with the state of the world in  $w$ . And since only truths can be discovered, normality structures allow us to easily model iterated discoveries (in contrast to formally similar models of theories of belief-revision like AGM [1], first developed in [22], which do not handle iterated belief-revision).<sup>6,7</sup>

## 2 Reducing normality to probability

Appealing to notions of comparative normality (or comparative plausibility) is, by now, a familiar idea in theorizing about knowledge and belief. The main advance of this paper is to explore the consequences of an analysis of these notions in terms of the result of conditioning a prior probability distribution on the agent's evidence.

Let us begin with the at-least-as-normal relation  $\succ$ . An initially attractive idea is that  $w \succ v$  iff  $w$  is at least as probable as  $v$ . Unfortunately, this simple proposal faces a number of problems. For example, the probability of a world depends on how finely we individuate worlds in our model in ways that intuitively shouldn't make a difference to what an agent knows or believes; also, natural ways of individuating worlds often make them all have the same probability, thereby trivializing inductive knowledge and belief. For these reasons, as well as others explained in section 5, we will model knowledge and belief

<sup>5</sup>Note that the Williamsonian definition is equivalent to the much simpler definition  $R_k(w) := \{v \in R_e(w) : \neg(w \succ v)\}$  in normality structures where  $\succ$  is a *total* preorder on  $E$  for all  $E \in \mathcal{E}$ , which includes all normality structures generated from probability structures in the way described below; see [18] (which considers a slightly more complicated Williamsonian definition to ensure that the set of epistemically accessible evidential possibilities is closed under  $\succ$ ).

<sup>6</sup>Normality structures are also related to models in dynamic epistemic logic, since instead of modeling accessibility as a relation between worlds, we could equivalently treat it as family of relations between states indexed by bodies of evidence, much like how accessibility is relativized to propositions in dynamic epistemic logic. Another formal precedent is [9], in which formulas are evaluated relative to a pair of a world and a set of worlds containing it. (Thanks to Aybüke Özgün for drawing our attention to this work.) Note that what “world” the agent is in changes as they get new evidence; those who prefer to reserve the word “world” for something unchanging can substitute “situation”, “case”, or “centered world”.

<sup>7</sup>The framework presented here sides with [18] over [19] by treating normality relations as holding between worlds rather than states. But it sides with [19] by presupposing that worlds can be factored into state/evidence pairs, which in turn implies that evidential accessibility is an equivalence relation. Appendix A explains how to modify the definitions below in order to model scenarios in which evidential accessibility is not an equivalence relation.

as relative to a contextually supplied question about the state of the world. Doing so allows for a more robust characterization of normality in terms of probability, as we will now explain.

Let a *probability structure* be a tuple  $\langle S, \mathcal{E}, W, Q, P, t \rangle$  such that:

1.  $S, \mathcal{E}, W$  satisfy clauses 1-3 of the definition of normality structures,
2.  $Q$  (the *question*) is a partition of  $S$ ,
3.  $P$  (the *prior*) is a probability distribution over  $S$  such that  $P(q|E)$  is defined for all  $q \in Q$  and  $E \in \mathcal{E}$ ,
4.  $t \in [0, 1]$  (the *threshold*).

We will now explain how to generate a normality structure from a probability structure. We identify the normality of a world with the evidential probability at that world of the true answer to  $Q$  at that world. Formally, letting  $[s]_Q$  (the *answer to  $Q$  in  $s$* ) be the cell of  $Q$  containing  $s$ ,  $P_w$  (the *evidential probability at  $w$* ) be  $P(\cdot|\pi_2(w))$ , and  $\lambda(w)$  (the *likeliness of  $w$* ) be  $P_w([\pi_1(w)]_Q)$ , we adopt the following definition of one world being at least as normal as another:

$$\text{NORMALITY AS LIKELINESS: } w \succcurlyeq v := \lambda(w) \geq \lambda(v) \text{ and } v \in R_e(w)$$

Next, let the *typicality* of a world be the evidential probability, at that world, that things are no more normal than they are at that world: formally,  $\tau(w) = P_w(\{\pi_1(v) : w \succcurlyeq v \text{ and } v \in R_e(w)\})$ . We will adopt the following definition of one world being sufficiently more normal than another:

$$\text{SUFFICIENCY: } w \gg v := 1 - \frac{\tau(v)}{\tau(w)} \geq t \text{ and } v \in R_e(w)$$

It is easy to verify that, so defined,  $\langle S, \mathcal{E}, W, \succcurlyeq, \gg \rangle$  is a normality structure.<sup>8</sup> In this normality structure, what the agent believes about the state of the world is the strongest disjunction of answers to  $Q$  that (i) includes the most probable answers, (ii) includes all answers at least as probable as any it includes, and (iii) has total probability at least  $t$ .<sup>9</sup> (This will also be what the agent knows about the state of the world if one of the most probable answers to  $Q$  is true.) NORMALITY AS LIKELINESS ensures (i) and (ii), while SUFFICIENCY ensures (iii), which may be more precisely stated as follows:<sup>10</sup>

$$\text{THRESHOLD: } P_w(\mathcal{R}_b(w)) \geq t \text{ for all } w \in W.$$

<sup>8</sup>The requirement that  $\succcurlyeq$  and  $\gg$  only relate evidentially accessible worlds is needed to validate conditions 5a and 5b of the definition of a normality structure, since worlds with the same likeliness but different evidence can have different typicality. For example, let  $S = \{1, \dots, 7\}$ ,  $\mathcal{E} = \{\{1, 2, 3\}, \{4, 5, 6, 7\}\}$ ,  $Q = \{\{s\} : s \in S\}$ ,  $P(1) = P(2) = P(4) = .2; P(3) = P(5) = P(6) = P(7) = .1; t = .5$ . Let  $w = \langle 3, \{1, 2, 3\} \rangle$  and  $v = \langle 5, \{4, 5, 6, 7\} \rangle$ .  $\lambda(w) = \lambda(v) = .2$ , but  $\tau(w) = .2 \neq \tau(v) = .6$ .

<sup>9</sup>Compare the theory of belief in [24], which implies (i) and a slight weakening of (ii) (with “more probable” in place of “at least as probable”), but not (iii). Rather than seeing this view as a competitor to ours, we prefer to see it as concerned with a weaker notion of belief that is less closely tied to knowledge. [26] has independently developed a theory of knowledge and belief very close to the one presented here, which he applies to the preface and St. Petersburg paradoxes.

<sup>10</sup>SUFFICIENCY does make some somewhat surprising predictions. Suppose Alice holds 999 tickets in a fair lottery and each of the million other entrants holds 1000 tickets. Assume  $P$  gives each ticket an equal probability of being chosen given the setup, that the question  $Q$  is *who will win*, and the threshold  $t = .99$ , and consider worlds in which the agent’s evidence is that this is the setup. If Alice will lose, can the agent know this? SUFFICIENCY predicts they can, which seems odd. (This might not seem so odd at first, since we don’t deny the agent can know this relative to a different question  $Q' = \text{how many tickets does the winner have}$ . To bring out the oddity, we can modify the example by adding another loser Bob who holds 1001 tickets; it seems odd that the agent would know that Alice will lose but not that Bob will lose.) To avoid this prediction, we might add a requirement that a world is sufficiently more normal than another only if it is also sufficiently more *likely*. (Compare the discussion of having no “*appreciably stronger reason*” (emphasis ours) to believe one entrant to a lottery will win compared to any other in [23, p.16].) We might implement this idea by adopting the stronger principle:

$$\text{SUFFICIENCY+: } w \gg v := 1 - \frac{\tau(v)}{\tau(w)} \geq t \text{ and } v \in R_e(w) \text{ and } 1 - \frac{\lambda(v)}{\lambda(w)} \geq t$$

This definition also determines a normality structure; while we actually find it more attractive than SUFFICIENCY, we will ignore it in the main text to simplify our presentation.

### 3 Inductive knowledge about the future

In this section and the next we'll review two cases of inductive knowledge involving coin flips from Dorr, Goodman and Hawthorne [10] to help illustrate the framework. The first case illustrates the possibility of inductive knowledge about the future:

**Flipping for Heads:** A coin flipper will flip a fair coin until it lands heads. Then he will flip no more.

Assume that the agent's evidence entails that this is the setup, and that they are watching the experiment unfold. We can model the case using the probability structure in which  $S = \{1, 2, \dots\}$ ,  $\mathcal{E} = \{\langle n, n+1, n+2, \dots \rangle : n \in S\}$ ,  $P$  is the probability function such that  $P(\{n\}) = 2^{-n}$  for all  $n$ ,  $Q = \{\{s\} : s \in S\}$ , and  $t = .99$ . Intuitively,  $n$  is the state in which the coin lands heads on the  $n$ th flip, the possible bodies of evidence are those compatible with having watched some initial sequence of zero or more flips that all landed tails, and  $Q$  is the question *on what flip does the coin land heads*. In the normality structure generated from this probability structure,  $\langle n, E \rangle \succ \langle m, E \rangle$  iff  $n \geq m$ , and  $\langle n, E \rangle \gg \langle m, E \rangle$  iff  $n \geq m + 7$ .<sup>11</sup> So if the agent has seen the coin land tails  $x$  times, what they believe is that it will land heads within the next 7 trials – i.e., on trials  $x + 1$  to  $x + 7$ . The predictions about knowledge match those of [10], on the Williamsonian definition of epistemic accessibility, and those of [19], on the Stalnakerian definition.

Notice that this model involves a kind of non-monotonic belief revision (which is prohibited by AGM). Let  $w = \langle 2, \{1, 2, \dots\} \rangle$ ,  $v = \langle 2, \{2, 3, \dots\} \rangle$ , and  $p = \{2, 3, \dots\}$ . So  $v$  is the result of discovering  $p$  in  $w$ . At the start of the experiment, the agent is in  $w$ ; after the first trial, the coin lands tails and the agent is in  $v$  (and, unbeknownst to them, the coin is about to land heads).  $\mathcal{R}_b(w) = \{1, \dots, 7\}$ , which is compatible with  $p$ , yet  $\mathcal{R}_b(w) \cap p \neq \mathcal{R}_b(v) = \{2, \dots, 8\}$ . Discovering something compatible with their prior beliefs leads the agent to give up some those beliefs. We think this is exactly the right prediction.<sup>12</sup>

### 4 Inductive knowledge of laws

The second case from [10] provides a simple model of inductive knowledge of lawful regularities:

**Heading for Heads:** You know a bag contains two coins: one fair, one double-headed. Without looking, you reach in and select a coin. You decide to flip it 100 times and observe how it lands.

If inductive knowledge of lawful regularities is ever possible, it should be possible here: that is, you can learn that the coin is double headed by seeing it land heads 100 times in a row. Our framework predicts this. Consider the probability structure with  $2^{100} + 1$  states, encoding the pattern of heads and tails and whether the coin is fair or double-headed. Let  $c$  be the state where the coin is fair but lands heads every time by coincidence, and  $d$  be the state where it is double-headed. For every state there are two worlds, corresponding to your evidence before and after flipping, so  $\mathcal{E} = \{S\} \cup \{c, d\} \cup \{\{s\} : s \in S \setminus \{c, d\}\}$ .  $P(\{d\}) = .5$  and  $P(\{s\}) = .5^{101}$  for  $s \in S \setminus \{d\}$ .  $Q = \{\{s\} : s \in S\}$  and  $t = .9999999$ . As desired, this structure predicts that, after seeing the coin land heads 100 times, you can know it is double headed.<sup>13</sup>

<sup>11</sup>  $\tau(\langle n, E \rangle) = 2^{1-n}$ , and  $1 - 2^{-6} < .99 < 1 - 2^{-7}$ , so  $1 - \frac{\tau(\langle m, E \rangle)}{\tau(\langle n, E \rangle)} \geq .99$  iff  $n \geq m + 7$ .

<sup>12</sup> Even more surprising behavior is possible if we modify the case by allowing the agent to get partial information about the result of the experiment after it is over, so that  $\{1, \dots, 7\} \in \mathcal{E}$ . Consider  $w = \langle 1, \{1, 2, \dots\} \rangle$ ,  $v = \langle 1, \{1, \dots, 7\} \rangle$ , and  $p = \{1, \dots, 7\}$ .  $\mathcal{R}_k(w) = \mathcal{R}_b(w) = p$  and  $v$  is the result of discovering  $p$  in  $w$ , yet  $\mathcal{R}_k(v) = \mathcal{R}_b(v) = \{1, \dots, 6\}$ : you can gain new beliefs and knowledge by discovering (i.e., gaining *evidential* knowledge of) something you already (inductively) knew.

<sup>13</sup> The theory similarly predicts that one can come to know how many sides of a die are painted red by observing the outcomes of a sufficient number of rolls of the die. It thus explains not only knowledge of law-like generalizations, but also knowledge of the objective chances associated with different physical processes.

Notice that, although  $\langle c, \{c, d\} \rangle \notin R_k(\langle d, \{c, d\} \rangle)$ , nevertheless  $\langle c, S \rangle \in R_k(\langle d, S \rangle)$ .<sup>14</sup> In other words: although you know the coin is double-headed after seeing it land heads every time, nevertheless, before flipping it, for all you knew the coin was fair and about to land heads every time by coincidence. This fact highlights a notable feature of the present framework that departs from our models in [19]: the comparative normality of two worlds is not a function only of those worlds' underlying states, but also depends on the agent's evidence. (This is perhaps more intuitive if we gloss  $\succ$  and  $\succcurlyeq$  as relations of comparative *plausibility*: what hypotheses about the state of the world are more or less plausible depends on what your evidence about the state of the world is.) The case also illustrates some more extreme departures from AGM-style dynamics of knowledge and belief, since discovering  $p$  (that the coin landed heads every time) can allow one to know  $q$  (that it isn't fair) even if, prior to the discovery,  $p \wedge \neg q$  was an epistemic possibility. While such behavior is unfamiliar (and claimed in [10] to be impossible), we again submit that in this case it is a plausible prediction.<sup>15</sup>

## 5 Inductive knowledge about multiple subject matters

Theories of inductive knowledge that accept THRESHOLD face a well-known challenge. Assuming inductive knowledge is possible at all, we should be able to know propositions whose evidential probability is less than 1. Now consider many independent subject matters about which we have such inductive knowledge, and the conjunction of everything we know concerning any one of these subject matters. If knowledge is closed under conjunction, as Hintikka semantics predicts, it follows that we know this conjunction. But if we are pooling knowledge across enough independent subject matters, then this conjunction is liable to have low evidential probability, even if the evidential probability of each of its conjuncts is high, thereby violating THRESHOLD.

Our proposal responds to this challenge by maintaining that “know” (and “believe”) are context-sensitive, with the question  $Q$  being the relevant parameter of context-sensitivity. For any given question, knowledge relative to that question is closed under conjunction. But there need be no single question relative to which an agent knows every proposition that they know relative to some question or other, and hence no question relative to which they know the conjunction of all such propositions. We will illustrate this aspect of the proposal using the following case we discuss in [19]:

**Racing for Heads:** Each of  $n$  coin flippers has a fair coin. Each will flip their coin until it lands heads.

If what the agent knows/believes about each coin flipper is like what they know/believe about the single coin flipper in **Flipping for Heads**, then as  $n$  grows the totality of what they know/believe can have arbitrarily low probability, violating THRESHOLD.<sup>16</sup> We will now explain our preferred treatment of the case, in terms of probability structures. We will show how its predictions about the agent's knowledge and beliefs depend on the contextually supplied question.

For concreteness, we will investigate the version of the case with 10 coin flippers. The probability structure is the one defined in the obvious way like in **Flipping for Heads**, with states modeled as

$${}^{14}1 - \frac{\tau(\langle c, \{c, d\} \rangle)}{\tau(\langle d, \{c, d\} \rangle)} = \frac{2^{100}}{2^{100}+1} \geq .9999999; \text{ but } 1 - \frac{\tau(\langle c, S \rangle)}{\tau(\langle d, S \rangle)} = .5 < .9999999.$$

<sup>15</sup>Note that the fact that, for all you know at the outset, the coin is fair and will land heads every time by coincidence, depends on the choice of question. Consider instead  $Q' = \textit{is the coin fair and how many times will it land heads}$ . Relative to this question, you do know at the outset that, if the coin is fair, it won't land heads every time. This is because the change in question from  $Q$  to  $Q'$  changes the typicality of  $\langle c, S \rangle$  from .5 to  $.5^{-100}$  – this is because fair-and-all-heads is less normal than all other states relative to  $Q'$  except for fair-and-all-tails (which is equally normal).

<sup>16</sup>[49] defends the existence of such THRESHOLD violations; [19] shows how they can be modeled using normality structures.

sequences of 10 positive integers, indicating how many trials it will take for each of the 10 coins to land heads, and  $P$  corresponding to the chances of various outcomes prior to the experiment.

What about  $Q$ ? In reflecting on **Racing for Heads** a number of natural questions suggest themselves. Some of these are the 10 different questions of the form *how many times will this particular coin land heads*. Relative to any such question, your knowledge is exactly like that in **Flipping for Heads** with respect to this particular coin, and trivial concerning every other coin. But other natural questions include (i) *what will the exact outcome of the whole experiment be*; (ii) *what will the shape of the outcome be* – that is, how many coins will take how long to land (the exact outcome up to isomorphism); (iii) *how many total tails will there be in the experiment as a whole*; (iv) *how long will it be before all the coins have landed heads*; and (v) *how many of the coins will ever land heads at the same time*. For each of these question, we can ask what you know and believe about a number of issues, such as how many total tails there will be, how long the experiment as a whole will last, and whether all the experiments will end on the same flip (a claim labelled ‘same end’ below). The table below records what the agent believes at the start of the experiment for different choices of  $Q$ , for thresholds  $t = .75$  and  $t = .95$ . This will also be what the agent knows in the most normal worlds.

$Q$	which worlds are most normal	$t$	min tails	max tails	min trials	max trials	same end?
(i) exact outcome	all coins land heads first time	.75	0	13	1	14	maybe
		.95	0	18	1	19	maybe
(ii) outcome shape	$6 \times 1$ flip, $3 \times 2$ flips, $1 \times 3$ flips	.75	1	15	2	8	no
		.95	0	22	1	12	maybe
(iii) how many total tails	8 or 9 total tails [tied]	.75	5	14	1	15	maybe
		.95	2	18	1	19	maybe
(iv) how long until over	ends on 4 <sup>th</sup> trial	.75	2	50	3	6	maybe
		.95	1	70	2	8	maybe
(v) how many end together	5 flippers get heads at once	.75	3	$\infty$	2	$\infty$	no
		.95	2	$\infty$	2	$\infty$	no

The table illustrates a general difference between relatively fine-grained questions (such as *exact outcome* or *outcome shape*) and more coarse-grained ones (such as *how many total tails*, *how long until over*, and *how many end together*). When the topic of our knowledge aligns with a coarse-grained question, we will generally know more relative to that question than we know relative to a more fine-grained question: for example, we know more about how many tails there will be relative to *how many total tails* than we do relative to *exact outcome* or *outcome shape*. But this increase in knowledge comes at a cost, since relative to a coarse-grained question we will know very little about the many topics that are orthogonal to that question: for example, we know little about how many tails there will be relative to *how long until over* or *how many end together*. These two facts share a common explanation. By treating all worlds that agree on the answer to a coarse-grained question as equally normal, we make it easier to exceed  $t$  as we add probabilities along the normality order while staying within a relatively restricted class of answers to that question, thus generating a lot of knowledge about that question. But in doing this, we will be including some worlds amongst the relatively normal ones in which things unfold in the least probable way they might with regard to some orthogonal subject matter.

## 6 Inductive knowledge from instrument readings

We often learn about the values of continuous quantities like weight and temperature by measuring them using less than perfectly reliable scales, thermometers, and so on. Modeling such knowledge requires two generalizations of the present framework. One concerns cases where there are a continuum of possible bodies of evidence, so  $P(E) = 0$  for some  $E \in \mathcal{E}$ . To handle such cases, we relax the requirement that evidential probabilities are always the result of conditioning a prior probability distribution on your evidence. Instead, we directly associate every  $E \in \mathcal{E}$  with a probability distribution  $P_E$  over  $E$ . A second problem concerns cases where  $Q$  has a continuum of answers all of which have evidential probability 0, yet we want to allow for non-trivial inductive knowledge. Here the natural solution is to generalize the operative notion of probability to *probability density*; see appendix B for the technical details. Our example in this section will illustrate both of these issues. Again, we will show that the framework allows us to derive (from purely probabilistic considerations) models of agents' knowledge and beliefs that have been defended in the literature on independent grounds.

Consider the kind of case made famous by Williamson [53]. You are going to glance at an unmarked modernist clock, with only an hour hand.  $S = [0, 2\pi) \times [0, 2\pi)$ , where  $\langle x, y \rangle$  is a state in which the hand's orientation is  $x$  (so that, e.g.,  $\frac{\pi}{2}$  represents 3-o'clock) and its apparent orientation when you look at it is  $y$ . We assume that, before looking at the clock, you have no idea how it will look or what time it is; after looking at the clock, your evidence is exhausted by how it appeared. That is,  $\mathcal{E} = \{S\} \cup \{\{s \in S : \pi_2(s) = y\} : y \in [0, 2\pi)\}$ , and  $P_S$  is uniform concerning both the hand's real and apparent orientations, in the sense that, for any interval  $I = [a, b] \subseteq [0, 2\pi)$ ,  $P_S(\{s : \pi_1(s) \in I\}) = P_S(\{s : \pi_2(s) \in I\}) = \frac{b-a}{2\pi}$ . Since your evidence after seeing the clock (i.e., that the apparent orientation was  $y$ ) had prior probability 0, your new evidential probabilities cannot be given by conditioning your prior evidential probabilities on your new discovery. Moreover, if the question is how the hand is oriented – i.e., if  $Q = \{\{s \in S : \pi_1(s) = x\} : x \in [0, 2\pi)\}$  – then, since your eyesight is imperfect, each of its answers will still have probability 0 after looking at the clock. The case thereby illustrates both issues described in the last paragraph.

To allow for non-trivial inductive knowledge concerning the position of the hand, we must modify NORMALITY AS LIKELINESS so that evidentially accessible worlds can differ in normality after looking at the clock. Fortunately, there is a natural way to do this. The key observation is that, after looking, not all *intervals* of orientations are on a par – their probabilities are no longer given merely by their length. Their probabilities are given instead by a non-uniform *probability density* function, a “bell curve” centered on the apparent orientation  $y$ . The area under this curve between two points gives the probability that the hand's true orientation is in that interval. This fact suggests modifying the definition of  $\succ$  as follows: rather than ordering worlds according to the respective *probabilities* of their answers to  $Q$ , we can instead order them by the respective *probability densities* (i.e., heights of the “bell curve”) of their answers to  $Q$ . The formal details are given in appendix B.

The normality structures generated in this way determine epistemic and doxastic accessibility relations of the same kind that have been defended in the literature. The agent's beliefs about the hand's orientation will be characterized by a non-trivial interval centered on its apparent orientation. Their knowledge will also be characterized by such an interval, in a way that may or may not always leave a “margin for error”, depending on whether we adopt a Williamsonian or a Stalnakerian definition of epistemic accessibility.<sup>17</sup> We believe that the present framework is strongly recommended by its ability to vindicate natural and anti-skeptical models of our knowledge in cases of this kind.

<sup>17</sup>Williamson [53] and Stalnaker [45] respectively defend such models of our knowledge about the unmarked clock. In [18], we discuss how other models in the literature arise from various combinations of definitions of epistemic accessibility and claims about the comparative normality of the relevant possibilities.

## 7 Conclusion

In this paper we have offered a new framework for modeling inductive knowledge and (rational) belief using resources congenial to philosophers in the Bayesian tradition. We did so by showing how the relations of comparative normality that have recently been used to model knowledge and belief can themselves be analyzed in probabilistic terms. The framework offers a unified account of our knowledge about chancy processes (section 3), lawful regularities (section 4), and imprecisely measured quantities (section 6). By positing a certain kind of context-sensitivity in “know” and “believe”, it offers a way of avoiding inductive skepticism while maintaining that only highly probable propositions are ever known or rationally believed (section 5). An urgent question for further research is how the contextually-supplied question that features in the probabilistic analysis of normality is determined.<sup>18</sup>

## Acknowledgements

Thanks to Cian Dorr and three anonymous referees for TARK for comments on a draft of this material, and to Kevin Dorst, John Hawthorne, Ben Holguín, and Harvey Lederman for very helpful discussion. Special thanks to Dorst for help with the coding required to calculate the values in the table in section 5.

## A Primitive evidential accessibility

Rather than modeling worlds as state/set-of-state pairs, we could treat them as unstructured points, and explicitly specify an evidential accessibility relation on them. This no longer allows us to model the notion of discovery, but it avoids the presupposition that there is a principled way of factoring worlds into state/set-of-state pairs. Moreover, it allows us to model evidential accessibility as an arbitrary reflexive relation. To handle cases where evidential accessibility is not an equivalence relation, we will need to relativize relations of comparative normality to a reference world – the world whose evidential probabilities are being used to assess the comparative normality of two other worlds.<sup>19</sup>

Let a *relativized normality structure* be a tuple  $\langle W, R_e, \succ, \succ_w \rangle$  such that:

1.  $W$  is a non-empty set,
2.  $R_e : W \rightarrow \mathcal{P}(W)$  such that  $w \in R_e(w)$  for all  $w \in W$ ,
3. For each  $w \in W$ ,  $\succ_w$  is a preorder on  $W$ ,
4. For each  $w \in W$ ,  $\succ_w$  is a well-founded relation on  $W$  such that, for any worlds  $w_1, w_2, w_3, w_4$ :
  - (a) If  $w_1 \succ_w w_2$ , then  $w_1 \succ_w w_2$ ;
  - (b) If  $w_1 \succ_w w_2 \succ_w w_3 \succ_w w_4$ , then  $w_1 \succ_w w_4$ .

We define  $R_b(w)$  and  $R_k(w)$  as in section 1, replacing  $\succ/\succ$  with  $\succ_w/\succ_w$ .

Now let a *worldly probability structure* be a tuple  $\langle W, R_e, Q, P, t \rangle$  such that:

<sup>18</sup>Note that relative to the question *is it true that p*, knowing that  $p$  requires only that  $p$  is true and has a high enough evidential probability. This might be considered objectionable for familiar reasons to do with Gettier cases [13]. If so, that would be one reason to deny that such questions are supplied by any context – although see [25] for arguments that there are contexts in which “knowledge” is this easy to come by.

<sup>19</sup>[48] argues that evidential accessibility is not an equivalence relation; [32] and [46] maintain that it is. [6] and [34] argue for a kind of world-relativity of (non-comparative) normality; see [18] for discussion in the context of comparative normality; ordering semantics for counterfactuals is a formal precedent, *cf.* [30].

1.  $W, R_e$  satisfy 1 and 2 in the definition of a relativized normality structure,
2.  $Q$  is a partition of  $W$ ,
3.  $P$  is a probability distribution over  $W$  such that  $P(q|R_e(w))$  is defined for all  $q \in Q$  and  $w \in W$ ,
4.  $t \in [0, 1]$ .

Let  $\lambda_w(v) := P([v]_Q|R_e(w))$  and  $\tau_w(v) := P(\{u : v \succ_w u \text{ and } u \in R_e(w)\}|R_e(w))$ . With these world-relative notions of likeliness and typicality in hand, we can now define  $v \succ_w u := \lambda_w(v) \geq \lambda_w(u)$  and  $v \gg_w u := 1 - \frac{\tau_w(u)}{\tau_w(v)} \geq t$ . It is easy to verify that these definitions yield a relativized normality structure that also obeys THRESHOLD (reformulated with  $R$  in place of the now ill-defined  $\mathcal{R}$ ).

## B Probability densities

Let a *density structure* be a tuple  $\langle S, \mathcal{E}, W, P, Q, m, f, t \rangle$  such that:<sup>20</sup>

1.  $S, \mathcal{E}, W, Q$ , and  $t$  are as in a probability structure and  $t > 0$ ,
2. For all  $E \in \mathcal{E}$ :
  - (a)  $P_E$  is a probability distribution over  $E$ ,
  - (b)  $P_E(q) = 0$  for all  $q \in Q$ ,
3.  $m : Q \rightarrow \mathbb{R}$  (the *measuring function*),
4.  $f : \mathcal{E} \rightarrow \mathbb{R}^{\mathbb{R}}$  such that  $f_E$  is the *density of  $P_E$  relative to  $m$* .

The intuitive idea behind a probability density function is that of a curve the area under which gives the associated probabilities. So, in particular,  $\int_a^b f_E(x)dx = P_E(\cup\{q \in Q : m(q) \in [a, b]\})$ . A formal characterization of  $f$  is given in a footnote; the role of  $m$  will be illustrated in appendix C.<sup>21</sup>

To understand this definition, let us return to the unmarked clock. As described in section 6,  $P$  satisfies clause 2 (where  $Q$  is *what is the hand's orientation*). Let  $m([s]_Q) = \pi_1(s)$  – i.e., it maps answers to  $Q$  (understood as sets of states) to corresponding real numbers in  $[0, 2\pi)$ .  $f_S(x) = \frac{1}{2\pi}$ , the constant function. By contrast,  $f_{\{s \in S : \pi_2(s) = y\}}$  – the probability density function determining your evidential probabilities after discovering that the hand's apparent orientation is  $y$  – will be a ‘bell curve’ centered on and symmetric around  $y$  (e.g. a ‘wrapped normal distribution’). In cases like this, rather than generating a normality order via NORMALITY AS LIKELINESS from a probability function, we instead generate it from this probability *density* function. Let  $d(w)$  (the *density of  $w$* ) be  $f_{\pi_2(w)}(m([\pi_1(w)]_Q))$ , and define being at least as normal as follows:

$$\text{NORMALITY AS DENSITY: } w \succ v := d(w) \geq d(v) \text{ and } v \in R_e(w)$$

As before,  $\gg$  is defined by SUFFICIENCY.<sup>22</sup>

To see this definition in action, we will consider a modification of the clock case that allows us to work with familiar Gaussian probability density functions (also known as ‘normal distributions’). We do so by considering a continuous quantity whose values are not confined to  $[0, 2\pi)$ , such as the difference in weight of two objects. Suppose  $Q$  is the question *how much heavier is this apple than this orange*,

<sup>20</sup>It would also be natural to require that, where possible,  $P$  behaves as if it were the result of conditioning a prior on the agent's evidence – i.e. that, for all  $E, E' \in \mathcal{E}$ , if  $P_E$  is defined on  $E \cap E'$ , then  $P_E(\cdot|E \cap E') = P_{E'}(\cdot|E \cap E')$ .

<sup>21</sup>Formally,  $f_E$  is the *density of  $P_E$  with respect to the reference measure  $\mu'$*  (defined as usual in terms of the Radon-Nikodym derivative), where  $\mu'(p) = \mu\{m(q) : q \subseteq p\}$  for all  $p$  on which  $P_E$  is defined, and  $\mu$  is the Lebesgue measure on  $\mathbb{R}$ .

<sup>22</sup>Requiring that  $t > 0$  ensures that  $\gg$  is well-founded; for an illustration of why this is needed, see  $d^l$  in appendix C.

and our scale reads  $\mu$  grams. Suppose, as a first approximation, that our evidential probabilities are now characterized by a Gaussian probability distribution over weight-difference in grams, with mean  $\mu$  and standard deviation  $\sigma$ .  $m$  will be the function from answers to  $Q$  to corresponding real numbers of grams and  $f_E(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ , where our evidence  $E = \{\langle x, \mu \rangle : x \in \mathbb{R}\}$ . (We model states as ordered pairs of actual and measured values of the quantity, as before.) By setting  $t = .9545\dots$ , we predict that  $\mathcal{R}_b(\langle \langle x, \mu \rangle, E \rangle) = \{\langle x', \mu \rangle : |x' - \mu| \leq 2\sigma\}$  for all  $x \in \mathbb{R}$ : we will believe that the scale reads  $\mu$  grams and that the true weight-difference is within two standard deviations of that. Predictions concerning knowledge depend on which of the two clauses for  $R_k$  are adopted, but in either case our knowledge will be non-trivial (and will coincide with what we believe when the scale is perfectly accurate – i.e., when  $x = \mu$ ). This model (simplified from [17]) shows how, at least in favorable circumstances, we can see familiar uses of confidence intervals in inferential statistic as corresponding to non-trivial knowledge and (rational) beliefs about the values of imprecisely measured quantities.<sup>23</sup>

Our view is not that normality relations are *always* determined by density structures. Sometimes they are determined by probability structures. It depends on whether the answers to  $Q$  have positive probabilities or only probability densities. When answers differ in this regard, we advocate a hybrid approach, with  $\succ$  determined by  $\lambda$  among pairs of answers one of which has positive probability and by  $d$  among pairs of answers with well-defined probability densities. Another generalization of density structures is also needed to handle multidimensional probability densities. In the case of  $n$  dimensions, we will then have  $f_E : \mathbb{R}^n \rightarrow \mathbb{R}$ . This generalization is needed when the relevant quantity is multidimensional (e.g., where a dart will land on a dartboard) and/or when  $Q$  concerns the outcomes of multiple independent noisy measurements of a given quantity (which will be modeled as a vector of real numbers).

## C Normality *de dicto* and *de se*

In the main text, we modelled a question as a partition of  $S$ . However, some natural questions cannot be modelled in this way, because they don't merely concern the state of the world. In **Flipping for Heads**, we might, for example, wonder *how many more times a coin will be flipped* – two worlds in which it is flipped a total of 5 times can differ on the answer to this question, because in one it is part of your evidence that the coin has already been flipped (only) 2 times while in another it is part of your evidence that it has been flipped 3 times. Following [31], we will think of these as *de se* questions, concerning not only the history of the world but also your place in it. Formally, we implement this idea by modeling  $Q$  as a partition of  $W$ , as in appendix A – recall that members of  $W$  should be thought of as *centered* worlds, two of which can agree on the complete history of the world while disagreeing on your evidence (because they concern your evidential situation at different times in that history).

Unlike appendix A, for present purposes we may keep  $P$  defined on subsets of  $S$  – this is natural in cases where we want  $P$  to conform to the prior objective chances which are (plausibly) only defined over histories of the world (i.e., states). Since  $P$  is now not defined on answers to  $Q$  (these being sets of worlds), we need to modify the definition of  $\lambda$ . For  $q \in Q$ , we let  $q_E = \{\pi_1(w) : w \in q \text{ and } \pi_2(w) = E\}$ ; then  $Q_E := \{q_E : q \in Q\}$  is a partition of  $E$ . So we can redefine  $\lambda(\langle s, E \rangle)$  as  $P([s]_{Q_E} | E)$ .<sup>24</sup>

<sup>23</sup>In the case of belief, our models correspond to the “minimum likelihood” method for generating confidence intervals (which yields the shortest possible intervals that have probability  $t$ ). The idea of ordering possibilities by probability density for this purpose goes back to [36]. This procedure has been criticized in the case of asymmetric distributions; in particular, in the case of  $\chi^2$  distributions, where it yields different verdicts from ordinary  $\chi^2$  tests [15, 37, 28]. We lack the space to address these criticisms here, except to note that the **Decay** example in appendix C illustrates why we find appealing the distinctive predictions that the minimum likelihood approach makes in the case of certain asymmetric distributions.

<sup>24</sup>This definition is a proper generalization of the previous one since, if we start with a partition  $Q^S$  of  $S$  and use it to generate

We will illustrate this generalization by showing how it allows us to model the following case (also discussed in [18]):

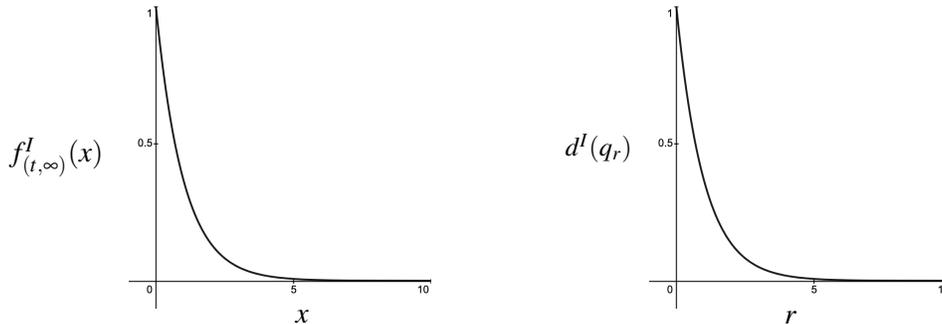
**Decay:** A radioactive atom is created; eventually, it will decay. The average time for an atom of this isotope to decay is one year.

Assume that the agent’s evidence entails that this is the set-up, that they are keeping track of time, and that, when the atom decays, they will be alerted of this fact. So states can be modeled as positive real numbers, specifying how many years from its creation it will be before the atom decays, and possible bodies of evidence (in which the atom has yet to decay) are intervals  $(t, \infty)$  for positive real  $t$ .

What should the agent believe about how many years after its creation the atom will decay? A natural thought is that, at every time  $t$  before the atom decays, the agent’s doxastic possibilities will be characterized by the interval  $[t + a, t + b]$ , for some  $0 < a < 1 < b$ . (Compare other events that tends to happen unexpectedly, like when you will next have an urge to sneeze, or when you will next get an email from an intermittent correspondent: you think they won’t happen in the next second, you think they will have happened within a decade, and the doxastically possible times form an interval, at least to a first approximation.) To ensure that the shape of this interval is the same at every time, we model the case using the *de se* question *how long after the current time will the atom decay*. To predict that the agent believes that the atom will take at least  $a$  years to decay, we need to choose the measuring function  $m$  of our density structure to reflect the fact that extremely short times to decay can be as far from average, in the relevant sense, as extremely long times to decay.

We will now make these ideas precise using the framework of density structures from appendix B (generalized to allow *de se* questions). Let  $S = \mathbb{R}^+$ ,  $\mathcal{E} = \{(t, \infty) : t \geq 0\}$ , and let  $P_{(t, \infty)}$  be given by the objective chances at  $t$  (which are entailed by the agent’s evidence at  $t$ , since it entails the setup and the time).  $Q = \{q_r : r \in \mathbb{R}^+\}$ , where  $q_r = \{(t', (t, \infty)) : t' - t = r\}$ . But this does not yet specify a density structure, since it is compatible with different choices of measuring functions  $m$  and corresponding probability density assignments  $f$ .

How should we associate answers to  $Q$  with real numbers? The simplest choice is the *index* function  $m^I$ , where  $m^I(q_r) = r$ . Given this choice, the corresponding density  $f_E^I(x) = e^{-x}$  (for all  $E \in \mathcal{E}$ , so the agent’s beliefs about how much longer it will be before the atom to decays don’t change if they see the atom hasn’t decayed yet), from which it follows that  $d^I(\langle t', (t, \infty) \rangle) = e^{t-t'}$ , since  $[\langle t', (t, \infty) \rangle]_Q = q_{t'-t}$ . Since the shortest lengths of time until decay correspond to the highest densities, it is always doxastically possible that the atom will decay arbitrarily soon.

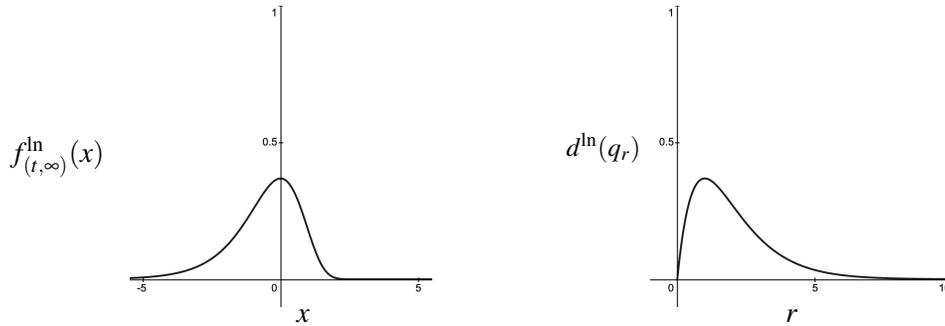


To avoid this prediction, we must choose a different measuring function. And there is a natural alternative: the *logarithmic* measuring function  $m^{\ln}$ , where  $m^{\ln}(q_r) = \ln(r)$ . This choice corresponds to

---

a partition  $Q^W$  of  $W$  in the obvious way, defining likeliness as above using  $Q^W$  yields the same results as the original definition of likeliness using  $Q^S$ .

a different density function  $f_E^{\ln} = e^x e^{-e^x}$  (for all  $E \in \mathcal{E}$ ), from which it follows that  $d^{\ln}(\langle t', (t, \infty) \rangle) = (t' - t)e^{t-t'}$ .<sup>25</sup> Intuitively, this choice of measuring function amounts to thinking of intervals between the present time and the time of decay in terms of the *order of magnitude* of their duration. In addition to being mathematically natural, this choice is psychologically well-motivated, as there is a large body of psychometric research showing that we perceive temporal duration in this way; see [14].



Unlike  $d^I(q_r)$  (which has its highest values as  $r$  approaches 0),  $d^{\ln}(q_r)$  approaches 0 as  $r$  approaches 0. This means that, for  $r < 1$ , smaller values of  $r$  make for increasing abnormality, and (centered) worlds in which the atom is going to decay extremely soon are hence doxastically inaccessible. More generally, the doxastically possible answers to  $Q$  are those whose value of  $d$  falls above a certain horizontal line (which depends on the probability threshold  $t$ ); inspection of the graph of  $d^{\ln}$  above shows that, as desired, this will always be an interval that excludes both possibilities in which the atom decays immediately and ones in which it won't decay for a very long time.

This model combines two ideas: the appeal to a *de se* question, and the appeal to a logarithmic measuring function. One might wonder whether the first of these is really necessary. Couldn't we have achieved the same effect by using a logarithmic measuring function on the *de dicto* question  $Q^*$ : *how long after its creation does the atom decay?* The answer is “no” – doing so achieves the same effect only at the moment the atom is created. This is because, unlike the probability densities of answers to  $Q$ , the probability densities of answers to  $Q^*$  change as new evidence becomes available: when the atom hasn't decayed 9 months after its creation, the probability that it will decay within the *next* year remains unchanged, but the probability that it will decay within its *first* year clearly decreases. (In particular, after a year, the answer to  $Q^*$  that implies that the atom will decay immediately will have the highest probability density even with a logarithmic measuring function; so possibilities in which the atom decays immediately will then be doxastically accessible.) Intuitively, it is unsurprising that a *de se* question is needed for the desired dynamical behavior in **Decay**. The idea that one is always entitled to believe that the atom isn't about to decay relies on there being something particularly odd about it decaying *right now* or *very soon*; but that oddity is clearly tied to *de se* notions, and cannot be articulated without them.

<sup>25</sup>By definition,  $f_E^{\ln}(x)$  is such that

$$\begin{aligned} \int_{-\infty}^y f_E^{\ln}(x) &= P_E\left(\bigcup\{q_r : m^{\ln}(q_r) \in (-\infty, y)\}\right) \\ &= \int_0^{e^y} e^{-x} dx \quad (\text{for all } E \in \mathcal{E}) \\ &= -e^{-e^y} + 1 \end{aligned}$$

Differentiating both sides with respect to  $y$  then yields  $f_E^{\ln}(x) = e^x e^{-e^x}$ . So  $d^{\ln}(\langle t', (t, \infty) \rangle) = f_{(t, \infty)}^{\ln}(m^{\ln}(\langle t', (t, \infty) \rangle_Q)) = f_{(t, \infty)}^{\ln}(\ln(t' - t)) = (t' - t)e^{t-t'}$ .

## References

- [1] Carlos Alchourrón, Peter Gärdenfors & David Makinson (1985): *On the Logic of Theory Change: Partial Meet Contraction and Revision Functions*. *Journal of Symbolic Logic* 50, pp. 510–530, doi:10.2307/2274239.
- [2] Andrew Bacon (2014): *Giving your Knowledge Half a Chance*. *Philosophical Studies* 171, pp. 373–397, doi:10.1007/s11098-013-0276-6.
- [3] Andrew Bacon (2020): *Inductive Knowledge*. *Noûs* 54, pp. 354–388, doi:10.1111/nous.12266.
- [4] Bob Beddor & Carlotta Pavese (2020): *Modal Virtue Epistemology*. *Philosophy and Phenomenological Research* 101, pp. 61–79, doi:10.1111/phpr.12562.
- [5] Kyle Blumberg & Harvey Lederman (2021): *Revisionist Reporting*. *Philosophical Studies* 178, pp. 755–783, doi:10.1007/s11098-020-01457-4.
- [6] Sam Carter (2019): *Higher order ignorance inside the margins*. *Philosophical Studies* 176, pp. 1789–1806, doi:10.1007/s11098-018-1096-5.
- [7] Sam Carter & Simon Goldstein (forthcoming): *The Normality of Error*. *Philosophical Studies*, doi:10.1007/s11098-020-01560-6.
- [8] Stewart Cohen & Juan Comesaña (2013): *Williamson on Gettier Cases and Epistemic Logic*. *Inquiry* 56, pp. 15–29, doi:10.1080/0020174X.2013.775012.
- [9] Andrew Dabrowski, Lawrence Moss & Rohit Parikh (1996): *Topological Reasoning and the logic of knowledge*. *Annals of Pure and Applied Logic* 78, pp. 73–110, doi:10.1016/0168-0072(95)00016-X.
- [10] Cian Dorr, Jeremy Goodman & John Hawthorne (2014): *Knowing against the odds*. *Philosophical Studies* 170, pp. 277–287, doi:10.1007/s11098-013-0212-9.
- [11] Julien Dutant (2016): *How to be an infallibilist*. *Philosophical Issues* 26, pp. 148–171, doi:10.1111/phis.12085.
- [12] Julien Dutant & Sven Rosenkranz (2020): *Inexact Knowledge 2.0*. *Inquiry* 63, pp. 812–830, doi:10.1080/0020174X.2020.1754286.
- [13] Edmund Gettier (1963): *Is Justified True Belief Knowledge?* *Analysis* 23, pp. 121–123, doi:10.1093/analysis/23.6.121.
- [14] John Gibbon, Russell M. Church & Warren H. Meck (1984): *Scalar timing in memory*. *Annals of the New York Academy of sciences* 423, pp. 52–77, doi:10.1111/j.1749-6632.1984.tb23417.x.
- [15] Jean Gibbons & John Pratt (1975): *P-Values: Interpretation and Methodology*. *The American Statistician* 29, pp. 20–25, doi:10.1080/00031305.1975.10479106.
- [16] Simon Goldstein & John Hawthorne (forthcoming): *Counterfactual Contamination*. *Australasian Journal of Philosophy*, doi:10.1080/00048402.2021.1886129.
- [17] Jeremy Goodman (2013): *Inexact Knowledge without Improbable Knowing*. *Inquiry* 56, pp. 30–53, doi:10.1080/0020174X.2013.775013.
- [18] Jeremy Goodman & Bernhard Salow: *Epistemology Normalized*. Unpublished ms.
- [19] Jeremy Goodman & Bernhard Salow (2018): *Taking a chance on KK*. *Philosophical Studies* 175, pp. 183–196, doi:10.1007/s11098-017-0861-1.
- [20] Nelson Goodman (1955): *Fact, Fiction, and Forecast*. Harvard UP, doi:10.2307/2964684.
- [21] Daniel Greco (2014): *Could KK be OK?* *Journal of Philosophy* 111, pp. 169–197, doi:10.5840/jphil2014111411.
- [22] Adam Grove (1988): *Two modellings for theory change*. *Journal of Philosophical Logic* 17, pp. 157–170, doi:10.1007/BF00247909.
- [23] John Hawthorne (2004): *Knowledge and Lotteries*. Oxford UP, doi:10.1093/0199269556.001.0001.
- [24] Ben Holguín: *Thinking, Guessing, and Believing*. Unpublished ms.

- [25] Ben Holguín (forthcoming): *Knowledge by Constraint. Philosophical Perspectives.*
- [26] Frank Hong (in preparation): *Uncertain Knowledge.* Ph.D. thesis, University of Southern California.
- [27] Sarit Kraus, Daniel Lehmann & Menachem Magidor (1990): *Nonmonotonic reasoning, preferential models, and cumulative logics.* *Artificial Intelligence* 44, pp. 167–207, doi:10.1016/0004-3702(90)90101-5.
- [28] Elena Kulinskaya (2008): *On two-sided p-values for non-symmetric distributions.* Available at <https://arxiv.org/abs/0810.2124>.
- [29] Hannes Leitgeb (2014): *The Stability Theory of Belief.* *Philosophical Review* 123, pp. 131–171, doi:10.1215/00318108-2400575.
- [30] David Lewis (1973): *Counterfactuals.* Blackwell, doi:10.2307/2273738.
- [31] David Lewis (1979): *Attitudes de dicto and de se.* *Philosophical Review* 88, pp. 513–543, doi:10.2307/2184843.
- [32] David Lewis (1996): *Elusive Knowledge.* *Australasian Journal of Philosophy* 74, pp. 549–567, doi:10.1080/00048409612347521.
- [33] Clayton Littlejohn & Julien Dutant (2020): *Justification, Knowledge, and Normality.* *Philosophical Studies* 177, pp. 1593–1609, doi:10.1007/s11098-019-01276-2.
- [34] Annina Loets (forthcoming): *Choice Points for a Theory of Normality.* *Mind.*
- [35] David Makinson (1993): *Five Faces of Minimality.* *Studia Logica* 52, pp. 339–379, doi:10.1007/BF01057652.
- [36] Jerzy Neyman & Egon Pearson (1931): *Further Notes on the  $\chi^2$  Distribution.* *Biometrika* 22, pp. 298–305, doi:10.1093/biomet/22.3-4.298.
- [37] Robert Radlow & Edward Alf (1975): *An Alternate Multinomial Assessment of the Accuracy of the  $\chi^2$  Test of Goodness of Fit.* *Journal of the American Statistical Association* 70, pp. 811–813, doi:10.1080/01621459.1975.10480306.
- [38] Jonathan Schaffer & Zoltán Szabó (2014): *Epistemic comparativism: a contextualist semantics for knowledge ascriptions.* *Philosophical Studies* 168, pp. 491–543, doi:10.1007/s11098-013-0141-7.
- [39] Martin Smith (2010): *What Else Justification Could Be.* *Noûs* 44, pp. 10–31, doi:10.1111/j.1468-0068.2009.00729.x.
- [40] Martin Smith (2016): *Between Probability and Certainty: What Justifies Belief.* Oxford UP, doi:10.1093/acprof:oso/9780198755333.001.0001.
- [41] Martin Smith (2017): *Why throwing 92 Heads in a row is not surprising.* *Philosophers' Imprint* 17, pp. 1–8. Available at <http://hdl.handle.net/2027/spo.3521354.0017.021>.
- [42] Martin Smith (2018): *Coin Trials.* *Canadian Journal of Philosophy* 48, pp. 726–741, doi:10.1080/00455091.2017.1381936.
- [43] Martin Smith (2018): *The Logic of Epistemic Justification.* *Synthese* 195, pp. 3857–3875, doi:10.1007/s11229-017-1422-z.
- [44] Robert Stalnaker (2006): *On Logics of Knowledge and Belief.* *Philosophical Studies* 128, pp. 169–99, doi:10.1007/s11098-005-4062-y.
- [45] Robert Stalnaker (2015): *Luminosity and the KK Thesis.* In S. Goldberg, editor: *Externalism, Self-Knowledge, and Skepticism*, Cambridge UP, pp. 17–40, doi:10.1017/CBO9781107478152.002.
- [46] Robert Stalnaker (2019): *Contextualism and the Logic of Knowledge.* In: *Knowledge and Conditionals*, Oxford UP, pp. 129–148, doi:10.1093/oso/9780198810346.003.0009.
- [47] Brian Weatherson (2013): *Margins and Errors.* *Inquiry* 56, pp. 63–76, doi:10.1080/0020174X.2013.775015.
- [48] Timothy Williamson (2000): *Knowledge and its Limits.* Oxford UP, doi:10.1111/j.1933-1592.2005.tb00537.x.
- [49] Timothy Williamson (2009): *Probability and Danger.* *The Amherst Lecture in Philosophy* 4, pp. 1–35. Available at <http://www.amherstlecture.org/williamson2009/>.

- [50] Timothy Williamson (2011): *Improbable Knowing*. In Trent Dougherty, editor: *Evidentialism and its Discontents*, Oxford UP, Oxford, pp. 147–164, doi:10.1093/acprof:oso/9780199563500.003.0010.
- [51] Timothy Williamson (2013): *Gettier Cases in Epistemic Logic*. *Inquiry* 56, pp. 1–14, doi:10.1080/0020174X.2013.775010.
- [52] Timothy Williamson (2013): *Response to Cohen, Comesaña, Goodman, Nagel, and Weatherson on Gettier Cases in Epistemic Logic*. *Inquiry* 56, pp. 77–96, doi:10.1080/0020174X.2013.775016.
- [53] Timothy Williamson (2014): *Very Improbable Knowing*. *Erkenntnis* 79, pp. 971–999, doi:10.1007/s10670-013-9590-9.
- [54] Timothy Williamson (forthcoming): *The KK principle and rotational symmetry*. *Analytic Philosophy*, doi:10.1111/phib.12203.
- [55] Stephen Yablo (2014): *Aboutness*. Princeton UP, doi:10.1515/9781400845989.
- [56] Seth Yalcin (2018): *Belief as Question-Sensitive*. *Philosophy and Phenomenological Research* 97, pp. 23–47, doi:10.1111/phpr.12330.

# Belief Inducibility and Informativeness

P. Jean-Jacques Herings\*

p.herings@maastrichtuniversity.nl

Dominik Karos†

dominik.karos@uni-bielefeld.de

Toygar Kerman‡

t.kerman@maastrichtuniversity.nl

We consider a group of receivers who share a common prior on a finite state space and who observe private correlated signals that are contingent on the true state of the world. We show that, while necessary, Bayes plausibility is not sufficient for a distribution over posterior belief vectors to be inducible, and we provide a characterization of inducible distributions. We classify communication strategies as minimal, direct, and language independent, and we show that any inducible distribution can be induced by a language independent communication strategy (LICS). We investigate the role of the different classes of communication strategies for the amount of higher order information that is revealed to receivers. We show that the least informative communication strategy which induces a fixed distribution over posterior belief vectors lies in the relative interior of the set of all language independent communication strategies which induce that distribution.

---

\*Department of Microeconomics and Public Economics (MPE), Maastricht University

†Center for Mathematical Economics, Bielefeld University

‡Department of Microeconomics and Public Economics (MPE), Maastricht University



# Measuring Violations of Positive Involvement in Voting\*

Wesley H. Holliday

University of California, Berkeley

wesholliday@berkeley.edu

Eric Pacuit

University of Maryland

epacuit@umd.edu

In the context of computational social choice, we study voting methods that assign a set of winners to each profile of voter preferences. A voting method satisfies the property of positive involvement (PI) if for any election in which a candidate  $x$  would be among the winners, adding another voter to the election who ranks  $x$  first does not cause  $x$  to lose. Surprisingly, a number of standard voting methods violate this natural property. In this paper, we investigate different ways of measuring the extent to which a voting method violates PI, using computer simulations. We consider the probability (under different probability models for preferences) of PI violations in randomly drawn profiles vs. profile-coalition pairs (involving coalitions of different sizes). We argue that in order to choose between a voting method that satisfies PI and one that does not, we should consider the probability of PI violation conditional on the voting methods choosing different winners. We should also relativize the probability of PI violation to what we call voter potency, the probability that a voter causes a candidate to lose. Although absolute frequencies of PI violations may be low, after this conditioning and relativization, we see that under certain voting methods that violate PI, much of a voter’s potency is turned against them—in particular, against their desire to see their favorite candidate elected.

## 1 Introduction

Voting provides a mechanism for resolving conflicts between the preferences of multiple agents in order to arrive at a group choice. Although traditional voting theory is largely motivated by the example of democratic political elections, the field of computational social choice [3, 4, 16] views voting theory as applicable to many other preference aggregation problems for multiagent systems.

One of the most basic ideas in voting is that an unequivocal increase in support for a candidate should not result in that candidate going from being a winner to being a loser. There are at least two ways to formalize this idea. First, there is the fixed-electorate axiom of *monotonicity*: if a candidate  $x$  is a winner given a preference profile  $\mathbf{P}$ , and  $\mathbf{P}'$  is obtained from  $\mathbf{P}$  by one voter moving  $x$  up in their ranking, then  $x$  should still be a winner given  $\mathbf{P}'$ . Second, there is the variable-electorate axiom of *positive involvement* [43, 41]: if a candidate  $x$  is a winner given  $\mathbf{P}$ , and  $\mathbf{P}^*$  is obtained from  $\mathbf{P}$  by adding a new voter who ranks  $x$  in first place, then  $x$  should still be a winner given  $\mathbf{P}^*$ . These axioms are logically independent. For example, Instant Runoff Voting (see Section 2.2) satisfies positive involvement but not monotonicity; and a number of well-known voting methods satisfy monotonicity but not positive involvement.

There are at least two reactions to a voting method’s violating monotonicity or positive involvement. One is that the method’s violating one of these axioms is a serious problem if and only if violations are sufficiently frequent (according to some probability model). Another reaction is that the method’s violating one of these axioms is a sign that the principle by which the voting method selects winners is fundamentally misconceived, which casts suspicion on its selection of winners in general, not just in those cases where the axiom is violated; but of course it is even worse for the method if in addition to having a misconceived principle for selecting winners, it frequently witnesses violations of the axiom.

---

\*We thank the three anonymous reviewers for TARK for their helpful comments.

In this paper, we discuss different ways of measuring the extent to which a voting method violates positive involvement (in particular, the eight different ways shown in Figure 1), using computer simulations. Our main aim is to determine appropriate measures of positive involvement violation and identify the main parameters affecting these measures—e.g., whether they are affected by variants of voting methods, numbers of candidates and voters, probability models, etc.—which is a necessary precursor to using positive involvement as part of an argument in favor of some voting methods over others in future work.

First, we make some stage-setting conceptual points in Sections 1.1-1.2, discuss related work in Section 1.3, and recall technical preliminaries in Section 2. Section 3 contains our main discussion, as well as results of our simulations. Our methodological points in Section 3 can be applied to measuring the extent of violation of other axioms in addition to positive involvement. We conclude in Section 4.

## 1.1 Strategic considerations

It is important to distinguish the perspective on positive involvement adopted above—we view it as an axiom, like monotonicity, that rules out perverse responses to unequivocal increases in support for a candidate—from a *strategic perspective* on positive involvement, which is also adopted in the literature. A voting method’s violating positive involvement may give a voter an incentive for *strategic abstention*: if a voter  $i$  knows that by casting a sincere ballot with her favorite candidate  $a$  ranked first, this would kick  $a$  out of the set of winners, then  $i$  may prefer not to vote rather than to cast a sincere ballot. This is so whenever  $a$  would be the unique winner were  $i$  not to vote but  $a$  would not be were  $i$  to vote sincerely.

However, there are two important qualifications. First, if we are considering strategic abstention, we may also consider *strategic voting*. For certain voting methods, whenever a candidate  $a$  would be the unique winner were  $i$  not to vote, there is always *some* linear order with  $a$  ranked first that  $i$  can cast as her ballot to keep  $a$  the unique winner,<sup>1</sup> so  $i$  has no incentive to abstain if she can vote strategically.<sup>2</sup> Second, we must consider the case where  $a$  would not be the *unique* winner were  $i$  not to vote. E.g., suppose that if  $i$  were not to vote, the result would be a tie between  $a$  and  $c$ , where  $c$  is  $i$ ’s least favorite candidate, whereas if  $i$  were to vote sincerely, the unique winner would be  $b$ , who is  $i$ ’s second favorite candidate. Depending on  $i$ ’s utility function,  $i$  may well prefer  $b$  to a tiebreaking process applied to  $\{a, c\}$ , in which case  $i$  would prefer to vote sincerely rather than to abstain. So not every situation witnessing a failure of positive involvement is one in which the voter would prefer to abstain rather than vote sincerely.

Thus, the problem with violating positive involvement is not just that it may incentivize strategic abstention. Even in a society in which all voters always vote and always vote sincerely, so there is no risk of strategic abstention instead of sincere voting, we still find violations of positive involvement perverse responses to unequivocal increases in support for a candidate.

## 1.2 Positive involvement vs. participation

It is also important to distinguish positive involvement from the axiom of *participation* [37]. Participation is usually stated for *resolute* voting methods that map each profile of voter preferences to a *unique* winning candidate.<sup>3</sup> A resolute voting method satisfies participation (and an arbitrary voting method

<sup>1</sup>When a voter casts a ballot with her true favorite  $a$  ranked first but with deviations from her sincere preference lower down on the ballot in order to get  $a$  elected, Dowding and Van Hees [14] call this a *sincere manipulation*.

<sup>2</sup>This is obvious if  $i$  can submit as her ballot a strict weak order instead of a linear order, since then she may submit a fully indifferent ballot, which for many voting methods is equivalent to abstention, or a ballot with  $a$  on top followed by an indifference class containing all other candidates. However, here we assume ballots are linear orders.

<sup>3</sup>Resolute voting methods (defined on any profile of voter preferences, as in Definition 3) either fail to treat voters equally—by violating the axiom of anonymity—or fail to treat candidates equally—by violating the axiom of neutrality (see [54, § 2.3]).

satisfies what could be called *resolute participation*) if adding a new voter who ranks  $x$  above  $y$  cannot result in a change from  $x$  being the unique winner to  $y$  being the unique winner.<sup>4</sup> Crucially, it is not required that the new voter ranks  $x$  in first place. Thus, the new voter may rank other candidates above  $x$ , thereby hurting  $x$ 's prospects. Participation says that ranking those other candidates above  $x$  should not hurt  $x$  more than it hurts  $y$ , so it should not result in  $y$  becoming the unique winner. But this is not so clear in profiles containing majority cycles. In the presence of cycles, the main threat to  $x$  may be a candidate  $z$  who does not threaten  $y$ . Then a new voter with a ranking of the form  $z \dots xy \dots$  may do more harm to  $x$  than to  $y$ . This is only a sketch of an argument raising doubts about the participation axiom, but the key point is this: unlike violations of positive involvement, some violations of participation are not perverse responses to unequivocal increases in support for a candidate  $x$ , as some violations involve adding a voter who ranks other candidates above  $x$ . Of course, violations of participation give a voter an incentive to abstain rather than to vote sincerely, but that is a separate strategic issue that we set aside.

The term “No Show Paradox” was introduced by Fishburn and Brams [23] for violations of what is now called *negative involvement*.<sup>5</sup> Later Moulin [37] changed the meaning of “No Show Paradox” to refer to violations of participation. Violations of positive and negative involvement are called instances of the “Strong No Show Paradox” by Pérez [41].<sup>6</sup> Pérez concludes that the Strong No Show Paradox is a common flaw of many *Condorcet consistent* voting methods, which are methods that always pick a Condorcet winner—a candidate who is majority preferred to every other candidate—if one exists.

### 1.3 Related Work

This study fits into a line of recent work using computer simulations to estimate the frequency of violations of various voting criteria for different voting methods [42, 25, 6, 5]. The most closely related previous study is that of Brandt et al. [6] on the frequency of the “No Show Paradox” in Moulin’s sense [37], i.e., the frequency of violations of the participation axiom. As the authors observe, “While it is known that certain voting rules suffer from this paradox in principle, the extent to which it is of practical concern is not well understood” (p. 520), a comment that also applies to violations of positive involvement. To fill the gap in the case of the No Show Paradox, Brandt et al. use computer simulations—as well as analytic results obtained using Ehrhart theory—for six Condorcet consistent voting methods, three of which overlap with our list: Baldwin, Copeland, and Nanson. They find that “for few alternatives, the probability of the NSP is rather small (less than 4% for four alternatives and all considered preference models, except for Copeland’s rule). As the number of alternatives increases, the NSP becomes much more likely and which rule is most susceptible to abstention strongly depends on the underlying distribution of preferences” (p. 520). This observation about alternatives matches what we find for positive involvement, where the frequency of violations significantly increases with the number of alternatives.

---

<sup>4</sup>For other definitions of participation for irresolute voting methods, see [41, 31, 44].

<sup>5</sup>The axiom of negative involvement [43, 41] states that adding a new voter who ranks a candidate last should not result in the candidate going from being a loser to a winner. The analysis of this paper can also be applied to negative involvement, but for the sake of space we focus on positive involvement.

<sup>6</sup>Plassman and Tideman [42] use “Strong No Show Paradox” to refer to violations of participation, contradicting Pérez’s use. Pérez [41] calls violations of positive involvement the “positive strong no show paradox” and violations of negative involvement the “negative strong no show paradox.” Felsenthal and Tideman [20] and Felsenthal and Nurmi [18, 19] call them the “P-TOP” and “P-BOT” paradoxes, respectively.

## 2 Preliminaries

### 2.1 Profiles

Fix infinite sets  $\mathcal{V}$  and  $\mathcal{X}$  of *voters* and *candidates*, respectively. A given election will use only finite subsets  $V \subseteq \mathcal{V}$  and  $X \subseteq \mathcal{X}$ . We consider elections in which each voter submits a ranking of all the candidates, which we assume is a strict linear order. For a set  $X$ , let  $\mathcal{L}(X)$  be the set of all strict linear orders on  $X$ . For  $P \in \mathcal{L}(X)$ , we write ‘ $xPy$ ’ for  $(x, y) \in P$ ; for  $a \in X$ , let  $\text{Rank}(a, P) = |\{b \in X \mid bPa\}| + 1$ .

**Definition 1.** A *profile* is a function  $\mathbf{P} : V \rightarrow \mathcal{L}(X)$  for some nonempty finite  $V \subseteq \mathcal{V}$ , which we denote by  $V(\mathbf{P})$  (called the set of *voters in  $\mathbf{P}$* ), and nonempty finite  $X \subseteq \mathcal{X}$ , which we denote by  $X(\mathbf{P})$  (called the set of *candidates in  $\mathbf{P}$* ). We call  $\mathbf{P}(i)$  voter  $i$ ’s *ballot* and write ‘ $\mathbf{P}_i$ ’ for  $\mathbf{P}(i)$ .

**Definition 2.** For a profile  $\mathbf{P}$  and set  $C \subseteq V(\mathbf{P})$  of voters, let  $\mathbf{P}_{-C}$  be the restriction of  $\mathbf{P}$  to  $V(\mathbf{P}) \setminus C$ . For a single voter  $i \in V(\mathbf{P})$ , let  $\mathbf{P}_{-i} = \mathbf{P}_{-\{i\}}$ . Given profiles  $\mathbf{P}$  and  $\mathbf{P}'$  such that  $X(\mathbf{P}) = X(\mathbf{P}')$  and  $V(\mathbf{P}) \cap V(\mathbf{P}') = \emptyset$ , we define the profile  $\mathbf{P} + \mathbf{P}' : V(\mathbf{P}) \cup V(\mathbf{P}') \rightarrow \mathcal{L}(X(\mathbf{P}))$  as follows: if  $i \in V(\mathbf{P})$ , then  $(\mathbf{P} + \mathbf{P}')(i) = \mathbf{P}(i)$ , and if  $i \in V(\mathbf{P}')$ , then  $(\mathbf{P} + \mathbf{P}')(i) = \mathbf{P}'(i)$ .

### 2.2 Voting Methods

**Definition 3.** A *voting method* is a function  $F$  on the domain of all profiles such that for any profile  $\mathbf{P}$ ,  $\emptyset \neq F(\mathbf{P}) \subseteq X(\mathbf{P})$ .

In this paper, we consider the following voting methods. We chose these methods because they give a broad representation of different classes of methods (e.g., scoring rules, iterative methods, and Condorcet methods, including C1 and C2 methods) and can be efficiently computed (except for Ranked Pairs).

**Positional scoring rules.** A *scoring vector* is a vector  $\langle s_1, \dots, s_n \rangle$  of numbers such that for each  $m \in \{1, \dots, n-1\}$ ,  $s_m \geq s_{m+1}$ . Given a profile  $\mathbf{P}$  with  $|X(\mathbf{P})| = n$ ,  $x \in X(\mathbf{P})$ , a scoring vector  $\vec{s}$  of length  $n$ , and  $i \in V(\mathbf{P})$ , define  $\text{score}_{\vec{s}}(x, \mathbf{P}_i) = s_r$  where  $r = \text{Rank}(x, \mathbf{P}_i)$ . Let  $\text{score}_{\vec{s}}(x, \mathbf{P}) = \sum_{i \in V(\mathbf{P})} \text{score}_{\vec{s}}(x, \mathbf{P}_i)$ . A voting method  $F$  is a *positional scoring rule* if there is a map  $\mathcal{S}$  assigning to each natural number  $n$  a scoring vector of length  $n$  such that for any profile  $\mathbf{P}$  with  $|X(\mathbf{P})| = n$ ,  $F(\mathbf{P}) = \text{argmax}_{x \in X(\mathbf{P})} \text{score}_{\mathcal{S}(n)}(x, \mathbf{P})$ . Two well-known examples of scoring rules are:

$$\mathbf{Plurality}: \mathcal{S}(n) = \langle 1, 0, \dots, 0 \rangle \quad \text{and} \quad \mathbf{Borda}: \mathcal{S}(n) = \langle n-1, n-2, \dots, 1, 0 \rangle.$$

**Instant Runoff** (also known as Alternative Vote, Ranked Choice, and Single Transferable Vote): Iteratively remove all candidates with the fewest number of voters who rank them first, until there is a candidate who is a majority winner (i.e., ranked first by a strict majority of voters).<sup>7</sup> If, at some stage of the removal process, all remaining candidates have the same number of voters who rank them first (so all candidates would be removed), then all remaining candidates are selected as winners.

**Coombs** [11, 27]: Iteratively remove all candidates with the most number of voters who rank them last, until there is a candidate who is a majority winner. If, at some stage of the removal process, all remaining candidates have the same number voters who rank them last (so all candidates would be removed), then all remaining candidates are selected as winners.

**Baldwin** [1]: Iteratively remove all candidates with the smallest Borda score, until there is a single candidate remaining. If, at some stage of the removal process, all remaining candidates have the same Borda score (so all candidates would be removed), then all remaining candidates are selected as winners.

<sup>7</sup>When there is more than one candidate with the fewest number of voters who rank them first, this definition of Instant Runoff, taken from [49, p. 7], eliminates all of them. For another way of handling such ties, see [51] and Footnote 14.

**Nanson** [38]: There are two versions of this voting method [40]. **Strict Nanson** (resp. **Weak Nanson**) iteratively removes all candidates whose Borda score is strictly less than (resp. less than or equal to) the average Borda score of the candidates remaining at that stage, until one candidate remains. If, at some stage of the removal process, all remaining candidates have the same Borda score (so all candidates would be removed), then all remaining candidates are selected as winners. Although Nanson seems to have intended **Weak Nanson** (see [40]), the literature on computational social choice usually interprets ‘Nanson’ as **Strict Nanson** (see, e.g., [39, 54, 21, 10]), so we focus on **Strict Nanson** in this paper.

**Bucklin** [28]: Given a candidate  $a$  in a profile  $\mathbf{P}$  and a positive integer  $n$ , say that  $a$  is an  $n$ -th level majority winner in  $\mathbf{P}$  if a strict majority of voters rank  $a$  in  $n$ -th place or higher (thus, a majority winner in the usual sense is a 1st level majority winner). Where  $k$  is the smallest positive integer for which there is at least one  $k$ -th level majority winner, Bucklin selects as winners the  $k$ -th level majority winners for whom the most voters rank them in  $k$ -th place or higher.<sup>8</sup>

The other methods that we study in this paper are from a broad class of so-called majority or margin-based methods. We need the following notation for defining these methods.

**Definition 4.** For a profile  $\mathbf{P}$  and  $a, b \in X(\mathbf{P})$ , let  $\text{Margin}_{\mathbf{P}}(a, b) = |\{i \in V(\mathbf{P}) \mid a\mathbf{P}_i b\}| - |\{i \in V(\mathbf{P}) \mid b\mathbf{P}_i a\}|$  be the margin of  $a$  over  $b$  in  $\mathbf{P}$ . The *margin graph* of  $\mathbf{P}$  is the weighted directed graph whose set of vertices is  $X(\mathbf{P})$  with an edge from  $a$  to  $b$  when  $\text{Margin}_{\mathbf{P}}(a, b) > 0$ , weighted by  $\text{Margin}_{\mathbf{P}}(a, b)$ .

**Copeland** [12]: The Copeland score of  $a \in X(\mathbf{P})$  is the number of  $b \in X(\mathbf{P})$  such that  $\text{Margin}_{\mathbf{P}}(a, b) > 0$  minus the number of  $b \in X(\mathbf{P})$  such that  $\text{Margin}_{\mathbf{P}}(b, a) > 0$ . Then  $\text{Copeland}(\mathbf{P})$  is the set of  $x \in X(\mathbf{P})$  with maximal Copeland score.<sup>9</sup>

**Top Cycle** [48, 46] (also known as Smith and GETCHA): For a profile  $\mathbf{P}$ , let  $a \succ_{\mathbf{P}} b$  if  $\text{Margin}_{\mathbf{P}}(a, b) > 0$ . Let  $\succ_{\mathbf{P}}^*$  be the transitive closure of  $\succ_{\mathbf{P}}$ . Then  $\text{TopCycle}(\mathbf{P}) = \{x \in X(\mathbf{P}) \mid \text{for all } y \in X(\mathbf{P}), x \succ_{\mathbf{P}}^* y\}$ .<sup>10</sup>

**Uncovered Set** (Gillies version) [26]: Given  $a, b \in X(\mathbf{P})$ ,  $a$  *Gillies covers*  $b$  in  $\mathbf{P}$  if  $\text{Margin}_{\mathbf{P}}(a, b) > 0$  and for all  $c \in X(\mathbf{P})$ , if  $\text{Margin}_{\mathbf{P}}(c, a) > 0$ , then  $\text{Margin}_{\mathbf{P}}(c, b) > 0$ . Then  $\text{UC}(\mathbf{P})$  is the set of candidates who are not Gillies covered in  $\mathbf{P}$ . This is called the Gillies Uncovered Set in [15].

**Ranked Pairs** [50]: For a profile  $\mathbf{P}$  and  $T \in \mathcal{L}(\{(x, y) \mid x \neq y \text{ and } \text{Margin}_{\mathbf{P}}(x, y) \geq 0\})$ , called the *tie-breaking ordering*, a pair  $(x, y)$  of candidates has a *higher priority* than a pair  $(x', y')$  of candidates according to  $T$  when either  $\text{Margin}_{\mathbf{P}}(x, y) > \text{Margin}_{\mathbf{P}}(x', y')$  or  $\text{Margin}_{\mathbf{P}}(x, y) = \text{Margin}_{\mathbf{P}}(x', y')$  and  $(x, y) T (x', y')$ . We construct a *Ranked Pairs ranking*  $\succ_{\mathbf{P}, T} \in \mathcal{L}(X)$  as follows:

1. Initialize  $\succ_{\mathbf{P}, T}$  to  $\emptyset$ .
2. If all pairs  $(x, y)$  with  $x \neq y$  and  $\text{Margin}_{\mathbf{P}}(x, y) \geq 0$  have been considered, then return  $\succ_{\mathbf{P}, T}$ . Otherwise let  $(a, b)$  be the pair with the highest priority among those with  $a \neq b$  and  $\text{Margin}_{\mathbf{P}}(a, b) \geq 0$  that have not been considered so far.
3. If  $\succ_{\mathbf{P}, T} \cup \{(a, b)\}$  is acyclic, then add  $(a, b)$  to  $\succ_{\mathbf{P}, T}$ ; otherwise, add  $(b, a)$  to  $\succ_{\mathbf{P}, T}$ . Go to step 2.

When the procedure terminates,  $\succ_{\mathbf{P}, T}$  is a linear order. The set  $\text{RP}(\mathbf{P})$  of Ranked Pairs winners is the set of all  $x \in X(\mathbf{P})$  such that  $x$  is the maximum of  $\succ_{\mathbf{P}, T}$  for some tie-breaking ordering  $T$ . This is the

<sup>8</sup>The **Simplified Bucklin** method selects as winners all  $k$ -th level majority winners. We ran our simulations with Simplified Bucklin as well as Bucklin; its frequency of violating positive involvement is similar to Bucklin but slightly worse.

<sup>9</sup>An equivalent definition of Copeland defines the score of  $a \in X(\mathbf{P})$  as the number of  $b \in X(\mathbf{P})$  with  $\text{Margin}_{\mathbf{P}}(a, b) > 0$  plus  $1/2$  times the number of  $b \in X(\mathbf{P})$  with  $\text{Margin}_{\mathbf{P}}(a, b) = 0$ . We also ran all of our simulations using the variant of Copeland known as Llull, where the score of  $a \in X(\mathbf{P})$  is the number of  $b \in X(\mathbf{P})$  with  $\text{Margin}_{\mathbf{P}}(a, b) > 0$  plus the number of  $b \in X(\mathbf{P})$  with  $\text{Margin}_{\mathbf{P}}(a, b) = 0$  (see [17]). Llull’s performance was very similar to that of Copeland.

<sup>10</sup>We also ran our simulations on the GOCHA method [46] that differs from Top Cycle only in profiles with margins of 0 between some candidates. Its performance was similar or slightly better than that of Top Cycle, depending on the parameters of the simulation.

“parallel universe” version of Ranked Pairs called RP in [8, 51], distinguished from other non-neutral, non-anonymous, or probabilistic versions of Ranked Pairs.

Since calculating  $RP(\mathbf{P})$  is an NP-complete problem [8], we also consider the non-anonymous version of Ranked Pairs proposed by Zavist and Tideman [53], which we call **Ranked Pairs ZT**. Zavist and Tideman propose to use a distinguish voter’s ranking to derive the tie-breaking ordering  $T$ . In particular, given  $i \in V(\mathbf{P})$ , let  $T(\mathbf{P}_i)$  be the lexicographic order on  $\{(x, y) \mid x \neq y \text{ and } Margin_{\mathbf{P}}(x, y) \geq 0\}$  derived from  $\mathbf{P}_i$ . Since different profiles have different sets of voters, we cannot use the same distinguished voter for all profiles. Given a linear order  $L$  of  $\mathcal{V}$ , for any profile  $\mathbf{P}$ , we define  $RPZT_L(\mathbf{P})$  to be the set of all  $x \in X(\mathbf{P})$  such that  $x$  is the maximum of  $\succ_{\mathbf{P}, T(\mathbf{P}_i)}$  where  $i$  is the minimal element of  $V(\mathbf{P})$  according to  $L$ .

**Beat Path** [45]: For  $a, b \in X(\mathbf{P})$ , a *path from  $a$  to  $b$  in  $\mathbf{P}$*  is a sequence  $\rho = x_1, \dots, x_n$  of distinct candidates in  $X(\mathbf{P})$  with  $x_1 = a$  and  $x_n = b$  such that for  $1 \leq k \leq n - 1$ ,  $Margin_{\mathbf{P}}(x_k, x_{k+1}) > 0$ . The *strength of  $\rho$*  is  $\min\{Margin_{\mathbf{P}}(x_k, x_{k+1}) \mid 1 \leq k \leq n - 1\}$ . Then  $a$  defeats  $b$  in  $\mathbf{P}$  according to Beat Path if the strength of the strongest path from  $a$  to  $b$  is greater than the strength of the strongest path from  $b$  to  $a$ .  $BP(\mathbf{P})$  is the set of undefeated candidates.

**Split Cycle** [29]: A *majority cycle in  $\mathbf{P}$*  is a sequence  $\rho = x_1, \dots, x_n$  of distinct candidates in  $X(\mathbf{P})$  except  $x_1 = x_n$  such that for  $1 \leq k \leq n - 1$ ,  $Margin_{\mathbf{P}}(x_k, x_{k+1}) > 0$ . The *strength of  $\rho$*  is defined as above for Beat Path. Then  $a$  defeats  $b$  in  $\mathbf{P}$  according to Split Cycle if  $Margin_{\mathbf{P}}(a, b)$  is positive and greater than the strength of the strongest majority cycle containing  $a$  and  $b$ .  $SC(\mathbf{P})$  is the set of undefeated candidates.

### 2.3 Positive Involvement

As explained in Section 1, our interest in this paper is the axiom of positive involvement.

**Definition 5.** A voting method  $F$  satisfies *positive involvement* (PI) if for any profile  $\mathbf{P}$  and  $x \in F(\mathbf{P})$ , if  $\mathbf{P}'$  is obtained from  $\mathbf{P}$  by adding a new voter who ranks  $x$  in first place, then  $x \in F(\mathbf{P}')$ .

PI is sometimes discussed in terms of the addition of several voters who rank  $x$  first. An obvious inductive argument shows that the coalitional version of PI follows from the single-voter version in Definition 5.

**Lemma 6.** If a voting method  $F$  satisfies PI, then it satisfies *coalitional* PI: for any profile  $\mathbf{P}$  and  $x \in F(\mathbf{P})$ , if  $\mathbf{P}'$  is a profile with  $X(\mathbf{P}) = X(\mathbf{P}')$ ,  $V(\mathbf{P}) \cap V(\mathbf{P}') = \emptyset$ , and every voter in  $\mathbf{P}'$  ranks  $x$  in first place, then  $x \in F(\mathbf{P} + \mathbf{P}')$ .

The voting methods in Section 2.2 can be classified by their satisfying or violating PI as follows.

**Proposition 7.** The voting methods Baldwin, Beat Path, Bucklin, Coombs, Copeland, Ranked Pairs, Strict Nanson, Top Cycle, and Uncovered Set all violate PI—see the Appendix for examples. All positional scoring rules, Instant Runoff, and Split Cycle satisfy PI (see [41, 29]).<sup>11</sup>

**Remark 8.** For results on violation or satisfaction of PI by other voting methods, see [41, 19]. Note that few known voting methods satisfy both PI and Condorcet consistency. Indeed, Pérez [41] observed that with the exception of the Minimax method [47, 33], which satisfies PI, “all the Condorcet correspondences that (to the best of our knowledge) are proposed in the literature” violate PI (p. 601). Other examples of Condorcet methods violating PI, besides those listed in Proposition 7 (Baldwin, Beat Path,

<sup>11</sup>Another well-known method satisfying PI is Plurality with Runoff [19, § 4.7.1]. This result depends on a specific way of breaking ties when more than two candidates qualify for the runoff (a candidate  $a$  qualifies for the runoff if either  $a$  has maximal plurality score or there is a unique candidate with maximal plurality score and  $a$  is among the candidates with the second highest plurality score). The version of Plurality with Runoff satisfying PI is the “parallel universe” version that considers all possible duels between candidates who qualify for the runoff; a candidate is a winner just in case they win in one of these duels. If instead all candidates who qualify for the runoff are promoted to a second round decided by Plurality, then the resulting version of Plurality with Runoff does not satisfy PI. Thanks to an anonymous referee for discussion of these points.

	profiles	profile-coalition pairs
absolute frequency	column 1 of Figure 2	column 1 of Figure 4
conditional on disagreement with $F_2$	columns 2-4 of Figure 2	columns 2-4 of Figure 4
relativized to voter potency	column 1 of Figure 3	column 1 of Figure 5
conditional on disagreement & relativized	columns 2-4 of Figure 3	columns 2-4 of Figure 5

Figure 1: ways of measuring PI violation

Copeland, Ranked Pairs, Strict Nanson, Top Cycle, and Uncovered Set), are Dodgson [13], Kemeny [32], and Young [52] (see [41, 19]).<sup>12</sup> However, Split Cycle is Condorcet consistent and satisfies PI.

### 3 Quantitative Analysis

We now turn to our discussion of different ways to quantify the extent to which a voting method  $F$  violates PI. They are summarized in Figure 1 and explained in the subsections to follow. The most obvious idea is to simply consider the probability that a randomly drawn profile  $\mathbf{P}$  (according to some probability model) *witnesses a violation of PI for  $F$* , meaning that there is some voter  $i$  in  $\mathbf{P}$  such that  $i$ 's favorite candidate wins in  $\mathbf{P}_{-i}$  according to  $F$  but not in  $\mathbf{P}$ . However, for medium to large sized electorates, we should expect this probability to be low for the mundane reason that the addition of any *single* voter will rarely cause someone to lose. We will explore several ways to deal with this issue.

One natural idea is to look at violations of *coalitional* PI: say that a profile  $\mathbf{P}$  *witnesses a violation of coalitional PI for  $F$*  if there is some coalition  $C$  of voters in  $\mathbf{P}$  with the same ranking (or at least the same top-ranked candidate), and their favorite candidate wins in  $\mathbf{P}_{-C}$  but not in  $\mathbf{P}$  according to  $F$ . Ideally, we would like to estimate the probability that a random profile witnesses a violation of coalitional PI (for various coalition sizes). Unfortunately, we find it too computationally expensive to check for every coalition of voters whether removing that coalition shows a violation of coalitional PI, especially for the thousands of profiles needed for reliable simulation results. To deal with this problem, in Section 3.2 we will assess the probability that a random *profile-coalition pair* witnesses a violation of PI.

However, we will begin in Section 3.1 with the probability that a random profile witnesses a violation of single-voter PI, since this provides a baseline with which to compare everything else that follows.

The probability model for profiles that will serve as our baseline is the Impartial Culture (IC) model: when sampling profiles with  $n$  candidates and  $m$  voters, the IC model gives every such profile equal probability. In Section 3.3, we consider several other probability models for sampling profiles. For each data point in our graphs, we sampled 25,000 profiles with the indicated even number  $m$  of voters and 25,000 profiles with the indicated odd number  $m + 1$  of voters, in order to have a mix of even and odd-sized electorates. We used the preflib [36] implementation for each of the probability models. Our code is in the online supplementary material at <https://github.com/epacuit/posinvolvement>.

<sup>12</sup>We did not include Dodgson, Kemeny, and Young in our simulations due to the computational complexity of determining winners for these methods (see [9] and [21, § 4.2]).

### 3.1 Random profiles

#### 3.1.1 Probability of PI violation

The leftmost column in Figure 2 shows, for several voting methods  $F$  that violate PI, the estimated probability that a random profile (according to IC) witnesses a violation of PI for  $F$ , as defined above. The important qualitative observations are that (i) the probability of PI violation increases as the number of candidates increases and (ii) the probability of PI violation decreases as the number of voters increases (above 20), for the mundane reason mentioned above—as the number of voters increases, any single voter has less influence in the election (a point to which we return in Section 3.1.3).

#### 3.1.2 Probability of PI violation conditional on voting method disagreement

Although the absolute frequency of PI violation in the leftmost column of Figure 2 is relevant, it is a mistake to think that if a voting method  $F_1$  has a low frequency of violating PI, then this undermines the use of PI as a reason to favor another voting method  $F_2$ , which satisfies PI, over  $F_1$ . First of all, when using PI to help choose between a method  $F_1$  that violates the axiom and a method  $F_2$  that satisfies it, we should ask: *in the profiles in which  $F_1$  and  $F_2$  disagree*, with what frequency does  $F_1$  violate PI?<sup>13</sup> That is, we should consider the following conditional probability, for a random profile  $\mathbf{P}$  (according to the given probability model, for a given number of candidates and voters):

$$Pr(\mathbf{P} \text{ witnesses a violation of PI for } F_1 \mid F_1(\mathbf{P}) \neq F_2(\mathbf{P})). \quad (1)$$

Columns 2-4 of Figure 2 show estimates of this conditional probability for several voting methods  $F_1$  that violate PI compared to three choices for  $F_2$ : Borda (2nd column), Instant Runoff (3rd column), and Split Cycle (4th column). We see a striking increase in the probability that  $F_1$  will violate PI when we condition on  $F_1$  disagreeing with the PI-satisfying method  $F_2$ . Note that if the probability of  $F_1$  violating PI is higher conditional on  $F_1$  disagreeing with  $F_2$  than it is conditional on  $F_1$  disagreeing with  $F'_2$ , then PI provides a stronger argument in the context of deciding between  $F_1$  and  $F_2$  than it does in the context of deciding between  $F_1$  and  $F'_2$ . E.g., let  $F_1$  be Baldwin,  $F_2$  be Split Cycle, and  $F'_2$  be Borda.

**Remark 9.** Since violation of (single voter) PI requires a close election in order for one voter to affect the outcome, we checked (i) to what extent the frequency of PI violation by Baldwin and Coombs depends on the manner of handling ties at intermediate stages of their elimination procedures and (ii) to what extent the frequency of PI violation by Uncovered Set depends on which of several variants of the Uncovered Set—which can disagree only when some candidates have a margin of 0—one chooses to analyze.

The tie-handling issue arises for Baldwin (resp. Coombs) in case there are multiple candidates with the smallest Borda score (resp. most last place votes) at a given stage of iteration. For each iterative elimination method  $F$ , there is a “parallel-universe tie-handling” (PUT) variant (cf. [24]): a candidate  $a$  wins under the PUT variant of  $F$  just in case there is some linear order  $L$  on the set of candidates such that  $a$  wins according to  $F_L$ , where  $F_L$  is defined in the same way as  $F$  except that if multiple candidates meet the criterion for elimination at some stage (e.g., have the smallest Borda score, or the most last place votes), only the  $L$ -minimal candidate among those candidates is eliminated. We found that the PUT versions of Baldwin and Coombs performed similarly or slightly better than the versions defined in

<sup>13</sup>We think that this methodology (which, as far as we know, is new) should be applied to the study of voting axioms in general, not only to PI.

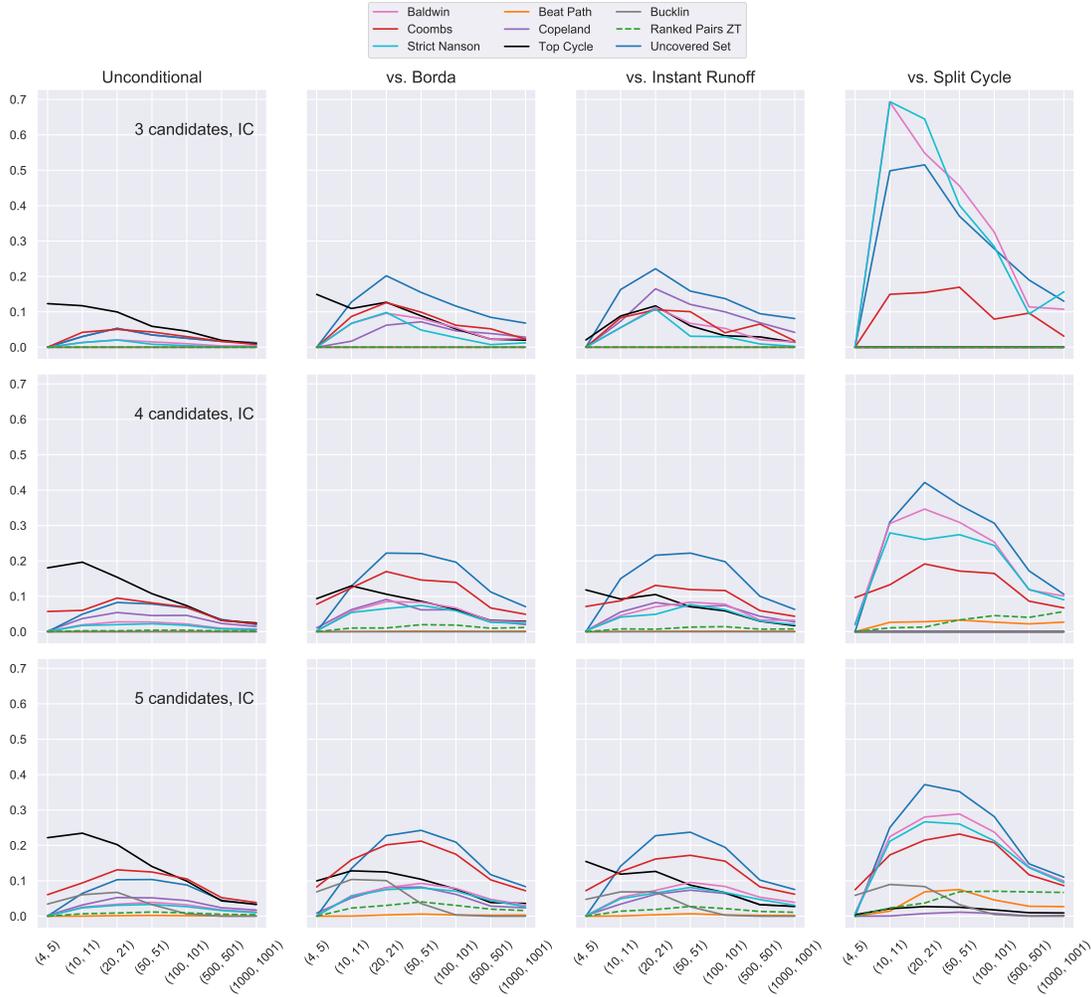


Figure 2: probability of a profile  $\mathbf{P}$  witnessing a violation of PI for  $F$  either unconditionally (far left), conditional on  $F$  disagreeing with Borda in  $\mathbf{P}$  (second from left), conditional on  $F$  disagreeing with Instant Runoff in  $\mathbf{P}$  (second from right), or conditional on  $F$  disagreeing with Split Cycle in  $\mathbf{P}$  (far right). Ranked Pairs was not included due to computational limitations (cf. [8]). Ranked Pairs ZT is shown with a dashed line to mark it out as our only non-anonymous voting method.

Section 2.2.<sup>14</sup> But given the computational difficulty of determining winners for the PUT versions, we use the versions defined in Section 2.2 for the rest of our analysis.

As for Uncovered Set, we considered three variants in addition to the Gillies Uncovered Set in Section 2.2: the Bordes Uncovered Set, the McKelvey Uncovered Set, and the Fishburn Uncovered Set.<sup>15</sup> Although the absolute frequencies of PI violation for these variants are similar (with Gillies and Fish-

<sup>14</sup>We also checked (for 3 and 4 candidates) the difference between conditioning on disagreement with Instant Runoff, as in Figure 2, and conditioning on disagreement with the PUT version of Instant Runoff. We found the results to be similar.

<sup>15</sup>Following Duggan [15], say that  $a$  Bordes covers  $b$  in  $\mathbf{P}$  if  $\text{Margin}_{\mathbf{P}}(a,b) > 0$  and for all  $c \in X(\mathbf{P})$ , if  $\text{Margin}_{\mathbf{P}}(c,a) \geq 0$ , then  $\text{Margin}_{\mathbf{P}}(c,b) \geq 0$ ; and say that  $a$  McKelvey covers  $b$  if a Gillies covers and Bordes covers  $b$ . For Fishburn’s [22] variant, say that  $a$  Fishburn covers  $b$  in  $\mathbf{P}$  if for all  $c \in X(\mathbf{P})$ , if  $\text{Margin}_{\mathbf{P}}(c,a) > 0$ , then  $\text{Margin}_{\mathbf{P}}(c,b) > 0$ , and there is a  $c \in X(\mathbf{P})$  such that  $\text{Margin}_{\mathbf{P}}(c,b) > 0$  and  $\text{Margin}_{\mathbf{P}}(c,a) \leq 0$ .

burn almost indistinguishable, and Bordes and McKelvey almost indistinguishable), the frequencies of PI violation conditional on disagreement with Borda, Instant Runoff, and Split Cycle are surprisingly different, with Gillies far worse than the others (see our online supplementary material). In this paper, we focus on the Gillies variant for a worst-case analysis of the family of Uncovered Set variants.

### 3.1.3 Probability of PI violation relativized to voter potency

Although the results of conditioning on voting method disagreement are striking, even those conditional probabilities are bound to decline as we increase the number of voters, given a single voter’s declining influence. What we should be asking, we think, is how likely a PI violation is compared to how likely it is that any single voter causes a candidate to lose. For this, we introduce the notion of voter *potency*.

**Definition 10.** For any profile  $\mathbf{P}$  and  $i \in V(\mathbf{P})$ , we say that  $i$  is *potent* in  $\mathbf{P}$  if  $F(\mathbf{P}_{-i}) \not\subseteq F(\mathbf{P})$ .<sup>16</sup>

Then our proposal is to measure how badly a voting method violates single-voter PI by the ratio

$$\frac{\Pr(\text{there is a voter } i \text{ triggering a PI violation, i.e., } \max(\mathbf{P}_i) \in F_1(\mathbf{P}_{-i}) \setminus F_1(\mathbf{P}))}{\Pr(\text{there is a voter } i \text{ who is potent in } \mathbf{P})}, \quad (2)$$

possibly with these probabilities conditioned on  $F_1$  disagreeing in  $\mathbf{P}$  with an  $F_2$  that satisfies PI. For voting methods violating PI, the numerator may be small (as opposed to 0 for methods that satisfy PI), but the denominator is also small; thus, the ratio may be surprisingly large. The results are shown in Figure 3.

**Remark 11.** One curious feature of the last column of Figure 3 is that for 3 or 4 candidates, the probability of Copeland and Top Cycle violating PI conditional on their disagreeing with Split Cycle in  $\mathbf{P}$  goes *down to zero*. This is initially puzzling: since Split Cycle satisfies PI, mustn’t Copeland and Top Cycle disagree with Split Cycle whenever they violate PI? The answer is ‘yes’, but they may disagree with Split Cycle in the *smaller profile*  $\mathbf{P}_{-i}$  rather than in  $\mathbf{P}$ ; indeed, this always happens when the cited methods violate single-voter PI for 3 or 4 candidates. Thus, there is an ambiguity in the idea of “conditioning on disagreement with  $F_2$ ”—we could condition on disagreement in  $\mathbf{P}_{-i}$ , in  $\mathbf{P}$ , in both, or in one or the other. So far we have only conditioned on disagreement in  $\mathbf{P}$ , since so far in our random sampling we only draw a profile  $\mathbf{P}$ ; but in Section 3.2.2, when we randomly sample profile-voter pairs, we will be able to conveniently condition on disagreement in  $\mathbf{P}_{-i}$  or  $\mathbf{P}$ .

## 3.2 Random profile-coalition pairs

Given the computational difficulty of checking whether a profile witnesses a violation of coalitional PI, we now turn to randomly sampling *profile-coalition pairs*, i.e., pairs  $(\mathbf{P}, \mathbf{P}_C)$  of a random profile  $\mathbf{P}$  (for a given number of candidates and voters) and a “coalitional” profile  $\mathbf{P}_C$  obtained by randomly selecting a single ranking and then assigning it to all voters in  $C$  (with  $C \cap V(\mathbf{P}) = \emptyset$ ). Such a profile-coalition pair *witnesses a violation of PI* just in case the unanimously top-ranked candidate in  $\mathbf{P}_C$  is a winner in  $\mathbf{P}$  but not in  $\mathbf{P} + \mathbf{P}_C$ . Note that even if the pair  $(\mathbf{P}, \mathbf{P}_C)$  does not witness a violation of PI, the profile  $\mathbf{P} + \mathbf{P}_C$  may witness a violation of coalitional PI by the removal of a different coalition  $C'$ . Thus, this approach misses violations of coalitional PI involving other coalitions. The probability that a random profile-coalition pair witnesses a violation of PI may be considerably lower than the probability that a random profile witnesses a violation of coalitional PI. Nonetheless, there are benefits of this approach, such as (i) our being able to feasibly study coalitions of more than one voter, (ii) our being able to search profiles up to 5,000 voters or up to 10 candidates, and (iii) our being able to conveniently condition on voting method disagreement before or after the new coalition of voters joins the election (see Remark 11).

<sup>16</sup>Note that  $i$  being potent in  $\mathbf{P}$  is stronger than  $i$  being *pivotal* in  $\mathbf{P}$  in the sense that  $F(\mathbf{P}_{-i}) \neq F(\mathbf{P})$ .

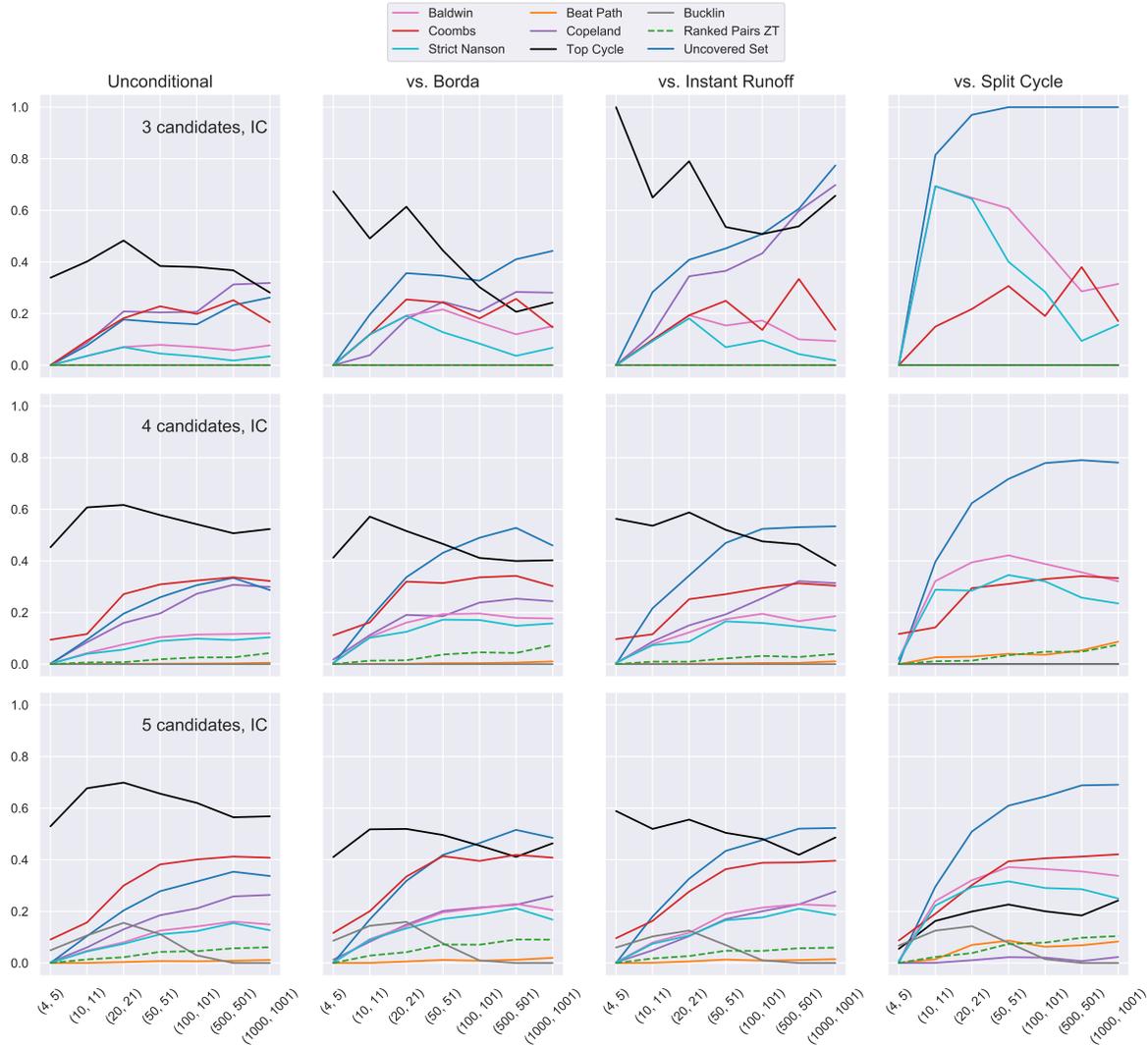


Figure 3: probability of a profile  $\mathbf{P}$  witnessing a violation of PI for  $F$ , divided by the probability of  $\mathbf{P}$  having a potent voter, either unconditionally (far left), conditional on  $F$  disagreeing with Borda in  $\mathbf{P}$  (second from left), conditional on  $F$  disagreeing with Instant Runoff in  $\mathbf{P}$  (second from right), or conditional on  $F$  disagreeing with Split Cycle in  $\mathbf{P}$  (far right).

### 3.2.1 Probability of PI violation

The leftmost column of Figure 4 shows estimates for the probability that a randomly selected profile-coalition pair  $(\mathbf{P}, \mathbf{P}_C)$  witnesses a violation of PI, depending on whether  $C$  is a coalition of a single voter (first row), a coalition of 0.25% of the total voter size (second row), a coalition of 0.5% (third row), or a coalition of 0.75% (fourth row). We show only the results for 500, 1,000, and 5,000 voter profiles, since with 100 voters or fewer, all the coalition sizes round to 1 voter. As expected, in the single voter case, the probability that a random profile-voter pair witnesses a violation of PI is much lower than the probability that a random profile witnesses a violation of single-voter PI (compare the leftmost column and third row of Figure 2). Indeed, the difference is roughly an order of magnitude. The probability does increase for

larger coalition sizes, but the probability of a random profile-0.5%-coalition pair witnessing a violation of PI is still less than half the probability of a random profile  $\mathbf{P}$  witnessing a violation of single-voter PI.

### 3.2.2 Probability of PI violation conditional on disagreement

As suggested by earlier results (Section 3.1.2), conditioning on the probability that in our random profile-coalition pair  $(\mathbf{P}, \mathbf{P}_C)$  the voting method  $F_1$  disagrees with the method  $F_2$  that satisfies PI—where disagreement now means that  $F_1(\mathbf{P}) \neq F_2(\mathbf{P})$  or  $F_1(\mathbf{P} + \mathbf{P}_C) \neq F_2(\mathbf{P} + \mathbf{P}_C)$  (recall Remark 11)—dramatically increases the probability that  $F_1$  violates PI. The results are shown in columns 2-4 of Figure 4.

### 3.2.3 Probability of PI violation relativized to voter potency

We can also apply the idea of relativizing to voter potency from Section 3.1.3 in the paradigm of sampling profile-coalition pairs. In particular, we want to estimate the value of the following ratio for a random profile-coalition pair  $(\mathbf{P}, \mathbf{P}_C)$ , a randomly chosen  $a \in F(\mathbf{P})$ , and a random new voter:

$$\frac{\Pr(\text{the new voters in } C \text{ cause } a \text{ to lose} \mid a \text{ is the favorite of the new voters in } C)}{\Pr(\text{the new voters in } C \text{ cause } a \text{ to lose})}. \quad (3)$$

Intuitively, the numerator should be 0, but as we know, for voting methods that violate PI it is not zero. It is small, but on the other hand, the denominator is also small, since small coalitions of voters have limited influence in any election. Thus, as in Section 3.1.3, the ratio itself can be surprisingly large. Ratios of the form  $\Pr(A \mid B)/\Pr(A)$ , as in (3), are used in confirmation theory to measure the degree of support that evidence provides for a hypothesis (see, e.g., [30, p. 54]). Note that  $\Pr(A \mid B)/\Pr(A) = \Pr(B \mid A)/\Pr(B)$ . Thus, we can phrase our question in one of two ways: If you learn that the new voters in  $C$  rank  $a$  first, to what extent does this support the hypothesis that the voters in  $C$  will cause  $a$  to lose? Or, alternatively, if you learn that the new voters in  $C$  caused  $a$  to lose, to what extent does this support the hypothesis that the voters in  $C$  ranked  $a$  first? Of course for methods satisfying PI the answer is “not at all.”

Results are shown in Figure 5, where columns 2-4 also condition the probabilities in (3) on  $F_1$  disagreeing with  $F_2$  in  $\mathbf{P}$  or  $\mathbf{P} + \mathbf{P}_C$ . Note the strikingly high ratios for some methods—especially Coombs and Top Cycle, but also Uncovered Set, Baldwin, Nanson, and Copeland—as we increase the candidates.

## 3.3 Other probability models

In addition to sampling profiles with the IC model, we tried several other probability models. In the Pólya-Eggenberger urn model [2], each voter in turn randomly draws a linear order from an urn. Initially the urn is  $\mathcal{L}(X)$ . If a voter randomly chooses  $L$  from the urn, we return  $L$  to the urn plus  $\alpha \in \mathbb{N}$  copies of  $L$ . IC is the special case where  $\alpha = 0$ . The Impartial Anonymous Culture (IAC) is the special case where  $\alpha = 1$ . We also considered  $\alpha = 10$  (as in [7, 6, 5]) for the model we call URN. In the Mallows’s model (see [34, 35]), given a reference ranking  $L_0 \in \mathcal{L}(X)$  and  $\phi \in (0, 1]$ , the probability that a voter’s ballot is  $L \in \mathcal{L}(X)$  is  $\Pr_{L_0}(L) = \phi^{\tau(L, L_0)}/C$  where  $\tau(L, L_0) = \binom{|X|}{2} - |L \cap L_0|$ , the Kendall-tau distance of  $L$  to  $L_0$ , and  $C$  is a normalization constant. We considered two reference rankings,  $L_0$  and its converse  $L_0^{-1}$  (e.g.,  $L_0$  ranks candidates from more liberal to more conservative, and  $L_0^{-1}$  vice versa), in which case the probability that a voter’s ballot is  $L$  is  $\frac{1}{2}\Pr_{L_0}(L) + \frac{1}{2}\Pr_{L_0^{-1}}(L)$ . We set  $\phi = 0.8$  (as in [6, 5]).

The most important finding concerning the different probability models is that as we deviate from the IC model, the main phenomena seen in the previous simulations do not disappear. Figure 6 shows the 5-candidate case of Figure 5 under these three probability models, plus IC again (for easy comparison). We see that the surprisingly high values for the ratio measure in (3) are not artifacts of the IC model.

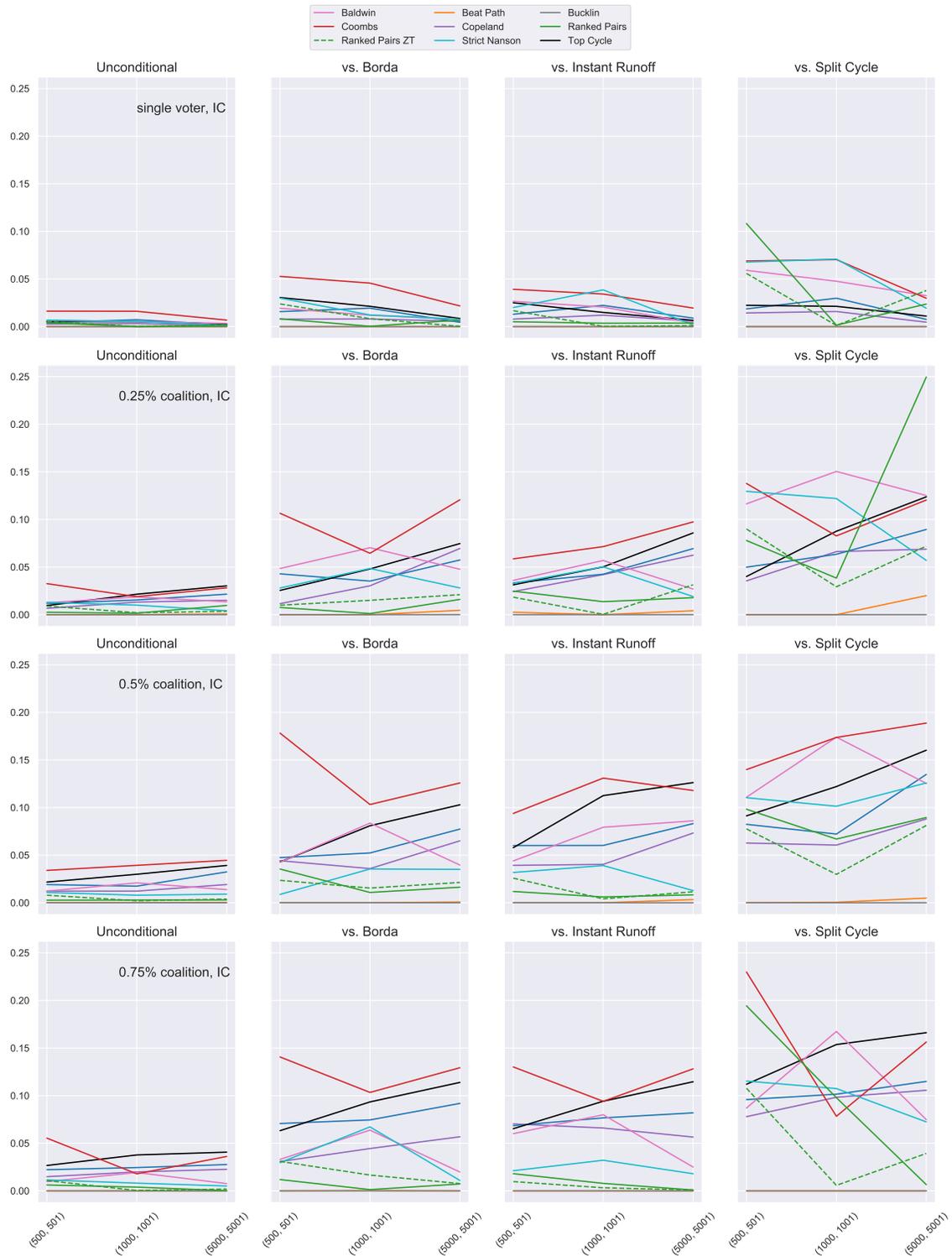


Figure 4: probability of a profile-coalition pair  $(\mathbf{P}, \mathbf{P}_C)$  witnessing a violation of PI either unconditionally (far left), conditional on disagreement in  $\mathbf{P}$  or  $\mathbf{P} + \mathbf{P}_C$  with Borda (second from left), with Instant Runoff (second from right), or with Split Cycle (far right). The first row is for a single voter coalition, the second for coalitions of 0.25% of initial total voter size, the third for 0.5%, and the fourth for 0.75%.

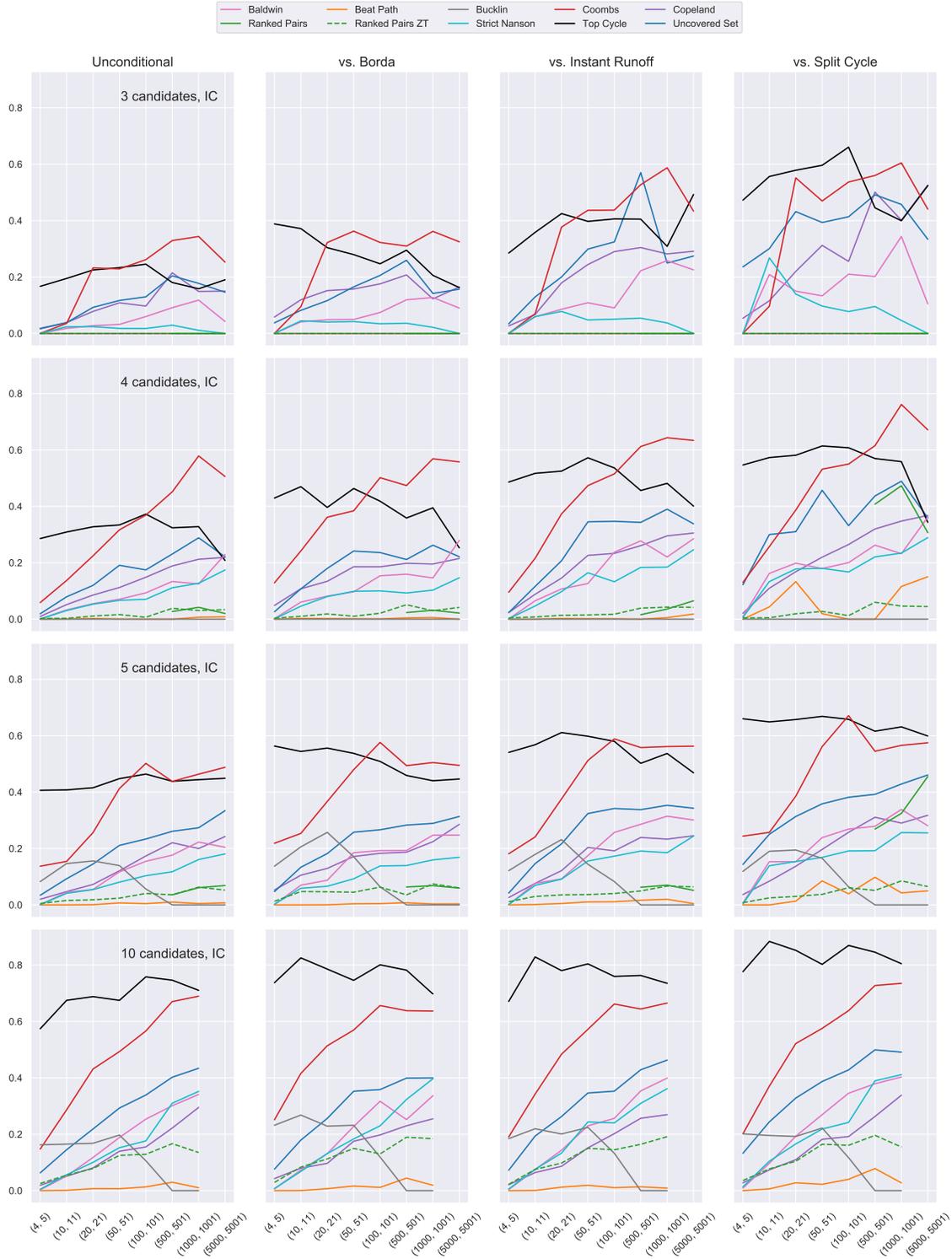


Figure 5: the probability ratio in (3) from Section 3.2.3 for profile-coalition pairs  $(\mathbf{P}, \mathbf{P}_C)$  with  $|C| = 1$ , either unconditionally (far left), conditional on disagreement with Borda in  $\mathbf{P}$  or  $\mathbf{P} + \mathbf{P}_C$  (second from left), conditional on disagreement with Instant Runoff in  $\mathbf{P}$  or  $\mathbf{P} + \mathbf{P}_C$ , or conditional on disagreement with Split Cycle in  $\mathbf{P}$  or  $\mathbf{P} + \mathbf{P}_C$  (far right). Due to computational limitations, results for 10 candidates and 5,000/5,001 voters were not obtained, and results for Ranked Pairs were obtained only for 3-5 candidates and 500-5,001 voters (as increasing the number of voters decreases the likelihood of tied margins).

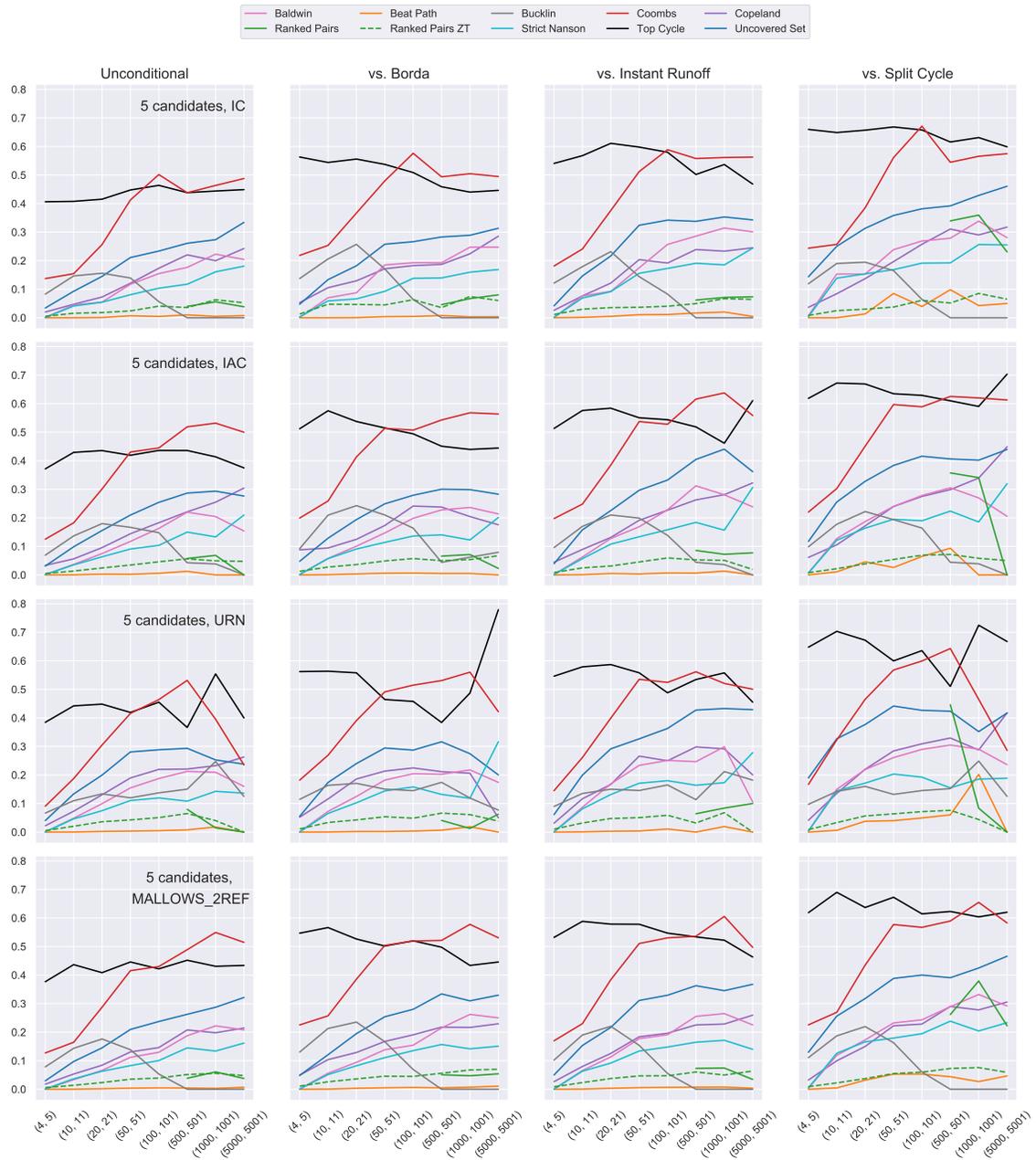


Figure 6: the same measures as in row 3 of Figure 5 but under different probability models. As before, due to computational limitations, results for Ranked Pairs were obtained only for 500-5,001 voters.

## 4 Conclusion

In their book on failures of monotonicity-type axioms for voting methods, Nurmi and Felsenthal [19] argue that violations of positive (and negative) involvement are the “more dramatic types of no show paradox,” wherein “the issues of legitimacy of outcomes and questionable voter incentives are far more obvious” (p. 86). In this paper, we proposed ways of measuring the extent to which these dramatic violations occur. The next step in future work is to marshal these measures when arguing in favor of some voting methods over others. Under certain voting methods that violate PI, much of a voter’s potency is turned against them—in particular, against their desire to see their favorite candidate elected. It is like adding insult to injury to be told that not only does your vote have little chance of influencing the election, but also your vote may cause your *favorite* candidate to lose, and *the probability of your vote doing so is not at all insignificant compared to the probability of it causing any candidate to lose*. The probabilities may be large enough to raise real concerns in small elections in committees and clubs, unless one somehow knows in advance that none of the probability models used here is relevant for the committee (e.g., one knows in advance that the committee has single-peaked preferences). But for large elections, defenders of voting methods that violate PI may simply respond: don’t worry—your vote probably won’t cause your favorite to lose, because it probably won’t influence the election at all. The same response could be used to try to dismiss concerns about violations of other single-voter axioms, such as monotonicity or single-voter strategyproofness. Such responses raise what is perhaps the ultimate paradox of voting: why do voters vote in large elections? The significance of single-voter axioms for large elections may turn on the answer to that question.

## A Appendix

The following are examples of profiles witnessing PI violation for the voting methods from Proposition 7. In the profiles, the new voter’s ranking is highlighted in grey.

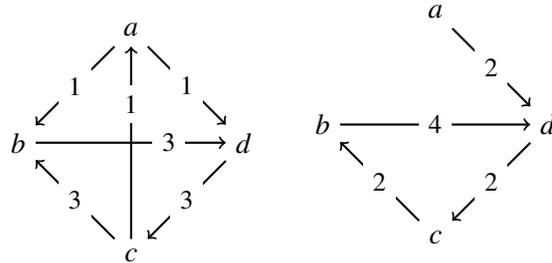
**Example 12** (Baldwin violates PI). Candidate  $c$  is the winner on the left (the order of elimination is  $a, d, b$ ) but not on the right (the order of elimination is  $d, b, c$ ):

1	2	1	1	1	1	1	2	1	1	1	1
$b$	$c$	$d$	$d$	$a$	$b$	$c$	$d$	$d$	$a$	$b$	$c$
$a$	$b$	$a$	$b$	$c$	$a$	$b$	$a$	$b$	$c$	$a$	$a$
$d$	$d$	$c$	$a$	$b$	$d$	$d$	$c$	$a$	$b$	$b$	$b$
$c$	$a$	$b$	$c$	$d$	$c$	$a$	$b$	$c$	$d$	$d$	$d$

Baldwin winners:  $\{c\}$       Baldwin winners:  $\{a\}$

Since there are no ties for the lowest Borda score at any stage, the same holds for Baldwin PUT.

**Example 13** (Beat Path violates PI). The left graph below can be realized as the margin graph of a profile with 11 voters; the right graph is obtained by adding a voter with the ranking  $b a c d$ . Then  $b$  is a winner on the left but not on the right:



Beat Path:  $\{a, b, c, d\}$

Beat Path:  $\{a\}$   
New ranking:  $b a c d$

**Example 14** (Bucklin violates PI). Candidate  $c$  is the winner on the left (there are no 1st or 2nd level majority winners;  $a$  and  $c$  are 3rd level majority winners; and more voters rank  $c$  in 3rd place or higher than  $a$ ) but not on the right (because  $e$  is the unique 2nd level majority winner):

1	1	1	1
$a$	$a$	$b$	$c$
$b$	$e$	$e$	$d$
$c$	$c$	$c$	$a$
$e$	$b$	$d$	$b$
$d$	$d$	$a$	$e$

1	1	1	1	1
$a$	$a$	$b$	$c$	$c$
$b$	$e$	$e$	$d$	$e$
$c$	$c$	$c$	$a$	$b$
$e$	$b$	$d$	$b$	$d$
$d$	$d$	$a$	$e$	$a$

Bucklin:  $\{c\}$

Bucklin:  $\{e\}$

The same example works for Simplified Bucklin, as defined in Footnote 8, only the winners on the left are  $a$  and  $c$ , and the winner on the right is  $e$ .

**Example 15** (Coombs violates PI). Candidate  $a$  is the winner on the left (because  $b$  and  $c$  are eliminated in the first round, and then  $a$  is the majority winner) but not on the right (because  $b$  is eliminated first, and then  $c$  is the majority winner):

1	2	1	1	1
$a$	$c$	$a$	$b$	$c$
$b$	$a$	$d$	$c$	$b$
$d$	$d$	$b$	$d$	$a$
$c$	$b$	$c$	$a$	$d$

1	2	1	1	1	1
$a$	$c$	$a$	$b$	$c$	$a$
$b$	$a$	$d$	$c$	$b$	$d$
$d$	$d$	$b$	$d$	$a$	$c$
$c$	$b$	$c$	$a$	$d$	$b$

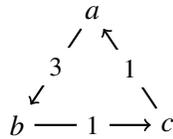
Coombs:  $\{a\}$

Coombs:  $\{c\}$

Under Coombs PUT, on the left, since  $b$  and  $c$  tie for most last place votes, we consider two cases: (i) if we eliminate  $b$  in the first round, then  $c$  is the majority winner; (ii) if we eliminate  $c$  in the first round, then  $a$  is the majority winner. Hence both  $a$  and  $c$  win in the profile on the left. In the profile on the right,  $c$  is the unique winner for Coombs PUT for the same reason as for Coombs.

**Example 16** (Copeland and Uncovered Set violate PI). Candidate  $b$  is a Copeland winner on the left but not on the right:<sup>17</sup>

2	1	2
$a$	$b$	$c$
$b$	$c$	$a$
$c$	$a$	$b$

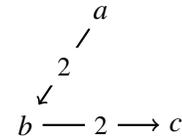


Copeland winners:  $\{a, b, c\}$

Gillies Uncovered Set winners:  $\{a, b, c\}$

Fishburn Uncovered Set winners:  $\{a, b, c\}$

2	1	2	1
$a$	$b$	$c$	$b$
$b$	$c$	$a$	$a$
$c$	$a$	$b$	$c$



Copeland winners:  $\{a\}$

Gillies Uncovered Set winners:  $\{a, c\}$

Fishburn Uncovered Set winners:  $\{a\}$

The same example works for the Gillies (resp. Fishburn) version of Uncovered Set (recall Remark 9): on the left, the winners are  $a$ ,  $b$ , and  $c$ , while on the right, the winners are  $a$  and  $c$  (resp.  $a$ ).

**Example 17** (Strict Nanson violates PI). Candidate  $c$  is the winner on the left (the order of elimination is  $b$ ,  $d$ ,  $a$ ) but not on the right (both  $a$  and  $b$  are eliminated in the first round, followed by  $c$ ):

1	1	3	3	1	1
$a$	$d$	$c$	$a$	$d$	$d$
$c$	$b$	$b$	$d$	$c$	$c$
$d$	$c$	$a$	$c$	$a$	$b$
$b$	$a$	$d$	$b$	$b$	$a$

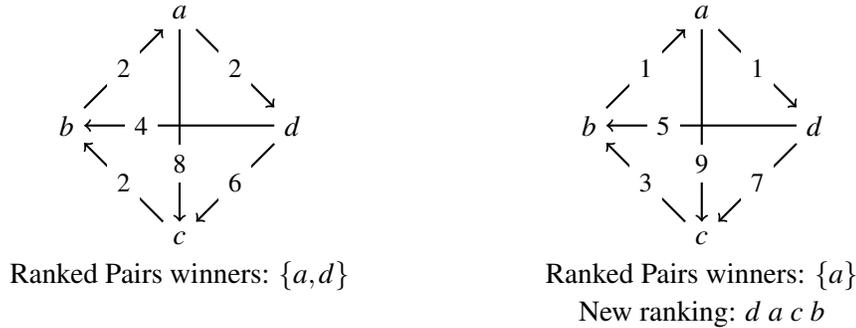
Strict Nanson winners:  $\{c\}$

1	1	3	3	1	1	1
$a$	$d$	$c$	$a$	$d$	$d$	$c$
$c$	$b$	$b$	$d$	$c$	$c$	$d$
$d$	$c$	$a$	$c$	$a$	$b$	$b$
$b$	$a$	$d$	$b$	$b$	$a$	$a$

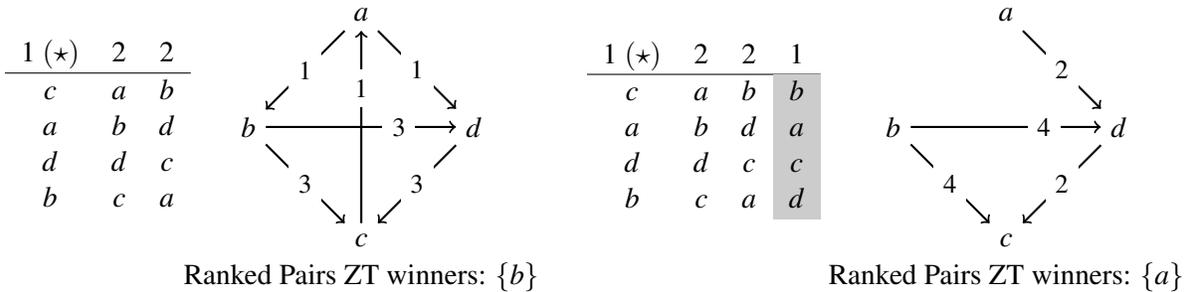
Strict Nanson winners:  $\{d\}$

<sup>17</sup>The same example provides a PI violation for the Llull method defined in Footnote 9. The Llull winners are the same as the Copeland winners in both profiles.

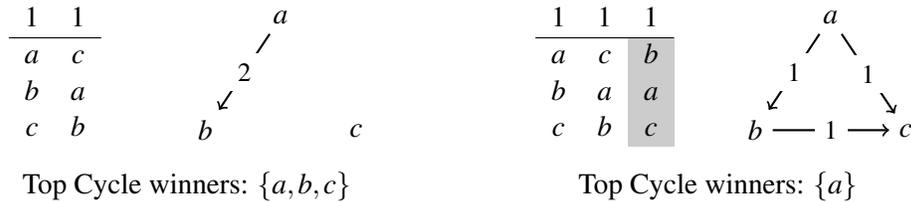
**Example 18** (Ranked Pairs violates PI). The left graph below can be realized as the margin graph of a profile with 20 voters; then the right graph below is obtained by adding a voter with the ranking  $d a b c$ . Then candidate  $d$  is a winner on the left (we first lock in the edges  $(a, c)$ ,  $(d, c)$ , and  $(d, b)$ , and then there is a choice of whether to prioritize the  $(b, a)$  edge, in which case  $d$  wins, or the  $(a, d)$  edge, in which case  $a$  wins) but not on the right (we first lock in the edges  $(a, c)$ ,  $(d, c)$ , and  $(d, b)$ , and then there is a choice of whether to prioritize  $(c, b)$  or  $(a, d)$ , but in either case  $(a, d)$  eventually gets locked in):



For Ranked Pairs ZT, in the profile on the left below, where the tiebreaking voter's ranking is indicated with  $(\star)$ ,  $b$  is the winner (we lock in  $(d, c)$ ,  $(b, c)$ ,  $(b, d)$ , and then  $(c, a)$ ). But on the right,  $a$  is the winner (there are no cycles, so all edges are locked in, and then the ranking  $(\star)$  ranks  $a$  above  $b$ ).



**Example 19** (Top Cycle violates PI). Candidate  $b$  is a winner on the left but not the right:



**References**

[1] J. M. Baldwin (1926): *A technique of the Nanson preferential majority system of election*. *Transactions and Proceedings of the Royal Society of Victoria* 39, pp. 45–52.  
 [2] Sven Berg (1985): *Paradox of voting under an urn model: The effect of homogeneity*. *Public Choice* 47, pp. 377–387, doi:10.1007/BF00127533.  
 [3] Felix Brandt, Vincent Conitzer & Ulle Endriss (2013): *Computational Social Choice*. In Gerhard Weiss, editor: *Multiagent Systems*, MIT Press, Cambridge, Mass., pp. 213–283.

- [4] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang & Ariel D. Procaccia, editors (2016): *Handbook of Computational Social Choice*. Cambridge University Press, New York, doi:10.1017/cbo9781107446984.003.
- [5] Felix Brandt, Christian Geist & Martin Strobel (2020): *Analyzing the Practical Relevance of the Condorcet Loser Paradox and the Agenda Contraction Paradox*. In M. Diss & V. Merlin, editors: *Evaluating Voting Systems with Probability Models: Essays by and in Honor of William Gehrlein and Dominique Lepelley*, Springer, Berlin, pp. 97–115, doi:10.1007/978-3-030-48598-6\_5.
- [6] Felix Brandt, Johannes Hofbauer & Martin Strobel (2019): *Exploring the No-Show Paradox for Condorcet Extensions Using Ehrhart Theory and Computer Simulations*. In: *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '19)*, International Foundation for Autonomous Agents and MultiAgent Systems, pp. 520–528.
- [7] Felix Brandt & Hans Georg Seedig (2014): *On the Discriminative Power of Tournament Solutions*. In M. Lübbecke, A. Koster, P. Letmathe, R. Madlener, B. Peis & G. Walther, editors: *Operations Research Proceedings 2014*, Springer, Cham, pp. 53–58, doi:10.1007/978-3-319-28697-6\_8.
- [8] Markus Brill & Felix Fischer (2012): *The Price of Neutrality for the Ranked Pairs Method*. In: *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12)*, AAAI Press, pp. 1299–1305.
- [9] Ioannis Caragiannis, Edith Hemaspaandra & Lane A. Hemaspaandra (2016): *Dodgson's Rule and Young's Rule*. In Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang & Ariel D. Procaccia, editors: *Handbook of Computational Social Choice*, Cambridge University Press, New York, pp. 103–126, doi:10.1017/cbo9781107446984.005.
- [10] Vincent Conitzer & Toby Walsh (2016): *Barriers to Manipulation in Voting*. In Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang & Ariel D. Procaccia, editors: *Handbook of Computational Social Choice*, Cambridge University Press, New York, pp. 127–145, doi:10.1017/cbo9781107446984.006.
- [11] Clyde Hamilton Coombs (1964): *A Theory of Data*. John Wiley and Sons, New York.
- [12] A. H. Copeland (1951): *A 'reasonable' social welfare function*. Notes from a seminar on applications of mathematics to the social sciences, University of Michigan.
- [13] Charles L. Dodgson (1995): *A Method of Taking Votes on More Than Two Issues*. In Iain McLean & Arnold Urken, editors: *Classics of Social Choice*, University of Michigan Press, Ann Arbor, pp. 288–298.
- [14] Keith Dowding & Martin Van Hees (2008): *In Praise of Manipulation*. *British Journal of Political Science* 38(1), pp. 1–15, doi:10.1017/S000712340800001X.
- [15] John Duggan (2013): *Uncovered Sets*. *Social Choice and Welfare* 41(3), pp. 489–535, doi:10.1007/s00355-012-0696-9.
- [16] Ulle Endriss, editor (2017): *Trends in Computational Social Choice*. AI Access.
- [17] Piotr Faliszewski, Edith Hemaspaandra, Lane A. Hemaspaandra & Jörg Rothe (2007): *Llull and Copeland Voting Broadly Resist Bribery and Control*. In: *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence (AAAI-07)*, AAAI Press, pp. 724–730.
- [18] Dan S. Felsenthal & Hannu Nurmi (2016): *Two types of participation failure under nine voting methods in variable electorates*. *Public Choice* 168, pp. 115–135, doi:10.1007/s11127-016-0352-5.
- [19] Dan S. Felsenthal & Hannu Nurmi (2017): *Monotonicity failures afflicting procedures for electing a single candidate*. Springer, Cham, doi:10.1007/978-3-319-51061-3.
- [20] Dan S. Felsenthal & Nicolaus Tideman (2013): *Varieties of failure of monotonicity and participation under five voting methods*. *Theory and Decision* 75, pp. 59–77, doi:10.1007/s11238-012-9306-7.
- [21] Felix Fischer, Olivier Hudry & Rolf Niedermeier (2016): *Weighted Tournament Solutions*. In Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang & Ariel D. Procaccia, editors: *Handbook of Computational Social Choice*, Cambridge University Press, New York, pp. 85–102, doi:10.1017/cbo9781107446984.004.
- [22] Peter C. Fishburn (1977): *Condorcet Social Choice Functions*. *SIAM Journal on Applied Mathematics* 33(3), pp. 469–489, doi:10.1137/0133030.

- [23] Peter C. Fishburn & Steven J. Brams (1983): *Paradoxes of Preferential Voting*. *Mathematics Magazine* 56(4), pp. 207–214, doi:10.2307/2689808.
- [24] Rupert Freeman, Markus Brill & Vincent Conitzer (2015): *General Tiebreaking Schemes for Computational Social Choice*. In: *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015)*, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, pp. 1401–1409.
- [25] William V. Gehrlein & Dominique Lepelley (2017): *Elections, Voting Rules and Paradoxical Outcomes*. Springer, Cham, doi:10.1007/978-3-319-64659-6.
- [26] Donald B. Gillies (1959): *Solutions to general non-zero-sum games*. In A. W. Tucker & R. D. Luce, editors: *Contributions to the Theory of Games*, Princeton University Press, Princeton, New Jersey, doi:10.1515/9781400882168-005.
- [27] Bernard Grofman & Scott L. Feld (2004): *If you like the alternative vote (a.k.a. the instant runoff), then you ought to know about the Coombs rule*. *Electoral Studies* 23(4), pp. 641–659, doi:10.1016/j.electstud.2003.08.001.
- [28] Clarence Hoag & George Hallett (1926): *Proportional Representation*. Macmillan, New York.
- [29] Wesley H. Holliday & Eric Pacuit (2020): *Split Cycle: A New Condorcet Consistent Voting Method Independent of Clones and Immune to Spoilers*. ArXiv:2004.02350.
- [30] Paul Horwich (2016): *Probability and Evidence*. Cambridge University Press, Cambridge, doi:10.1017/CBO9781316494219.
- [31] José L. Jimeno, Joaquín Pérez & Estefanía García (2009): *An extension of the Moulin No Show Paradox for voting correspondences*. *Social Choice and Welfare* 33(3), pp. 343–359, doi:10.1007/s00355-008-0360-6.
- [32] John G. Kemeny (1959): *Mathematics without Numbers*. *Daedalus* 88(4), pp. 577–591.
- [33] Gerald H. Kramer (1977): *A dynamical model of political equilibrium*. *Journal of Economic Theory* 16(2), pp. 310–334, doi:10.1016/0022-0531(77)90011-4.
- [34] C. L. Mallows (1957): *Non-Null Ranking Models. I*. *Biometrika* 44(2), pp. 114–130, doi:10.2307/2333244.
- [35] John Marden (1995): *Analyzing and Modeling Rank Data*. CRC Press, New York, doi:10.1201/b16552.
- [36] Nicholas Mattei & Toby Walsh (2013): *PrefLib: A Library of Preference Data*. In: *Proceedings of Third International Conference on Algorithmic Decision Theory*, Springer, pp. 259–270, doi:10.1007/978-3-642-41575-3\_20.
- [37] Hervé Moulin (1988): *Condorcet's Principle Implies the No Show Paradox*. *Journal of Economic Theory* 45(1), pp. 53–64, doi:10.1016/0022-0531(88)90253-0.
- [38] E. J. Nanson (1882): *Methods of election*. *Transactions and Proceedings of the Royal Society of Victoria* 19, pp. 197–240.
- [39] Nina Narodytska, Toby Walsh & Lirong Xia (2011): *Manipulation of Nanson's and Baldwin's Rules*. In: *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, AAAI Press, pp. 713–718.
- [40] Emerson M. S. Niou (1987): *A Note on Nanson's Rule*. *Public Choice* 54(2), pp. 191–193, doi:10.1007/BF00123006.
- [41] Joaquín Pérez (2001): *The Strong No Show Paradoxes are a common flaw in Condorcet voting correspondences*. *Social Choice and Welfare* 18(3), pp. 601–616, doi:10.1007/s003550000079.
- [42] Florenz Plassmann & T. Nicolaus Tideman (2014): *How frequently do different voting rules encounter voting paradoxes in three-candidate elections?* *Social Choice and Welfare* 42, pp. 31–75, doi:10.1007/s00355-013-0720-8.
- [43] Donald G. Saari (1995): *Basic Geometry of Voting*. Springer, Berlin, doi:10.1007/978-3-642-57748-2.
- [44] M. Remzi Sanver & William S. Zwicker (2012): *Monotonicity properties and their adaptation to irresolute social choice rules*. *Social Choice and Welfare* 39(2/3), pp. 371–398, doi:10.1007/s00355-012-0654-6.

- [45] Markus Schulze (2011): *A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method*. *Social Choice and Welfare* 36, pp. 267–303, doi:10.1007/s00355-010-0475-4.
- [46] Thomas Schwartz (1986): *The Logic of Collective Choice*. Columbia University Press, New York, doi:10.7312/schw93758.
- [47] Paul B. Simpson (1969): *On Defining Areas of Voter Choice: Professor Tullock on Stable Voting*. *The Quarterly Journal of Economics* 83(3), pp. 478–490, doi:10.2307/1880533.
- [48] John H. Smith (1973): *Aggregation of Preferences with Variable Electorate*. *Econometrica* 41(6), pp. 1027–1041, doi:10.2307/1914033.
- [49] Alan D. Taylor & Allison M. Pacelli (2008): *Mathematics and Politics: Strategy, Voting, Power, and Proof*, 2nd edition. Springer, New York, doi:10.1007/978-0-387-77645-3.
- [50] T. Nicolaus Tideman (1987): *Independence of Clones as a Criterion for Voting Rules*. *Social Choice and Welfare* 4, pp. 185–206, doi:10.1007/bf00433944.
- [51] Jun Wang, Sujoy Sikdar, Tyler Shepherd Zhibing Zhao, Chunheng Jiang & Lirong Xia (2019): *Practical Algorithms for Multi-Stage Voting Rules with Parallel Universes Tiebreaking*. In: *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, AAAI Press, doi:10.1609/aaai.v33i01.33012189.
- [52] H. P. Young (1977): *Extending Condorcet's Rule*. *Journal of Economic Theory* 16, pp. 335–353, doi:10.1016/0022-0531(77)90012-6.
- [53] T. M. Zavist & T. Nicolaus Tideman (1989): *Complete Independence of Clones in the Ranked Pairs Rule*. *Social Choice and Welfare* 6, pp. 167–173, doi:10.1007/bf00303170.
- [54] William S. Zwicker (2016): *Introduction to the Theory of Voting*. In Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang & Ariel D. Procaccia, editors: *Handbook of Computational Social Choice*, Cambridge University Press, New York, pp. 23–56, doi:10.1017/cbo9781107446984.003.



# Algorithmic Randomness, Bayesian Convergence and Merging

Simon Huttegger

University of California, Irvine  
shuttegg@uci.edu

Sean Walsh

University of California, Los Angeles  
walsh@ucla.edu

Francesca Zaffora Blando

Carnegie Mellon University  
fzaffora@andrew.cmu.edu

## Abstract

Convergence-to-the-truth results and merging-of-opinions results are part of the basic toolkit of Bayesian epistemologists. In a nutshell, the former establish that Bayesian agents expect their beliefs to almost surely converge to the truth as the evidence accumulates. The latter, on the other hand, establish that, as they make more and more observations, two Bayesian agents with different subjective priors are guaranteed to almost surely reach inter-subjective agreement, provided that their priors are sufficiently compatible. While in and of themselves significant, convergence to the truth with probability one and merging of opinions with probability one remain somewhat elusive notions. In their classical form, these results do not specify which data streams belong to the probability-one set of sequences on which convergence to the truth or merging of opinions occurs. In particular, they do not reveal whether the data streams that ensure eventual convergence or merging share any property that might explain their conduciveness to successful learning. Thus, a natural question raised by these classical results is whether the kind of data streams that are conducive to convergence and merging for Bayesian agents are uniformly characterizable in an informative way.

The results presented in this paper provide an answer to this question. The driving idea behind this work is to approach the phenomena of convergence to the truth and merging of opinions from the perspective of computability theory and, in particular, the theory of algorithmic randomness—a branch of computability theory concerned with characterizing the notion of a sequence displaying no effectively detectable patterns. We restrict attention to Bayesian agents whose subjective priors are computable probability measures and whose goal, in the context of convergence to the truth, is estimating quantities that can be effectively approximated. These are natural restrictions to impose when studying the inductive performance of more realistic, computationally limited learners. Crucially, they also allow to provide a more fine-grained analysis of both convergence to the truth and merging of opinions. Our results establish that, in this setting, the collections of data streams along which convergence and merging occur are indeed uniformly characterizable in an informative way: they are exactly the algorithmically random data streams.



# Game-Theoretic Models of Moral and Other-Regarding Agents (extended abstract)

Gabriel Istrate

West University of Timișoara

gabrielistrate@acm.org

We investigate Kantian equilibria in finite normal form games, a class of non-Nashian, morally motivated courses of action that was recently proposed in the economics literature. We highlight a number of problems with such equilibria, including computational intractability, a high price of miscoordination, and problematic extension to general normal form games. We give such a generalization based on concept of *program equilibria*, and point out that that a practically relevant generalization may not exist. To remedy this we propose some general, intuitive, computationally tractable, other-regarding equilibria that are special cases Kantian equilibria, as well as a class of courses of action that interpolates between purely self-regarding and Kantian behavior.

## 1 Introduction

Game Theory is widely regarded as the main conceptual foundation of strategic behavior. The promise behind its explosive development (at the crossroads of Economics and Computer Science) is that of understanding the dynamics of human agents and societies and, equally importantly, of guiding the engineering of artificial agents, ultimately capable of realistic, human-like, courses of action. Yet, it is clear that the main models of Game Theory, primarily based on the self-interested, rational actor model, and exemplified by the concept of Nash equilibria, are not realistic representations of the richness of human interactions. Concepts such as *bounded rationality* [63], and the limitations they impose on the computational complexity of agents' cognitive models [56] can certainly account for some of this difference. But this is hardly the only possible explanation: People behave differently from ideal economic agents not because they would be irrational [2], but since many human interactions are cooperative, rather than competitive [68], guided by social norms such as *reciprocity*, *fairness* and *inequity-aversion* [13], often involving *networked minds*, rather than utility maximization performed in isolation [27], driven by moral considerations [69] or by other not purely self-regarding behaviors, e.g. *altruism* [36] and *spite* [17, 16].

Moral considerations (should) interact substantially with game theory: indeed, the latter field has been used to propose a reconstruction of moral philosophy [9, 8, 10]; conversely, some philosophers have gone as far as to claim that we need a *moral equilibrium theory* [65]. Whether that's true or not, it is a fact that *homo economicus*, the Nash optimizer of economics, is increasingly complemented by a rich emerging typology of human behavior [28], that also contains (in Gintis's words) "*homo socialis*, the other-regarding agent who cares about fairness, reciprocity, and the well-being of others, and *homo moralis*"<sup>1</sup> ... the Aristotelian bearer of nonconsequentialist character virtues".<sup>2</sup> These claims are well-documented experimentally: for instance, Fischbacher et al. [23] investigated the percent of people

<sup>1</sup>since our agents are not necessarily human, we will use alternate names such as "moral agent" for this type of behavior.

<sup>2</sup>Gintis proposes a taxonomy of behavior with three distinct types of preferences: *self-regarding*, *other regarding* and *universalist*; a further relevant distinction is between so-called *private* and *public personas*, that leads to further types of behavior such as *homo Parochialis*, *homo Universalis* and *homo Vertus*. See [28] for further details.

having self-regarding preferences in a public goods game, showing that it is in the range of 30-40%, while the remaining were either other-regarding or moral agents. Since artificial agents (will) interact with humans, such concerns are highly relevant to the design of multiagent systems and justify the study of alternative other-regarding notions, e.g. Rong and Halpern's [32, 55] "cooperative equilibria" or *dependency theory* [62, 30]. Other-regarding considerations could be encoded (e.g. [22]) as *externalities* into agents' perceived utilities, that may lead them away from straightforward maximization of their material payoffs. However, keeping them explicit may be important for agent implementations.

The purpose of this paper is to contribute to the emerging literature on non-Nashian, morally inspired game theoretic concepts and, equally important, to bring its concerns and methods to the attention of the various communities represented in TARK. We are inspired by what we believe is one of the most intriguing classes of equilibrium concepts that can be seen as morally grounded: *Kantian (a.k.a. Hofstadter) equilibria* [54]. This notion emerged from three separate lines of research converging on an identical mathematical definition, but justifying it, however, from several very different perspectives: *superrationality* [38, 25], *team reasoning* [4], and *Kantian optimization*, respectively [54].

The common framework (most crisply developed for symmetric coordination games) only considers as relevant the action profiles where all agents choose *the same action*, choosing the action  $x$  that, if played by everyone, maximizes agents' (identical) utility functions. The justification of this restriction depends on the perspective: *superrationality* assumes that if rationality constrains an agent to choose a specific course of action  $x$ , then the same reasoning compels **all** agents (at least in the case of symmetric games, when all agents are positionally indistinguishable from the original agent) to also choose  $x$ .<sup>3</sup> In contrast, *Kantian optimization* justifies the limitation to symmetric profiles in a very different manner: Roemer [52] suggested that agents often ignore the potential for action of the other players, acting instead according to the *Kantian categorical imperative* [58] "act only according to that maxim whereby you can, at the same time, will that it should become a universal law", that is, choose a course of action that, if adopted by every agent, would bring all agents the highest payoff.<sup>4</sup> One way to formalize this idea, employed e.g. in [1], is to decouple the *material payoffs* agents receive from their (perceived) utility, which agents maximize in order to select the action. Specifically, assume the given agent  $i$  plays strategy  $x$  against action profile  $y$ . We assume that the material payoff the agent receives is  $\pi_i(x, y)$ . On the other hand the utility the agent uses to evaluate alternative  $x$  may not be equal to  $\pi_i(x, y)$  and may in fact, have in fact nothing to do with  $y$  at all! Instead,  $u_i(x, \mathbf{y}) = \pi_i(x, \bar{x}_{-i})$ , where  $\bar{x}_{-i}$  is the action profile where all agents other than  $i$  play  $x$  as well. That is, the agent evaluates the desirability of action  $x$  in isolation from the actions of the other players, as if choosing  $x$  could somehow "magically" determine the other players to adopt the same strategy.<sup>5</sup> Alternatively,  $\pi(x, \bar{x}_{-i})$  measures the extent to which action  $x$  is "the morally best course of action". **Such a justification is cognitively plausible:** experiments have shown [45] that people often employ such "universalization" arguments when judging the morality of a given behavior.

<sup>3</sup>To cite Hofstadter: "If reasoning dictates an answer, then everyone should independently come to that answer. Seeing this fact is itself the critical step in the reasoning toward the correct answer [...]". Though superrationality does away with the assumption of counterfactual independence of Nash equilibria, it is otherwise compatible with a particular version of homo economicus that requires some very strong assumptions on agent rationality (see [25] for a discussion).

<sup>4</sup>As recognized by Roemer himself and discussed e.g. in [14], the connection of Kantian equilibria to actual Kantian ideas is quite loose. Another possible interpretation is that Kantian equilibria embody *rule utilitarianism* [35]. Finally, see [60] for a discussion of the normative aspects of Kantian equilibria.

<sup>5</sup>Frank [26] refers to this as *voodoo causation*. Elster [21] argues that Kantian optimization seems to be rooted in a form of *magical thinking*, "causing agents to act on the belief (or act as if they believed) that they can have a causal influence on outcomes that are effectively outside their control". We take a descriptive, rather than normative position: such reasoning is something people **simply do**; understanding its implications is strategically valuable.

The questions we attempt to start answering in this paper are:

1. *Can we extend the definition of Kantian equilibria to cover all natural cases of "Kantian behavior"?*

However, we are **not** simply looking in our generalization for yet another equilibrium notion of primarily mathematical interest, but for one satisfying specific tractability requirements that ensure easy implementation in computational agents. Specifically, a target concept should at least be:

- I. **expressive**, i.e. indicative of realistic behavior of human agents in sufficiently typical situations
- II. **cognitively plausible**: the equilibrium should **not** be justifiable in terms of expensive epistemic assumptions (the way common knowledge of rationality can be used to justify Nash equilibria [3]);
- III. **logically tractable**: proposed equilibria should be easy to specify formally, in a way that translates to efficient implementations.
- IV. **computationally tractable**: equilibria should be easy to compute [56], since bounded rational agents are assumed to compute (and play) them.

**The main message of the paper is that such an extension is possible, but any general notion of Kantian equilibria may be of theoretical interest only:** while we give an interesting extension for certain symmetric games (Sec. 5) inspired by the concept of *program equilibria*, it's not clear how to further extend it. Together with intractability (Thm. 2) this suggests that **a general, practically relevant, notion of Kantian equilibrium might not exist.**

2. *What is the relation between Kantian equilibria and Bacharach's (informally defined) team-reasoning equilibria [4]?* The answer is that Kantian equilibria are a proper subset of team-reasoning equilibria.
3. *Given that the answer to Q1 is negative, are there more specialized equilibria related to Kantian optimization that satisfy (I)-(IV)?* We will show that there exist, indeed, several more restrictive equilibrium notions, satisfying tractability and plausibility constraints, and relate them to Kantian equilibria.
4. *Real people are seldom purely selfish or purely Kantian. (How) can we formalize this?* We give such a definition, and motivate it through the case of Prisoners' Dilemma.

The outline of the paper is as follows: In Section 3 we review some basic notions. In Section 4 we obtain some further results on (and highlight some limitations of) Kantian equilibria: first of all, we point out that finding a mixed Kantian equilibrium is computationally intractable even for two-player symmetric games (Theorem 2). Second, multiple Kantian equilibria may exist, and lack of coordination on the same equilibrium may be detrimental to players, even with all of them playing a common linear combination of Kantian actions. In Section 5 we discuss the problem of extending Kantian equilibria to non-symmetric games, giving a proposal based on the concept of program equilibria. As such, our proposal inherits the problems of this concept. Given these problems, in Section 6 we propose several other-regarding equilibria.<sup>6</sup> We show (Theorem 6) that these equilibria can be computed efficiently, that they are indeed Kantian equilibria (according to our generalized definition), and that they yield Kantian equilibria for symmetric coordination games. Finally, in Section 7 we relax the assumption that the agents are other-regarding: we assume that agents have a degree of greed, zero for Kantian agents, infinite for Nashian agents. We show (Theorem 8) how our definition applies to Prisoners' Dilemma.

For reasons of space, **most proof details are deferred to a longer version of the paper, available on arXiv [40]**. So have we done, for reasons of abundance of technical details, with some of the results: e.g. the ones the proper definition and characterization of Kantian program equilibria (Theorems 9, 10 in [40]), which also clarify the connection between Kantian and *team reasoning*.

<sup>6</sup>Generally, ethical egoism and its variant, rational egoism, are not accepted as a basis of moral behavior; counterexamples exist, [50]; however, it's fair to say that such positions are controversial, and somewhat marginal. In contrast, moral and other-regarding behaviors are better aligned, with other-regarding behavior often a consequence of moral play.

## 2 Related Work

The literature on other-regarding game-theoretic models is quite large, and a short section like this one cannot do justice to all the related, relevant work. Instead we have chosen to highlight a modicum of references directly relevant to our work.

The major impetus for this work was *Kantian optimization*. It was developed in [52, 53], developing early ideas of Laffont [43]. The current status of the theory is consolidated in the recent book [54]. A recent special issue of the *Erasmus Journal of Economics* is devoted to discussing and situating Roemer's contribution. Particularly valuable articles in this collection include [14, 60].

The other strand of ideas relevant to our work concerns the concept of *program equilibria*, defined in [66] and further investigated in [24, 41, 37, 44, 6, 19, 48]. There are several other related (and relevant) models, such as the *translucent player* model of [15, 31], or *mediated equilibria* [46].

The two other paradigms leading to the same concept for two-player symmetric coordination games, *superrationality* and (especially) *team reasoning* are, of course, relevant to our approach. Superrationality is rather different, though, and we only reiterate recommendations of [38, 25]. The main reference for team reasoning is still [5]. We also recommend papers [64, 18, 29].

Notions of symmetry in games have been insufficiently investigated, and they play an important role in defining Kantian programs. We refer to [33, 67] for such studies.

Finally, an impressive amount of work on behaviorally relevant game-theoretic notions related to moral behavior is summarized in [20]. While it is by no means comprehensive (especially with respect to the computer science literature), it is an excellent starting point.

## 3 Preliminaries

We assume knowledge of basic results of game theory at the level of a textbook such as, e.g. [49], in particular with concepts such as normal form games, best response strategy, and mixed (Nash) equilibria. All the games  $G$  we consider are normal form and, unless mentioned otherwise, have identical action sets  $Act_G$  for all players. Given a finite set  $S$ , we will define  $\Delta(S)$  to be the set of probability distributions on  $s$ . Elements of  $\Delta(S)$  are functions  $c : S \rightarrow [0, 1]$  satisfying  $\sum_{i \in S} c(i) = 1$ .  $\Delta^n := \Delta(\{1, 2, \dots, n\})$  is, geometrically, a  $(n - 1)$ -dimensional simplex. When  $G$  is a normal-form game and  $k$  a player in the game we will denote by  $\Delta_G^k$  the set of mixed actions available to player  $k$ , identified with some simplex  $\Delta^n$  with a suitable dimension. We will occasionally drop  $k$  from the notation and simply write  $\Delta_G$  instead when the player is clear from the context, or when all agents have the same action set. Given vectors  $x = (x_1, \dots, x_n)$  and  $y = (y_1, \dots, y_n)$ , we say that  $x$  *dominates*  $y$  iff  $x_i \geq y_i$  for all  $i = 1, \dots, n$ . The domination is *strict* if at least one inequality is. When comparing (mixed) action profiles  $\mathbf{a}$  and  $\mathbf{b}$ , the domination relation may apply to the vectors of agent utilities  $(u_1(\mathbf{a}), \dots, u_n(\mathbf{a}))$  and  $(u_1(\mathbf{b}), \dots, u_n(\mathbf{b}))$ , respectively. Action profiles that are strictly dominated may be assumed not to occur in game play.

A game with identical action sets is *diagonal* if every pure action profile is Pareto dominated by some profile on the diagonal, the set of action profiles where all players play the same action. A particular class of diagonal games are *coordination games*, where all player utilities are zero outside the diagonal. Such a game is *symmetric* if, additionally, agent utilities are identical for all action profiles on the diagonal.

**Definition 1.** Let  $G$  be a game with common action set  $A$ . A variation function is a function  $\phi : \Xi \times A \rightarrow A$ , for some set of parameters  $\Xi$ .<sup>7</sup> A Kantian (Hofstadter) equilibrium is a pure strategy profile

<sup>7</sup>The precise form of this definition follows [60], and is motivated by Roemer's definition of *additive/multiplicative Kantian*

		Player 2	
		C	D
Player 1	C	2, 2	0, 3
	D	3, 0	1, 1

		Player 2	
		B	S
Player 1	B	2, 3	0, 0
	S	1, 1	3, 2

Figure 1: a. Prisoners' Dilemma. b. BoS game as modified by Roemer.

$x^{opt} = (x_1^{opt}, \dots, x_n^{opt})$  that maximizes the material payoff of each agent, should everyone deviate similarly. Formally, for every agent  $i$  and  $r \in \Xi$ ,

$$V_i(x_1^{opt}, \dots, x_n^{opt}) \geq V_i(\phi(r, x_1^{opt}), \dots, \phi(r, x_n^{opt})).$$

**Example 1.** One of the original applications of Kantian equilibria was Prisoners' Dilemma (PD, Fig. 1 (a)). Kantian equilibria provide an elegant solution to the paradox: Kantian agents coordinate on action profile (C,C), as jointly doing so gives them a higher payoff than the Nash equilibrium (D,D).

Kantian equilibria are easiest to justify for symmetric diagonal games, since in this case they dominate all other action profiles, thus can be properly seen as "best course of action for all". There are symmetric (nondiagonal) games, though, where no pure strategy Kantian equilibrium is adequate, and which seem to compel us to considering mixed-strategy Kantian equilibria. An example is Hofstadter's "Platonian's Dilemma" [38], a special case of the *market entry games* of Selten and Güth [59]:

**Definition 2.** In *Platonian Dilemma*  $n$  agents (say,  $n = 20$ ) are offered to win a prize. Agents may choose to send their name to a referee. An agent wins the prize if and only if it is **the only one submitting their name**: if zero or at least two agents send their names then noone wins anything.

It is easy to see that both pure strategies, sending/not sending their name, are equally bad if adopted by all agents: they get zero payoff. A better option is to allow independent randomization:

**Definition 3.** Given a game  $G$  with identical actions, a **mixed Kantian agent** will choose a mixed strategy  $X^{OPT} \in \Delta_G$  that maximizes its expected utility, should everyone play  $X$ . For two-player symmetric games with game matrix  $A$  and variation function  $\phi(b, a) = b$  we have  $X^{OPT} = \operatorname{argmax}\{y^T A y : y \in \Delta_G\}$ .

**Lemma 1.** In *Platonian Dilemma* the probabilistic strategy where each agent independently submits their name with probability  $p$  brings an expected profit to every agent equal to  $p(1-p)^{n-1}$ . This quantity is maximized for  $p = \frac{1}{n}$ . Thus the strategy with  $p = \frac{1}{n}$  is a mixed Kantian equilibrium.

*Proof.* Let  $f(p) = p(1-p)^{n-1}$ .  $f'(p) = (1-p)^{n-1} - (n-1)p(1-p)^{n-2} = (1-p)^{n-2}(1-np)$ , so  $f$  is increasing on  $[0, 1/n]$  and decreasing on  $[1/n, 1]$ .  $\square$

## 4 Limitations of Mixed Kantian Equilibria

In this section we note some properties of mixed Kantian equilibria. They are mostly negative: finding a mixed Kantian equilibrium is intractable. Also, such equilibria may be vulnerable to miscoordination.

---

*equilibria*, with action set  $A = \mathbb{R}_+$  and variation functions  $\phi(r, a) = a+r$ ,  $a \cdot r$ , respectively. **We will mostly be concerned with variation functions of the type "change (everyone's) current action to  $b$ "** (for  $b \in A$ ). Formally,  $\Xi = A$  and  $\phi(b, a) = b$ .

#### 4.1 The Computational Intractability of Mixed Kantian Equilibria

We make an easy observation concerning the computational complexity of mixed Kantian equilibria in symmetric two player games. To our knowledge this has not been discussed before. Note: such equilibria are guaranteed to exist, since the  $(n - 1)$ -dimensional simplex of mixed strategies is a compact set and the common utility function is continuous. First of all, finding a mixed Kantian equilibrium is easy in symmetric coordination games, as all such equilibria coincide with pure Kantian equilibria.

**Theorem 1.** *Consider a finite symmetric coordination game. Then mixed Kantian equilibria coincide with pure Kantian equilibria. Hence one can compute mixed Kantian equilibria in polynomial time.*

Platonian Dilemma with  $n = 2$  shows that Theorem 1 does not extend to general symmetric games. This is no coincidence: in this case finding (or just detecting) the optimal mixed strategy is intractable:

**Theorem 2.** *The following problem, called MIXED KANTIAN EQUILIBRIUM, is NP-hard:*

*INPUT: A two-player symmetric game  $G$ , and an aspiration level  $r \in \mathbb{Q}$ .*

*TO DECIDE: Is there a mixed strategy profile  $x = (x_1, \dots, x_N)$  such that the utility of every player under common mixed action  $x_1a_1 + x_2a_2 + \dots + x_m a_m$  is at least  $r$ ?*

*Proof.* We point out to the existence of a reduction from CLIQUE to MIXED KANTIAN EQUILIBRIUM, that shows that the latter problem is NP-hard. In fact the reduction will only consider symmetric games with 0/1 payoffs.

Consider, indeed, a graph  $g$ . Let  $k$  be an integer and  $(g, k)$  be the corresponding instance of CLIQUE.

Define the symmetric two-player game  $G$  whose payoff matrix is the adjacency matrix  $A$  of  $g$ .

Mixed Kantian equilibria  $x = (x_1, \dots, x_N)$  of  $G$  correspond to optimal solutions of the following quadratic program:

$$\begin{cases} \max(x^T A x) \\ x_1 + \dots + x_N = 1 \\ x_1, \dots, x_N \geq 0. \end{cases} \quad (1)$$

This is a problem that has been called [12] *the standard quadratic optimization problem*, and has been investigated substantially in the global optimization literature (see e.g. [11]). A beautiful result due to Motzkin and Straus [47] can be restated as claiming that for programs whose matrix  $A$  is the adjacency matrix of a graph  $g$ , if  $o$  is the optimum of problem (1) then  $\frac{1}{1-o}$  is the size of the maximum clique in  $g$ .

Hence  $(g, k) \in \text{CLIQUE}$  if and only if  $(G, \frac{k-1}{k}) \in \text{MIXED-KANTIAN-EQUILIBRIUM}$ . □

#### 4.2 Multiple Equilibria and miscoordination

Optimal diagonal action profiles may fail to be unique. If the agents are not communicating (and no implicit coordination mechanisms are acting, e.g., one of the action profiles being a *focal point*, such as in the Hi-Lo game from [5]), agents may reach a suboptimal action profile due to their lack of coordination on the same optimal action: Consider, indeed, the game in Figure 2.  $(C, C)$  and  $(E, E)$  are equally good pure (and mixed) Kantian equilibria. But if one player plays  $C$  and the other plays  $E$  the resulting outcomes are the worst possible for both of them, being dominated by every single possible strategy profile! Randomizing among Kantian actions might not help either: miscoordination impacts even "Kantian" scenarios, where players, lacking a salient equilibrium to coordinate on, play a joint mixed strategy formed of Kantian actions.<sup>8</sup> We quantify the degradation in performance as follows:

<sup>8</sup>Such a scenario is, of course, not justifiable from a usual rational choice perspective. But **it is justifiable in a Kantian setting where every player believes that choosing a pure action  $a$  will immediately make all other players do the same:** a

		P1 2		
		C	D	E
P1 1	C	5, 5	3, 6	1, 2
	D	6, 3	4, 4	6, 3
	E	2, 1	3, 6	5, 5

		P1 2	
		C	D
C	C	10, 1	0, 0
	D	0, 0	4, 2

		P1 2	
		B	S
B	B	6, 1	0, 0
	S	0, 0	3, 2

		P1 2	
		C	S
C	C	10, 10	100, 200
	S	200, 100	6, 6

Figure 2: (a). A game with multiple Kantian equilibria. (b). Modified BoS (Example 3). (c). Modified BoS (Example 2). (d). An anti-coordination game.

**Definition 4.** For a symmetric game  $G$  with strictly positive payoffs let  $NC$  be the set of mixed action profiles composed of Kantian actions only. The price of miscoordination of  $G$  is the ratio  $p(G) = \sup_{a \in NC} \frac{u_i(X^{OPT})}{u_i(a)}$ . Because of symmetry this does not depend on the particular choice of player  $i$ .

The following result shows that the price of miscoordination can be arbitrarily large:

**Theorem 3.** Let  $G$  be a symmetric diagonal game with  $k \geq 2$  players and  $r \geq 1$  pure Kantian actions. Then the price of miscoordination of  $G$  is in the range  $[1, r^{k-1}]$ . Both bounds are tight and can be reached in settings where players choose a Kantian action uniformly at random.

The merit of this simple result is to point out that the definition of generalized Kantian equilibria needs to include scenarios where randomness is correlated, as in *correlated equilibria* (see e.g. [61]).

*Proof.* The price of miscoordination is insensitive to dividing all utilities by the same factor  $\lambda$ , so w.l.o.g. one may assume that the utilities agent receive on pure Kantian equilibrium profiles is 1. For the mixed action  $\mathbf{a}$  where players play the  $r$  Kantian actions (w.l.o.g.  $1, 2, \dots, r$ ) with probabilities  $p_1, p_2, \dots, p_r$  (which add up to 1), its expected utility is  $E[u_i(\mathbf{a})] = \sum u_i(i_1, i_2, \dots, i_k) \cdot p_{i_1} p_{i_2} \dots p_{i_k} \geq \sum_{i=1}^r p_i^k \geq r \cdot \frac{1}{r} = \frac{1}{r^{k-1}}$ , by Jensen’s inequality. The upper bound is obtained when off-diagonal action profiles formed of Kantian actions only have utilities equal to 0. As for the lower bound, for diagonal games by domination we have  $u_i(i_1, i_2, \dots, i_k) \leq 1$ , so  $E[u_i(\mathbf{a})] = \sum u_i(i_1, i_2, \dots, i_k) \cdot p_{i_1} p_{i_2} \dots p_{i_k} \leq \sum p_{i_1} p_{i_2} \dots p_{i_k} = (p_1 + p_2 + \dots + p_r)^k = 1$ . A game realizing the lower bound is the one where agent utilities on all pure action profiles are equal to 1. □

## 5 Kantian Program Equilibria in (Pareto) Symmetric Games

Definition 1 of Kantian equilibria makes the most sense in symmetric coordination games, but does not capture all the intuitive cases of Kantian behavior. Indeed, let us consider the BoS game, as modified by Roemer (Fig. 1 (b)).<sup>9</sup> Intuitively, agents would perhaps agree that the following **protocol** could be called Kantian, in that it is symmetric and both players benefit if they both follow it: flip a fair coin; if it comes out *heads*, they (both) play  $B$ , else both play  $S$ . As described, the protocol requires the centralized choice of a random bit, but it could easily be implemented in a distributed manner by making each of the

---

player may use the Kantian imperative to restrict itself to pure Kantian equilibria, then use the assumption to justify playing a convex combination of pure Kantian equilibria it is indifferent between.

<sup>9</sup>Roemer ([54], Proposition 2.3) argues that  $(S, B)$  is a simple Kantian equilibrium. His argument is, however, ad-hoc, based on making this profile "diagonal" by flipping the order of  $B$  and  $S$  for the second player, and the conclusion that  $(B, S)$  is Kantian is, we feel, unintuitive, since  $(B, B), (S, S)$  strictly dominate it. Our protocol plays  $\frac{1}{2}(B, B) + \frac{1}{2}(S, S)$ , different from (and better than) what Roemer calls the mixed Kantian equilibrium, where row player plays  $\frac{3}{8}B + \frac{5}{8}S$  and the column player plays  $\frac{3}{8}S + \frac{5}{8}B$ .

two agents flip a (fair) coin and taking their XOR. The implementation of the protocol (Algorithm 5.1) assumes that each agent is parameterized by an *agent ID*  $i \in \{1, 2\}$  (not needed in this particular case) and vector  $(myb, otherb)$  of random bit choices, one for each player, and shared between players.

An even more dramatic case is that of the game from Figure 2 (d), where the best outcomes are not symmetric. In these cases it even seems irrational for the agents to play symmetric action profiles, since these action profiles are dominated by all the other action profiles! Rather, it is plausible that agents would agree that they need to **anticoordinate**, but they have different preferences for the joint action profile to coordinate upon. A "best for all" solution would jointly play a random anti-coordinated profile  $\frac{1}{2}(C, S) + \frac{1}{2}(S, C)$ . As in the previous example, this course of action can be implemented by the two agents in a distributed manner, by jointly playing according to the protocol in Algorithm 5.2. In this example, in addition to the extra bit *otherb* communicated by the other player, the protocol of each agent **makes explicit use of the agents' own *id***,  $i \in \{1, 2\}$ .

**Algorithm 5.1:** (*BoS*)

$BOS(i :: ID, myb :: BIT, otherb :: BIT)$

Randomly choose a bit  $myb \in \{0, 1\}$   
 communicate *mybit* to the  
 other player as its *otherb*.  
**if** [ $myb \oplus otherb == 0$ ]  
     **then** play *B*  
     **else** play *S*

**Algorithm 5.2:** (*Anticoord*)

$Anticoord(i :: ID, myb :: BIT, otherb :: BIT)$

Randomly choose a bit  $myb \in \{0, 1\}$   
 communicate *myb* to the  
 other player as its *otherb*.  
**if** [ $myb \oplus otherb \equiv i \pmod{2}$ ]  
     **then** play *C*  
     **else** play *S*

The intuitive conclusion of these two examples is simple: Definition 1 is not sufficient. Some simple games may have coordinated **protocols** that could properly be called "Kantian". In this section **we give a somewhat more general definition<sup>10</sup> of Kantian equilibria, but not for general games, only for a class of "symmetric" games**. There are multiple definitions of game symmetry in the literature [33, 67]; the most important one requires that for every player  $i$ , action profile  $(x_1, x_2, \dots, x_n)$  and permutation  $\sigma \in S_n$ , we have  $u_{\sigma(i)}(x_1, x_2, \dots, x_n) = u_i(x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(n)})$ . We use a slightly less demanding definition (we call our version *Pareto symmetry*, see Definition 16 in the longer version [40]) to capture some asymmetric games like Romer's version of BoS. This game is asymmetric, but in an inconsequential way: all the asymmetries concern dominated profiles. Our extension is inspired by the concept of **program equilibria** [39, 66]. To define these equilibria we associate to every finite normal-form game an extended game whose actions correspond to *programs*. Agents' programs have access to own and others' sources<sup>11</sup>, and can act on them. A program equilibrium is a Nash equilibrium in the extended game. We extend this idea to Kantian equilibria:

**Definition 5.** *Given a (Pareto symmetric) game  $G$  with identical actions sets for all players, a Kantian program equilibrium in  $G$  is a probability distribution  $p$  on the action profiles of  $G$  such that: (a).  $p$  has*

<sup>10</sup>other plausible alternatives are discussed (and ruled out) in Section 13 of the longer version [40]. Of special interest is the relation between Kantian equilibria and *team-reasoning equilibria*.

<sup>11</sup>this aspect was important for Nashian optimization. It will be less so for us, since deviations are not present in the definition of Kantian equilibria. On the other hand, a Kantian agent playing program  $P$  can make sure it is not taken advantage upon by the other players, either alone, as in [66], by reading other players' programs and only playing  $P$  when all do, or with the help of a *mediator*, which implements on behalf of all players the following protocol: if all agents follow  $P$  then the mediator will simulate  $P$  on behalf of the agent; otherwise it will play in a Nashian way.

its support on the set of action profiles that are Pareto optimal, that is not Pareto dominated by any other action profile. (b).  $p$  is implemented by **agents playing a common program  $P$**  in the extended game. (c). there exists no probability distribution  $q$  with properties a. and b. such that the vector of expected utilities  $(E[u_i(q)])$  Pareto dominates the vector  $(E[u_i(p)])$ .

The assumption that *players have the same action sets* is motivated by the "common program" requirement of Def. 5 (b). Point (a). encodes a simple rationality condition. Points (b). and (c). embody a generalized version of the Kantian categorical imperative: (b). encodes the constraint of identical behavior, (c). encodes the fact that implementing  $p$  is "a best action" for all players.

**We only need to formalize what we mean by *program* in this definition.** The semantics is inspired by the one in [66], but the full formalization is somewhat subtle. We defer a full presentation to Section 15 of the longer version [40]. A couple of technical points are, however, worth stating:

- The semantics of programs in [66] does not allow for any synchronization between different versions of the same program, other than testing whether they are syntactically equivalent. Since (as recognized in Theorem 3) we need to include correlated randomness, we need to extend the semantics of programs from [66], where this is not possible. There are many ways to do it, but one way is to allow *correlated sampling from distributions*: all the programs get an identical sample from a given distribution.
- **However, simply adding correlated sampling of action profiles to the semantics of programs from [66] leads (see Theorem 9 in the long version [40]) to paradoxical results:** every convex combination of Pareto optimal strategy profiles would be a Kantian program equilibrium. This is an issue: in Prisoners' Dilemma profiles  $(C, D), (D, C)$  could perhaps be justified from a "team reasoning" perspective where one player "sacrifices itself" so that the other one walks free. To accept them as "Kantian equilibria" seems, however, problematic (see also the discussion in sections 13.3 and 14 of [40])
- If, on the other hand, **agent programs didn't communicate at all, used no private randomness, or used no specific ID/payoff information** then they would run identically for all agents, coordinating on the same action (excluding, thus, scenarios like that of Example 3, that we want to model).
- We will take a middle-ground approach, and assume that agents can use their ID and the information about the game payoffs in a very limited way, that makes the program act "identically with respect to a group of symmetries acting transitively on the set of agents". This requires us to restrict ourselves to the class of Pareto symmetric games of Definition 16 in the longer version [40], whose set of Pareto dominant action profiles has such symmetries. The precise technical details are spelled out in Section 15 of the longer version [40], where we prove (Theorem 10) a characterization of Kantian equilibria for Pareto symmetric games which also shows that Kantian program equilibria are a strict subset of the class of team-reasoning equilibria.

Algorithms 4.1 and 4.2 lend some credibility to the intuition that Kantian equilibria are somehow related to some "symmetric" notion of correlated equilibria. This intuition is correct: in Definition 19 in the longer version [40] we define a notion of "correlated symmetric equilibrium". We then prove:

**Theorem 4.** *Correlated symmetric equilibria of symmetric games are Kantian program equilibria.*

Kantian program equilibria allow players to obtain a better expected payoff in Platonia Dilemma:

**Theorem 5.** *Algorithm 5.3 implements a Kantian program equilibrium for Platonia Dilemma.*

*Proof.* Points (a). and (b). from the definition of Kantian program equilibria are clear, the only one that merits a discussion is point (c).

The expected utility of each player under Algorithm 5.3 is equal to  $1/n$ . Since the sum of utilities of all players under a particular set of random choices is equal to 1, no vector of expected utilities can strictly dominate the vector  $(1/n, 1/n, \dots, 1/n)$  of expected utilities for the Algorithm.  $\square$

**Algorithm 5.3:** CHOOSE-WINNER( $i, b_1, b_2, \dots, b_n$ )

```

Randomly choose an integer  $b_i \in \mathbb{Z}_n$ 
if  $[\sum_{j=1}^n b_j \equiv i \pmod{n}]$ 
  then S(UBMIT)
  else D(ON'T)

```

## 6 Some computationally efficient other-regarding equilibria

As defined in the previous section, Kantian program equilibria for games with identical action sets inherit some of the definitional problems of "ordinary" program equilibria. Among them:

- *fragility*: (Kantian) program equilibria are sensitive (see e.g. [48]) to the precise specification of programs: do we insist that all agent programs are syntactically identical, or just "do the same thing"? See [44, 37] for some attempted solutions for program equilibria that could be adapted to our setting.
- *lack of generality*: Definition 5 it is only applicable to (some of the) games with identical action sets. To further generalize it to all finite normal-form games one would need to specify what it means for two agents to "take the same course of action" in settings with differing action sets.
- *lack of predictive power*: There may be multiple (even infinitely many) Kantian program equilibria.

Given these objections, and with constraints (I)-(IV) in mind, we propose in the sequel a substantially more modest approach: Rather than seeking a general definition of Kantian equilibria we propose instead *several* other-regarding equilibria. **They all correspond intuitively to real-life situations, are tractable, can be justified by team reasoning and are related, for symmetric coordination games, to Kantian equilibria.** One was independently suggested in [42], the other ones are first introduced here:

**Definition 6.** A Rawlsian equilibrium is a probability distribution over Pareto optimal profiles maximizing the egalitarian social welfare (the expected utility of the worst-off player) and is strictly dominated by no other profile with this property. Such equilibria implement the idea of justice as fairness [51].

**Example 2.** We modify the BoS example as in Fig. 2 (c): perhaps 1 is a classical music lover, that gets a higher utility than the other player by going, together with its partner, to any of the two concerts. Then  $(S,S)$  is the (unique) Rawlsian equilibrium. Choosing such an equilibrium is an example of altruistic behavior from player 1, since it maximizes the payoff of its non-music-lover partner.

**Definition 7.** A Bentham-Harsányi equilibrium is a probability distribution on Pareto optimal profiles maximizing the sum of expected payoffs. See [34] for a philosophical motivation. A best-off equilibrium is a prob. distrib. on Pareto optimal profiles maximizing the largest expected payoff, and strictly dominated by no profile with this property. E.g., in Exp. 2  $(B,B)$  is the unique Bentham-Harsányi/best-off equilibrium.

Although a best-off equilibrium may not seem "fair", there exist real-life "team reasoning" situations that elicit behavior suggestive of such an equilibrium: one such example is, for instance, scenarios where members of a team "sacrifice" for one of their members (e.g. parents for a child).

The equilibrium notions we introduced so far implicitly assumed that player utility is given by material payoffs. Sometimes the frustration a player feels is derived by counterfactually comparing its realized payoff with all possible ones. There are many implementations of this idea. The following notion quantifies the extent to which a given profile is worse for the given player than a random profile.

**Definition 8.** *The percentile index of profile  $a$  for player  $i$  is the percentage of Pareto optimal profiles that would get  $i$  a strictly better payoff than  $a$ . A Rawlsian percentile equilibrium is a profile minimizing the largest expected percentile index of all players, and strictly dominated by no profile with this property.*

**Example 3.** *Consider the game shown in Figure 2 (b). Then percentile indices of Pareto optimal profiles are  $(0, 100)$  for  $(C, C)$ , and  $(100, 0)$  for  $(D, D)$ , respectively. Profile  $\frac{1}{2}(C, C) + \frac{1}{2}(D, D)$  is a Rawlsian percentile equilibrium. Player 1 gets average utility 7 while player 2 gets average utility  $\frac{3}{2}$ .*

An even less cognitively sophisticated model of agent frustration relies on classifying outcomes as "happy/not happy". The following is a simple example of such a notion:

**Definition 9.** *The natural expectation point of player  $i$  is the median (over all undominated pure strategy profiles) payoff. If there are two medians then the average value is taken. A player is happy in a pure strategy profile  $a$  iff its payoff is larger or equal than its natural expectation point and unhappy otherwise.*

*An aspiration equilibrium is a mixed strategy profile that minimizes the largest probability of unhappiness among all players and is strictly dominated by no other profile with this property.*

**Example 4.** *Take a coordination game with payoffs  $(C, C) \rightarrow (10, 1)$ ,  $(D, D) \rightarrow (9, 2)$ ,  $(E, E) \rightarrow (8, 3)$ ,  $(F, F) \rightarrow (4, 7)$ . The natural expectation points of players are 8.5 and 2.5, respectively. The first player is happy in  $(C, C)$  and  $(D, D)$ , the second in  $(E, E)$ ,  $(F, F)$ . Hence in  $\frac{1}{4}(C, C) + \frac{1}{4}(D, D) + \frac{1}{4}(E, E) + \frac{1}{4}(F, F)$  the players are happy 50% of the time and no mixed action profile can do any better.*

Unlike general Kantian program equilibria, the equilibria we defined are computationally tractable:

**Theorem 6.** *Rawlsian, Rawlsian percentile, Bentham-Harsányi, best-off, aspiration equilibria exist and can be found by solving a sequence of linear programs (hence in polynomial time).*

We now connect our other regarding equilibria to Kantian equilibria in symmetric coordination games. We call an equilibrium point *extremal* if it cannot be written as a nontrivial convex combination of other (similar) equilibria. We show that extremal self-regarding equilibria generalize Kantian pure equilibria. Extremality is needed, since our equilibria are closed under convex combinations (such combinations are justifiable from a magical thinking perspective, see footnote 7), while pure Kantian equilibria are not. Because of Thm. 2 no similar connection is likely for mixed Kantian equilibria:

**Theorem 7.** *In symmetric diagonal games Rawlsian, Bentham-Harsányi, best-off, Rawlsian percentile, aspiration equilibria coincide with convex combinations of Kantian pure equilibria.*

## 7 Agents with bounded greed

So far we have assumed that people are other-regarding. In reality people are not unrestricted optimizers, nor are they perfect Kantian moralists. Alger and Weibull [1] attempted to interpolate between utilitarian agents and Kantian ones, by defining *homo moralis* to be an agent whose utility has the form  $u_i(x, y) = (1 - k)\pi(x, y) + k\pi(x, x)$ , where  $k \in [0, 1]$  is the so-called *degree of morality* of the agent. They showed that evolutionary models with assortative mixing and incomplete information favor a particular kind of *homo moralis*, those whose degree of morality coincides with the degree of assortativity of the matching process. Interesting as this result is, it has some weaknesses. For instance [1], *homo moralis* behaves like *homo economicus* in Prisoners' Dilemma and all constant-sum games when  $k \neq 1$ . In other words, agent behavior is not sensitive to the degree of morality, as long as the agent is not Kantian.

We give (for symmetric games, but the idea can be extended to general ones, via Kantian program equilibria) a definition with the same overall intention, but capturing a slightly different agent behavior:

**Definition 10.** Let  $\lambda \in [1, \infty]$ . Agent  $i$  is called  $\lambda$ -utilitarian if, for every action profile  $(a_i, b)$ , its utility  $u_i(a_i, b)$  is (a).  $\pi_i(a_i, (\bar{a}_i)_{-i})$  if  $a_i$  is a Kantian action. (b). 0 if  $a_i$  is not Kantian and  $\pi_i(a_i, b) \leq \lambda \cdot \pi_i(X^{OPT})$ ; (c).  $\pi_i(a, b)$  if  $a_i$  is not Kantian and  $\pi_i(a_i, b) \leq \lambda \cdot \pi_i(X^{OPT})$ . I.e., a  $\lambda$ -utilitarian agent deviates from its Kantian action  $X^{OPT}$  **only** if the utility it obtains is more than  $\lambda$  times larger.

We call the number  $\frac{1}{\lambda-1}$  the *greed index* of  $i$ . It varies between 0 (Kantian agents) and  $\infty$  (purely utilitarian ones). The natural equilibrium concept for such agents is no longer Kantian, but Nash equilibrium. Definition 10 allows giving an empirically plausible justification of all possible outcomes in PD:

**Theorem 8.** All pure action profiles in PD are Nash equilibria of agents with varying degrees of greed.

*Proof.* Bounded-greed agents still coordinate on the Kantian equilibrium  $(C, C)$  as long as both their greed indices are  $< 2$  (i.e. they would need at least a twofold increase in payoff to deviate). If one of them has greed index  $< 2$  and the other one has greed index  $\geq 2$ , then the latter one will defect. If both agents have greed indices  $\geq 2$ , then they will coordinate, just as if utilitarian agents would do, on the Nash equilibrium  $(D, D)$ .  $\square$

## 8 Conclusions

Our main contribution is bringing Kantian equilibria (and related concepts) to the attention of TARK community, showing that this notion is theoretically interesting, but that the road to implementable behaviors probably goes through less general equilibrium concepts. Many of the notions we introduced, on the other hand, including Kantian program equilibria and bounded greed agents, deserve further investigation. For instance a justification like that of Theorem 8 could be used as a rationality criterion. One could look for evolutionary justifications of bounded greed agents along the lines of [1]. One could use such agents in relation to work on the concept of *price of anarchy* [57]. On a more conceptual level, the use of *frames* in game theory [5, 7] and how this interacts with equilibrium notions deserves further study. Finally, several open problems remain: Can we find algorithms for our equilibria that bypass the need for solving multiple LP's? Is the problem from Theorem 2 NP-complete (i.e. in NP)?

## References

- [1] Ingela Alger & Jörgen W Weibull (2013): *Homo moralis - preference evolution under incomplete information and assortative matching*. *Econometrica* 81(6), pp. 2269–2302, doi:10.3982/ECTA10637.
- [2] Dan Ariely (2010): *Predictably irrational*. Harper.
- [3] Robert Aumann & Adam Brandenburger (1995): *Epistemic conditions for Nash equilibrium*. *Econometrica: Journal of the Econometric Society*, pp. 1161–1180, doi:10.2307/2171725.
- [4] Michael Bacharach (1999): *Interactive team reasoning: A contribution to the theory of co-operation*. *Research in economics* 53(2), pp. 117–147, doi:10.1006/reec.1999.0188.
- [5] Michael Bacharach (2006): *Beyond individual choice: teams and frames in game theory*. Princeton University Press, doi:10.1515/9780691186313.
- [6] Mihaly Barasz, Paul Christiano, Benja Fallenstein, Marcello Herreshoff, Patrick LaVictoire & Eliezer Yudkowsky (2014): *Robust Cooperation in the Prisoner's Dilemma: Program Equilibrium via Provability Logic*. *arXiv preprint https://arxiv.org/abs/1401.5577*.
- [7] José Luis Bermúdez (2021): *Frame it Again: New Tools for Rational Decision-making*. Cambridge University Press.
- [8] Kenneth G. Binmore (1994): *Game theory and the social contract: just playing*. M.I.T. Press.

- [9] Kenneth G. Binmore (1994): *Game theory and the social contract: playing fair*. M.I.T. Press.
- [10] Kenneth G. Binmore (2005): *Natural justice*. Oxford University Press, USA, doi:10.1093/acprof:oso/9780195178111.001.0001.
- [11] Immanuel M Bomze (1997): *Evolution towards the maximum clique*. *Journal of Global Optimization* 10(2), pp. 143–164, doi:10.1023/A:1008230200610.
- [12] Immanuel M Bomze (1998): *On standard quadratic optimization problems*. *Journal of Global Optimization* 13(4), pp. 369–387, doi:10.1023/A:1008369322970.
- [13] Samuel Bowles & Herbert Gintis (2013): *A cooperative species: Human reciprocity and its evolution*. Princeton University Press.
- [14] Matthew Braham & Martin van Hees (2020): *Kantian Kantian Optimization*. *Erasmus Journal for Philosophy and Economics* 13(2), pp. 30–42, doi:10.23941/ejpe.v13i2.513.
- [15] Valerio Capraro & Joseph Y Halpern (2019): *Translucent players: Explaining cooperative behavior in social dilemmas*. *Rationality and Society*, pp. 371–408, doi:10.1177/2F1043463119885102.
- [16] Jing Chen & Silvio Micali (2016): *Auction revenue in the general spiteful-utility model*. In: *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pp. 201–211, doi:10.1145/2840728.2840741.
- [17] Po-An Chen & David Kempe (2008): *Altruism, selfishness, and spite in traffic routing*. In: *Proceedings of the 9th ACM conference on Electronic Commerce*, pp. 140–149, doi:10.1145/1386790.1386816.
- [18] Andrew M Colman & Natalie Gold (2018): *Team reasoning: Solving the puzzle of coordination*. *Psychonomic Bulletin & Review* 25(5), pp. 1770–1783, doi:10.1080/10002003098538748.
- [19] Andrew Critch (2019): *A parametric, resource-bounded generalization of Löb’s theorem, and a robust cooperation criterion for open-source game theory*. *The Journal of Symbolic Logic*, pp. 1–15, doi:10.1017/jsl.2017.42.
- [20] Sanjit Dhami (2016): *The foundations of behavioral economic analysis*. Oxford University Press.
- [21] Jon Elster (2017): *On seeing and being seen*. *Social Choice and Welfare* 49(3-4), pp. 721–734, doi:10.1007/s00355-017-1029-9.
- [22] Ernst Fehr & Klaus M Schmidt (1999): *A theory of fairness, competition, and cooperation*. *The quarterly journal of economics* 114(3), pp. 817–868, doi:10.1162/003355399556151.
- [23] Urs Fischbacher, Simon Gächter & Ernst Fehr (2001): *Are people conditionally cooperative? Evidence from a public goods experiment*. *Economics Letters* 71(3), pp. 397–404, doi:10.1016/S0165-1765(01)00394-9.
- [24] Lance Fortnow (2009): *Program equilibria and discounted computation time*. In: *Proceedings of the 12th Conference on Theoretical Aspects of Rationality and Knowledge*, pp. 128–133, doi:10.1145/1562814.1562833.
- [25] Ghislain Fourny (2020): *Perfect Prediction in normal form: Superrational thinking extended to non-symmetric games*. *Journal of Mathematical Psychology* 96, p. 102332, doi:10.1016/j.jmp.2020.102332.
- [26] Robert H Frank (2004): *What Price the Moral High Ground? Ethical Dilemmas in Competitive Environments*. Princeton University Press.
- [27] Herbert Gintis (2016): *Individuality and entanglement: the moral and material bases of social life*. Princeton University Press, doi:10.2307/j.ctvc779cx.
- [28] Herbert Gintis (2016): *A Typology of Human Morality*. In David S. Wilson & Alan Kirman, editors: *Complexity and Evolution: Towards a New Synthesis for Economics*, M.I.T. Press, doi:10.7551/mitpress/9780262035385.003.0007.
- [29] Natalie Gold & Andrew M Colman (2020): *Team reasoning and the rational choice of payoff-dominant outcomes in games*. *Topoi* 39(2), pp. 305–316, doi:10.1007/s11245-018-9575-z.
- [30] Davide Grossi & Paolo Turrini (2012): *Dependence in games and dependence games*. *Autonomous Agents and Multi-Agent Systems* 25(2), pp. 284–312, doi:10.1007/s10458-011-9176-3.

- [31] Joseph Y Halpern & Rafael Pass (2018): *Game theory with translucent players*. *International Journal of Game Theory* 47(3), pp. 949–976, doi:10.1007/s00182-018-0626-x.
- [32] Joseph Y Halpern & Nan Rong (2010): *Cooperative equilibrium*. In: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pp. 1465–1466. Available at <http://ifaamas.org/Proceedings/aamas2010/pdf/02%20Extended%20Abstracts/Red/R-49.pdf>.
- [33] Nicholas Ham (2013): *Notions of Symmetry for Finite Strategic-Form Games*. arXiv preprint <https://arxiv.org/abs/1311.4766>.
- [34] John C Harsanyi (1955): *Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility*. *Journal of Political Economy* 63(4), pp. 309–321, doi:10.1086/257678.
- [35] John C Harsanyi (1977): *Rule utilitarianism and decision theory*. *Erkenntnis* 11(1), pp. 25–53, doi:10.1007/BF00169843.
- [36] Martin Hoefer & Alexander Skopalik (2013): *Altruism in atomic congestion games*. *ACM Transactions on Economics and Computation (TEAC)* 1(4), pp. 1–21, doi:10.1145/2542174.2542177.
- [37] Wiebe van der Hoek, Cees Witteveen & Michael Wooldridge (2013): *Program equilibrium—a program reasoning approach*. *International Journal of Game Theory* 42(3), pp. 639–671, doi:10.1007/s00182-011-0314-6.
- [38] Douglas Hofstadter (1985): *Dilemmas for Superrational Thinkers, Leading up to a Luring Lottery*. In: *Metamagical Themas: Questing for the Essence of Mind and Pattern*, Basic Books.
- [39] John V Howard (1988): *Cooperation in the Prisoner’s Dilemma*. *Theory and Decision* 24(3), p. 203, doi:10.1007/BF00148954.
- [40] Gabriel Istrate (2021): *Game-theoretic Models of Moral and Other-Regarding Agents*. arXiv preprint <http://arxiv.org/abs/2012.09759v2>.
- [41] Adam Tauman Kalai, Ehud Kalai, Ehud Lehrer & Dov Samet (2010): *A commitment folk theorem*. *Games and Economic Behavior* 69(1), pp. 127–137, doi:10.1016/j.geb.2009.09.008.
- [42] Ioannis Kordonis (2020): *A Model for Partial Kantian Cooperation*. In: *Advances in Dynamic Games*, Springer, pp. 317–346, doi:10.1007/978-3-030-56534-3\_13.
- [43] Jean-Jacques Laffont (1975): *Macroeconomic constraints, economic efficiency and ethics: An introduction to Kantian economics*. *Economica* 42(168), pp. 430–437, doi:10.2307/2553800.
- [44] Patrick LaVictoire, Benja Fallenstein, Eliezer Yudkowsky, Mihaly Barasz, Paul Christiano & Marcello Herreshoff (2014): *Program equilibrium in the Prisoner’s Dilemma via Löb’s theorem*. In: *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*. Available at [https://mipc.inf.ed.ac.uk/2014/papers/mipc2014\\_lavictoire\\_etal.pdf](https://mipc.inf.ed.ac.uk/2014/papers/mipc2014_lavictoire_etal.pdf).
- [45] Sydney Levine, Max Kleiman-Weiner, Laura Schulz, Joshua Tenenbaum & Fiery Cushman (2020): *The logic of universalization guides moral judgment*. *Proceedings of the National Academy of Sciences* 117(42), pp. 26158–26169, doi:10.1073/pnas.2014505117.
- [46] Dov Monderer & Moshe Tennenholtz (2009): *Strong mediated equilibrium*. *Artificial Intelligence* 173(1), pp. 180–195, doi:10.1016/j.artint.2008.10.005.
- [47] Theodore S Motzkin & Ernst G Straus (1965): *Maxima for graphs and a new proof of a theorem of Turán*. *Canadian Journal of Mathematics* 17, pp. 533–540, doi:10.4153/CJM-1965-053-6.
- [48] Caspar Oesterheld (2019): *Robust program equilibrium*. *Theory and Decision* 86(1), pp. 143–159, doi:10.1007/s11238-018-9679-3.
- [49] Martin Osborne & Ariel Rubinstein (1994): *A Course in Game Theory*. M.I.T. Press.
- [50] Ayn Rand (1964): *The virtue of selfishness*. Penguin.
- [51] John Rawls (2001): *Justice as fairness: A restatement*. Harvard University Press.

- [52] John E Roemer (2010): *Kantian equilibrium*. *Scandinavian Journal of Economics* 112(1), pp. 1–24, doi:10.1111/j.1467-9442.2009.01592.x.
- [53] John E Roemer (2015): *Kantian optimization: A microfoundation for cooperation*. *Journal of Public Economics* 127, pp. 45–57, doi:10.1016/j.jpubeco.2014.03.011.
- [54] John E Roemer (2019): *How We Cooperate: A Theory of Kantian Optimization*. Yale University Press, doi:10.2307/j.ctvfc52jk.
- [55] Nan Rong & Joseph Y Halpern (2013): *Towards a deeper understanding of cooperative equilibrium: characterization and complexity*. In: *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pp. 319–326. Available at <http://www.ifaamas.org/Proceedings/aamas2013/docs/p319.pdf>.
- [56] Iris van Rooij, Mark Blokpoel, Johan Kwisthout & Todd Wareham (2019): *Cognition and intractability: A guide to classical and parameterized complexity analysis*. Cambridge University Press, doi:10.1093/comjnl/bxm038.
- [57] Tim Roughgarden (2005): *Selfish Routing and the Price of Anarchy*. M.I.T. Press.
- [58] Sally Sedgwick (2008): *Kant's groundwork of the metaphysics of morals: an introduction*. Cambridge University Press, doi:10.1017/CBO9780511809538.
- [59] Reinhard Selten & Werner Güth (1982): *Equilibrium point selection in a class of market entry games*. In: *Games, economic dynamics, and time series analysis*, Springer, pp. 101–116, doi:10.1007/978-3-662-41533-7\_6.
- [60] Itai Sher (2020): *Normative Aspects of Kantian Equilibrium*. *Erasmus Journal for Philosophy and Economics* 13(2), pp. 43–84, doi:10.23941/ejpe.v13i2.514.
- [61] Yoav Shoham & Kevin Leyton-Brown (2009): *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press.
- [62] Jaime Simão Sichman & Rosaria Conte (2002): *Multi-agent dependence by dependence graphs*. In: *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*, pp. 483–490, doi:10.1145/544741.544855.
- [63] Herbert Alexander Simon (1997): *Models of bounded rationality: Empirically grounded economic reason*. 3, M.I.T. Press, doi:10.7551/mitpress/4711.001.0001.
- [64] Robert Sugden (2003): *The logic of team reasoning*. *Philosophical explorations* 6(3), pp. 165–181, doi:10.1080/10002003098538748.
- [65] William J Talbott (1998): *Why We Need a Moral Equilibrium Theory*. In P. Danielson, editor: *Modeling Rationality, Morality and Evolution*, Oxford University Press.
- [66] Moshe Tennenholtz (2004): *Program equilibrium*. *Games and Economic Behavior* 49(2), pp. 363–373, doi:10.1016/j.geb.2004.02.002.
- [67] Fernando A Tohmé & Ignacio D Viglizzo (2019): *Structural relations of symmetry among players in strategic games*. *International Journal of General Systems* 48(4), pp. 443–461, doi:10.1080/03081079.2019.1573228.
- [68] Michael Tomasello (2009): *Why we cooperate*. M.I.T. Press, doi:10.7551/mitpress/8470.001.0001.
- [69] Michael Tomasello (2016): *A natural history of human morality*. Harvard University Press, doi:10.4159/9780674915855.



# Understanding transfinite eliminations of non-best replies

Stephan Jagau

IMBS, University of California, Irvine

In auction theory, industrial organization, and other fields of game theory, it is often convenient to let infinite strategy sets stand in for large finite strategy sets. A tacit assumption is that results from infinite games generally translate back to their finite counterparts. Transfinite eliminations of non-best replies pose a radical challenge here, suggesting that common belief in rationality in infinite games strictly refines up to  $k$ -fold belief in rationality for all finite  $k$ . I provide a general characterization of common belief in rationality for finite and infinite games that fully restores the equivalence to up to  $k$ -fold belief in rationality for all finite  $k$ . By means of eliminating non-best replies and supporting beliefs, my characterization entirely avoids transfinite eliminations. Hence, rather than revealing new depths of reasoning, transfinite eliminations signal an inadequacy of eliminating non-best replies as a general description of strategic rationality.



# Persuading Communicating Voters

Toygar Kerman \*

t.kerman@maastrichtuniversity.nl

Anastas P. Tenev<sup>†</sup>

ap.tenev@maastrichtuniversity.nl

This paper studies a multiple-receiver Bayesian persuasion model, where a sender communicates with receivers who have homogeneous beliefs and aligned preferences. The sender wants to implement a proposal and commits to a communication strategy which sends private (possibly) correlated messages to the receivers, who are in an exogenous and commonly known network. Receivers can observe their neighbors' private messages and after updating their beliefs, vote sincerely on the proposal. We examine how networks of shared information affect the sender's gain from persuasion and find that in many cases it is not restricted by the additional information provided by the receivers' neighborhoods. Perhaps surprisingly, the sender's gain from persuasion is not monotonically decreasing with the density of the network.

---

\*Department of Microeconomics and Public Economics (MPE), Maastricht University

<sup>†</sup>Department of Microeconomics and Public Economics (MPE), Maastricht University



# Knowing How to Plan

Yanjun Li

College of Philosophy,  
Nankai University, Tianjin, China

Yanjing Wang\*

Department of Philosophy,  
Peking University, Beijing, China

Various *planning-based know-how logics* have been studied in the recent literature. In this paper, we use such a logic to do *know-how-based planning* via model checking. In particular, we can handle the *higher-order epistemic planning* involving know-how formulas as the goal, e.g., find a plan to make sure  $p$  such that the adversary does not know how to make  $p$  false in the future. We give a PTIME algorithm for the model checking problem over finite epistemic transition systems and axiomatize the logic under the assumption of perfect recall.

## 1 Introduction

Standard Epistemic Logic (EL) mainly studies reasoning patterns of *knowing that*  $\varphi$ , despite early contributions by Hintikka on formulating other *know-wh* expressions such as *knowing who* and *why* using first-order and higher-order modal logic. In recent years, there is a resurgence of interest on epistemic logics of *know-wh* powered by the new techniques for fragments of first-order modal logic based on the so-called *bundle modalities* packing a quantifier and a normal epistemic modality together [26, 24, 21]. Within the varieties of logics of *know-wh*, the logics of know-how received the most attention in AI (cf. e.g., [25, 8, 20, 16]).

Besides the inspirations from philosophy and linguistics (e.g., [23]), the idea of *automated planning* in AI also plays an important role in the developments of various logics of know-how. The core idea is (roughly) to interpret *knowing how to achieve*  $\varphi$  as a *de re* statement of knowledge: “*there is a plan such that you know that this plan will ensure*  $\varphi$ ”. Here, depending on the exact notion of *plans* and how much we want to “ensure” that the plan works, there can be different semantics based on ideas from conformant planning and contingent planning in AI. However, as shown by Li and Wang [16], there is a logic core which is independent from the exact notions of plans underlying all these notions of know-how. We can actually unify different planning-based approaches in a powerful framework.

In this paper, we show that the connection between planning and know-how is not merely one way in terms of *planning-based know-how*, it also makes perfect sense to do *know-how-based planning* which generalizes the notion of planning to incorporate know-how based goals to be explained in the next subsection. As observed in [16], the typical epistemic planning problem given explicit models can be viewed as model checking problems in our framework with the know-how modality  $K_h$  in the language. For example,  $K_{h_1}(K_2\varphi \wedge \neg K_3K_2\varphi)$  captures the epistemic planning problem for agent 1 to ensure that agent 2 knows that  $\varphi$  and keep it a secret from agent 3. Such epistemic planning can also be done in other epistemic approaches, for example, by using dynamic epistemic logic (cf. e.g., [3, 5, 4, 7]). However, we can do much more with the  $K_h$  modality in hand, as explained below.

---

\*Corresponding author. Yanjing Wang thanks the support of NSSF grant 19BZX135.

## 1.1 Higher-order epistemic planning

A distinct feature of modal logic is that we can bring some notions of the meta-language to the object-language. This is not merely formalizing the existing meta-language concepts more precisely since the object language can open new possibilities. Consider the following multi-agent epistemic language of know-that and know-how:

$$\varphi ::= \top \mid p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_i\varphi \mid Kh_i\varphi$$

Since we have the know-how operators  $Kh_i$  in the fully compositional logical language, the goal of planning can involve the (Boolean combinations and nestings of) know-how formulas as well, and we call such planning problems **higher-order epistemic planning**, i.e., *planning about planning*. By *higher-order* we do not mean higher-order logic or higher-order epistemic goals in terms of nested know-that formulas.

Although in the single-agent scenario,  $Kh_1Kh_1\varphi$  is equivalent to  $Kh_1\varphi$  under reasonable conditions as shown in [8, 16], it still makes sense to do higher-order planning. For example, the planning problem to achieve  $\varphi$  with the future control to turn it off is expressed by  $Kh_1(\varphi \wedge Kh_1\neg\varphi)$ , which is not reducible to a formula without the nesting of  $Kh_1$ .<sup>1</sup>

In a multi-agent setting, given that different agent may have different abilities, the importance of higher-order planning is self-evident. Actually, it is a characteristic human instinct to plan in such a higher-order way. A one-year-old baby girl may not know how to open a bottle but she knows how to use her parents' knowledge-how to achieve her goal.<sup>2</sup> As a more concrete example about academic collaborations, it often happens that both researchers do not know how to prove a theorem independently, but one knows how to show a critical lemma which can simplify the original problem and thus enable the other to prove the final theorem using her expertise about the simplified statement. In our language, the situation can be expressed as  $\neg Kh_1\varphi \wedge \neg Kh_2\varphi \wedge Kh_1Kh_2\varphi$ . Of course, it depends on the cooperation of agent 2 to finally ensure the goal  $\varphi$  of agent 1. The nesting of know-how can go arbitrarily deep and also be interactive such as  $Kh_1Kh_2Kh_1Kh_2\varphi$ . Such planning based on others' knowledge-how is at the core of the arts of leadership and management in general.

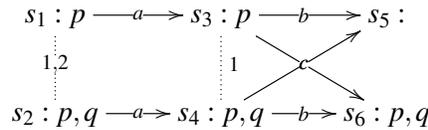
Higher-order planning also makes perfect sense in non-cooperative settings. For example, it is sometimes important to ensure not only the goal  $\varphi$  but also that the adversary cannot spoil it in the future. This is expressed by  $Kh_1(\varphi \wedge \neg Kh_2\neg\varphi)$  in our language of knowing how. It is also interesting to have mixed goals with both  $K$  and  $Kh$ , such as showing a commitment by achieving  $\varphi$  while letting the other know you will not be able to change it afterwards:  $Kh_1K_2(\varphi \wedge \neg Kh_1\neg\varphi)$ . In the more interactive game-like scenarios,  $Kh_1\neg Kh_2\neg Kh_1\varphi$  describe a winning strategy-like plan. We will come back to the related work using *Alternating-time Temporal Logic* (ATL)-like logics at the end of this introduction. Let us first see a concrete example.

**Example 1** Suppose a patient is experiencing some rare symptom  $p$ . To know the cause ( $q$  or  $\neg q$ ), Doctor 1 in the local hospital suggests the patient first take a special fMRI scan ( $a$ ) which is only available in the local hospital at the moment, and bring the result to Doctor 2 in a hospital in another city, who is more experienced in examining the result from the scanner. Then Doctor 2

<sup>1</sup>In contrast to normal modal logic,  $(Kh_i\varphi \wedge Kh_i\psi) \leftrightarrow Kh_i(\varphi \wedge \psi)$  is *invalid* thus  $Kh_iKh_i\varphi \leftrightarrow Kh_i\varphi$  cannot be applied, cf. e.g., [16].

<sup>2</sup>As new fathers, both authors know this well and sometimes it can be tiring to be a fully cooperative parent. At the same time, the parents are eager to know how to let their children acquire relevant knowledge-how as early as possible.

will then know the cause, depending on which different treatments ( $b$  or  $c$ ) should be performed by Doctor 2 to cure the patient. Intuitively, Doctor 1 knows how to let Doctor 2 know how to cure the patient although neither Doctor 1 nor Doctor 2 knows how to cure it without the help of the other. The situation is depicted below as a model with both epistemic relations (dotted lines labelled by 1,2) and action relations (solid lines labelled by  $a, b, c$ ). Note that only 1 can execute action  $a$  and only 2 can perform  $b$  or  $c$ . Reflexive epistemic arrows are omitted.



We would like to have  $\neg Kh_1 \neg p \wedge \neg Kh_2 \neg p \wedge Kh_1((K_2 q \vee K_2 \neg q) \wedge Kh_2 \neg p)$  hold at  $s_1$  according to the semantics to be introduced.

In this paper, to facilitate such higher-order planning using model checking, we extend the single-agent framework in [16] to a genuine multi-agent setting and study the model checking complexity in details. Technical and computational complications arise due to the introduction of multiple agents with different abilities and knowledge to base their own plans. To stay as close as the intended application of automated planning, we will restrict ourselves to finite models with perfect recall.<sup>3</sup> The model checking algorithm is based on a correspondence between execution trees in the models and the syntactically specified knowledge-based plans. It turns out that this correspondence result also enables us to significantly simplify the proofs of the soundness and completeness of the proof system compared with the method in [8, 16].

We summarize our contributions below.

- A multi-agent framework of logic of knowing how, which can handle the diversity of agents' abilities and formally specified plans based on each agent's own knowledge.
- A PTIME model checking algorithm for the full language.
- A complete axiomatization over finite models with perfect recall to ensure the semantics works as desired.<sup>4</sup>

## 1.2 Related work

Doing automated planning using model checking techniques is not a new idea (cf. e.g., [10]). The most distinctive feature of our framework, compared to other approaches to (epistemic) planning such as [5], is that we can express the planning problem in the object language and incorporate higher-order epistemic planning. Following [9], we base our framework on explicit-state models to lay out the theoretical foundation, instead of using more practical compact representations of models and actions which can generate the explicit-state models with certain properties (cf. [15, 16] for the discussions on the technical connections of the two). In contrast to the informal, model-dependent notions of plans such as policies, we use a plan specification language to formalize knowledge-based plans inspired by [13, 14, 27], but with more expressive

<sup>3</sup>cf. [16] for the discussions on the implicit assumption of perfect recall in epistemic planning and the technical difficulty of capturing it by axioms.

<sup>4</sup>Axiomatizations can also help us to do abstract syntactic reasoning about know-how without fixing a model.

branching conditions and more general constructors. Instead of the general structural operational semantics in [16], we give a more intuitive semantics for our knowledge-based plans in this paper.

Comparing to the coalition-based know-how approaches such as [20], we do not have any group know-how modality, but use a much more general notion of plans than the single-step joint action there. Besides the axiomatizations, which are the main focus in the previous work on the logic of know-how, we study the model checking complexity that is crucial for the application of our work to planning. A *second-order know-how modality*  $H_C^D$  was proposed in [19], where  $H_C^D \varphi$  says the coalition  $C$  knows how coalition  $D$  can ensure  $\varphi$  by a single-step joint action, though  $D$  may not know how themselves. Note that  $H_{\{i\}}^{\{j\}} \varphi$  is neither  $\text{Kh}_i \text{Kh}_j \varphi$  nor  $\text{K}_i \text{Kh}_j \varphi$  in our setting, where  $i$  may not know the plan that others may use to reach the goal.

The  $\text{Kh}_i$  operator is clearly also related to the strategy-based *de re* variants of the (single-agent) ATL operator  $\langle\langle i \rangle\rangle$  under imperfect information (cf. e.g., [22, 12]). In the setting of concurrent games of ATL, the strategies are functions from finite histories (or epistemic equivalence classes of them) to available actions, which induce the resulting plays of the given game that are usually infinite histories, on which temporal properties can be verified. This “global” notion of strategies leads to the problem of revocability of the strategies when nesting the  $\langle\langle i \rangle\rangle$  operators: can you change the strategy when evaluating  $\langle\langle i \rangle\rangle \varphi$  in the scope of another such operator [11, 1]? In contrast, it is crucial that our witness plans for know-how are not global and always terminate, and there is no controversy in the nested form of  $\text{Kh}_i$  operators in our setting. It is also crucial that the plans are executed sequentially in our setting, e.g.,  $\text{Kh}_i \text{Kh}_j \varphi$  simply means agent  $i$  knows how to make sure agent  $j$  knows how to make sure  $\varphi$  *after agent  $i$  finishes executing the witness plan for its know-how*. Moreover, while the *de re* variants of ATL are useful in reasoning about (imperfect information) concurrent games (cf. e.g., [18, 17]), our framework is closer to the standard automated planning in AI over transition systems. We leave it to the full version of the paper for a detailed technical comparison.

**Structure of the paper** In the rest of the paper, we first introduce briefly an epistemic language in Section 2 to be used to specify the branching conditions of our plan specification language in Section 3. In Section 4 we introduce the semantics of our know-how language and give a PTIME model checking algorithm. Finally, we obtain a complete axiomatization in Section 5 and conclude with future directions in Section 6.

Due to the strict space limitation, we have to omit some proofs.

## 2 Preliminaries

To specify the knowledge-based plans, we need the following formal language EAL to express the knowledge conditions.

**Definition 2 (EAL Language)** *Let  $\mathbf{P}$  be a set of propositional letters,  $\mathbf{I}$  be a set of agents, and  $\mathbf{A}$  be a set of atomic actions. The language of Epistemic Action Logic ( $\text{EAL}_1^{\mathbf{A}}$ ) is defined as follows:*

$$\varphi ::= \top \mid p \mid \neg\varphi \mid (\varphi \wedge \psi) \mid \text{K}_i \varphi \mid [a] \varphi$$

where  $p \in \mathbf{P}$ ,  $i \in \mathbf{I}$ , and  $a \in \mathbf{A}$ .

The auxiliary connectives and modalities  $\rightarrow, \vee, \langle a \rangle$  are defined as abbreviations as usual.

The semantics of EAL formulas is defined on finite models with both epistemic relations and transitions labelled by atomic actions, which will also be used for the know-how logic. In contrast with the single-agent model of [16], for each agent  $i$  we have a special set of actions  $\mathbf{A}_i$  that are executable by  $i$ . Note that an action may be executable by multiple agents. Moreover, we implicitly assume each agent is aware of all the actions, even for those not executable by itself.

**Definition 3 (Model)** An Epistemic Transition System (i.e., a model)  $\mathcal{M}$  is a tuple

$$\langle W, \{\sim_i \mid i \in \mathbf{I}\}, \{\mathbf{A}_i \mid i \in \mathbf{I}\}, \{Q(a) \mid a \in \bigcup_{i \in \mathbf{I}} \mathbf{A}_i\}, V \rangle$$

where:  $W$  is a non-empty set;  $\mathbf{A}_i$  is a set of actions for each  $i \in \mathbf{I}$ ;  $\sim_i \subseteq W \times W$  is an equivalence relation for each  $i \in \mathbf{I}$ ;  $Q(a) \subseteq W \times W$  is a binary relation for each  $a \in \bigcup_{i \in \mathbf{I}} \mathbf{A}_i$ ;  $V : W \rightarrow 2^{\mathbf{P}}$  is a valuation. A frame is a model without the valuation function.

**Notations** In the rest of the paper, we use  $\mathbf{A}$  to denote  $\bigcup_{i \in \mathbf{I}} \mathbf{A}_i$  for notational simplicity. We use  $[s]^i$  to denote the set  $\{t \in W \mid s \sim_i t\}$  and  $[W]^i$  to denote the set  $\{[s]^i \mid s \in W\}$ . We sometimes call the equivalence class  $[s]^i$  an  $i$ -belief-state. A model captures the agents' abilities to act and the uncertainty about the states.

$EAL_1^{\mathbf{A}}$  formulas are given truth values on pointed epistemic transition systems.

**Definition 4 (Semantics)** Given a pointed model  $(\mathcal{M}, s)$  where  $\mathcal{M} = \langle W, \{\sim_i \mid i \in \mathbf{I}\}, \{\mathbf{A}_i \mid i \in \mathbf{I}\}, \{Q(a) \mid a \in \mathbf{A}\}, V \rangle$  and a formula  $\varphi \in EAL_1^{\mathbf{A}}$ , the semantics is given below:

$\mathcal{M}, s \models \top$	$\Leftrightarrow$	always
$\mathcal{M}, s \models p$	$\Leftrightarrow$	$s \in \mathcal{V}(p)$
$\mathcal{M}, s \models \neg \varphi$	$\Leftrightarrow$	$\mathcal{M}, s \not\models \varphi$
$\mathcal{M}, s \models (\varphi \wedge \psi)$	$\Leftrightarrow$	$\mathcal{M}, s \models \varphi$ and $\mathcal{M}, s \models \psi$
$\mathcal{M}, s \models K_i \varphi$	$\Leftrightarrow$	for all $t$ , if $s \sim_i t$ then $\mathcal{M}, t \models \varphi$
$\mathcal{M}, s \models [a] \varphi$	$\Leftrightarrow$	for all $t$ , if $(s, t) \in Q(a)$ then $\mathcal{M}, t \models \varphi$

A formula is valid on a frame if it is true on all the pointed models based on that frame.

Let  $X$  be a set of states. We sometimes use  $\mathcal{M}, X \models \varphi$  to denote that  $\mathcal{M}, s \models \varphi$  for all  $s \in X$ . Thus  $\mathcal{M}, X \not\models \varphi$  denotes that  $\mathcal{M}, s \not\models \varphi$  for some  $s \in X$ , e.g., we may write  $\mathcal{M}, [s]^i \models K_i \varphi$  or  $\mathcal{M}, [s]^i \not\models K_i \varphi$ .

Here are some standard results about bisimulation that will be useful in the later technical discussion (cf. e.g., [2]).

**Definition 5 (Bisimulation)** Given a model  $\mathcal{M} = \langle W, \{\sim_i \mid i \in \mathbf{I}\}, \{\mathbf{A}_i \mid i \in \mathbf{I}\}, \{Q(a) \mid a \in \bigcup_{i \in \mathbf{I}} \mathbf{A}_i\}, V \rangle$ , a non-empty symmetric binary relation  $Z$  on the domain of  $\mathcal{M}$  is called a bisimulation iff whenever  $(w, v) \in Z$  the following conditions are satisfied:

- (1). For any  $p \in \mathbf{P} : p \in V(w) \iff p \in V(v)$ .
- (2). for any  $i \in \mathbf{I}$ , if  $w \sim_i w'$  for some  $w' \in W$  then there exists a  $v' \in W$  such that  $v \sim_i v'$  and  $w'Zv'$ .
- (3). for any  $a \in \bigcup_{i \in \mathbf{I}} \mathbf{A}_i$ , if  $(w, w') \in Q(a)$  for some  $w' \in W$  then there exists a  $v' \in W$  such that  $(v, v') \in Q(a)$  and  $w'Zv'$ .

The pointed models  $(\mathcal{M}, w)$  and  $(\mathcal{M}, v)$  are said to be bisimilar  $(\mathcal{M}, w \dot{\sim} \mathcal{M}, v)$  if there is a bisimulation  $Z$  such that  $(w, v) \in Z$ .

We use  $\mathcal{M}, [s]^i \Leftrightarrow \mathcal{M}, [s']^i$  to denote that for each  $t' \in [s']^i$ , there exists  $t \in [s]^i$  such that  $\mathcal{M}, t \Leftrightarrow \mathcal{M}, t'$  and vice versa. Given two pointed models  $(\mathcal{M}, s)$  and  $(\mathcal{M}, s')$ , we use  $\mathcal{M}, s \equiv_{EAL_{\mathbf{I}}^A} \mathcal{M}, s'$  to denote that for each  $\varphi \in EAL_{\mathbf{I}}^A$ ,  $\mathcal{M}, s \models \varphi$  if and only if  $\mathcal{M}, s' \models \varphi$ . Similarly,  $\mathcal{M}, s \equiv_{EAL_{\mathbf{I}}^A | K_i} \mathcal{M}, s'$  denotes that for each  $K_i \varphi \in EAL_{\mathbf{I}}^A$ ,  $\mathcal{M}, s \models K_i \varphi$  if and only if  $\mathcal{M}, s' \models K_i \varphi$ .

Please note that  $\mathcal{M}, s \Leftrightarrow \mathcal{M}, s'$  implies  $\mathcal{M}, [s]^i \Leftrightarrow \mathcal{M}, [s']^i$ , but the other way around does not work in general.

As an adaption of the Hennessy-Milner theorem for modal logic (cf. e.g., [2]), we have:

**Proposition 6** *Given a finite model  $\mathcal{M}$ , for each state  $s$  in  $\mathcal{M}$ , we have that  $\mathcal{M}, s \equiv_{EAL_{\mathbf{I}}^A} \mathcal{M}, s'$  iff  $\mathcal{M}, s \Leftrightarrow \mathcal{M}, s'$ .*

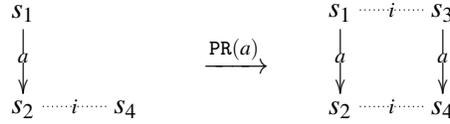
Moreover, if we only look at the  $K_i \varphi$  formulas, we have:

**Proposition 7** *Given a finite model  $\mathcal{M}$ , for each state  $s$  in  $\mathcal{M}$ , we have that  $\mathcal{M}, [s]^i \Leftrightarrow \mathcal{M}, [s']^i$  iff  $\mathcal{M}, s \equiv_{EAL_{\mathbf{I}}^A | K_i} \mathcal{M}, s'$ .*

Most approaches of epistemic planning implicitly assume the property of perfect recall (cf. [16, Sec. 6.5]), here we will also consider this extra property which will play a role in model checking.

**Definition 8 (Perfect recall)** *Given a model  $\mathcal{M}$ , for each  $i \in \mathbf{I}$  and each  $a \in \mathbf{A}_i$ , if  $(s_1, s_2) \in Q(a)$  and  $s_2 \sim_i s_4$  then there must exist a state  $s_3$  such that  $s_1 \sim_i s_3$  and  $(s_3, s_4) \in Q(a)$ .*

In words, if the agent cannot distinguish two states after doing  $a$  then it could not distinguish them before. Perfect recall (PR) corresponds to the following property depicted below:



### 3 A specification language for knowledge-based plans

Inspired by [13], we only consider knowledge-based plans for each agent. We introduce the following specification language which is a multi-agent variant of a fragment of the programming language introduced in [16].

**Definition 9 (Knowledge-based specification language KBP( $i$ ))** *Based on  $EAL_{\mathbf{I}}^A$ , KBP( $i$ ) is defined as follows:*

$$\pi ::= \varepsilon \mid a \mid (\pi; \pi) \mid \text{if } K_i \varphi \text{ then } \pi \text{ else } \pi$$

where  $\varepsilon$  is the empty plan,  $a \in \mathbf{A}_i$  and  $\varphi \in EAL_{\mathbf{I}}^A$ .

Note that the atomic action  $a$  in the above definition must belong to  $i$ , but the condition  $K_i \varphi$  can contain epistemic modalities and actions of other agents, e.g.,  $K_i \neg K_j [b] p$  is a legitimate condition for a knowledge-based plan for  $i$  even when  $b \notin \mathbf{A}_i$ .

We will abbreviate  $\text{if } K_i \varphi \text{ then } \pi_1 \text{ else } \pi_2$  as  $K_i \varphi ? \pi_1 : \pi_2$ . We use  $\pi^n$  to denote the program  $\underbrace{\pi; \dots; \pi}_n$ , and  $\pi^0$  is the empty program  $\varepsilon$ .

Let  $\mathcal{M}$  be an epistemic transition system,  $s$  be a state in  $\mathcal{M}$ , and  $\pi$  be a program in  $\text{KBP}(i)$ . The state set  $Q(\pi)(s)$ , which is the set of states on which executing  $\pi$  on  $s$  will terminate, is defined by induction on  $\pi$  as follows:

$$\begin{aligned} Q(\varepsilon)(s) &= \{s\} \\ Q(a)(s) &= \{t \mid (s, t) \in Q(a)\} \\ Q(\pi_1; \pi_2)(s) &= \{v \mid \text{there is } t \in Q(\pi_1)(s) : v \in Q(\pi_2)(t)\} \\ Q(\text{K}_i\varphi? \pi_1 : \pi_2)(s) &= \begin{cases} Q(\pi_1)(s) & \mathcal{M}, s \models \text{K}_i\varphi \\ Q(\pi_2)(s) & \mathcal{M}, s \not\models \text{K}_i\varphi \end{cases} \end{aligned}$$

We use  $Q(\pi)([s]^i)$  to denote the set  $\bigcup_{s' \in [s]^i} Q(\pi)(s')$ . Given the property of perfect recall, the resulting states of a knowledge-based plan have a nice property.

**Proposition 10** *If  $\mathcal{M}$  has perfect recall and  $\pi \in \text{KBP}(i)$ , then  $Q(\pi)([s]^i)$  is closed over  $\sim_i$ .*

Next we will define a notion of strong executability which is a generalized version of the one in [25].

**Definition 11 (Strong executability)** *We define a program  $\pi$  to be strongly executable on a pointed model  $(\mathcal{M}, s)$  by induction on  $\pi$  as follows: (1)  $\varepsilon$  is always strongly executable on  $\mathcal{M}, s$ ; (2)  $a$  is strongly executable on  $\mathcal{M}, s$  if  $Q(a)(s) \neq \emptyset$ ; (3)  $\pi_1; \pi_2$  is strongly executable on  $\mathcal{M}, s$  if  $\pi_1$  is strongly executable on  $\mathcal{M}, s$  and  $\pi_2$  is strongly executable on each  $t \in Q(\pi_1)(s)$ ; (4)  $\text{K}_i\varphi? \pi_1 : \pi_2$  is strongly executable on  $\mathcal{M}, s$  if either  $\mathcal{M}, s \models \text{K}_i\varphi$  and  $\pi_1$  is strongly executable on  $\mathcal{M}, s$ , or  $\mathcal{M}, s \not\models \text{K}_i\varphi$  and  $\pi_2$  is strongly executable on  $\mathcal{M}, s$ . We say that  $\pi$  is strongly executable on a set  $X$  of states if it is strongly executable on  $\mathcal{M}, s$  for each  $s \in X$ .*

To define the model checking algorithm later, we need to construct the tree of possible executions of a plan where each node is essentially an epistemic equivalence class (usually called a *belief state* in automated planning).

**Definition 12 (Execution tree)** *Given a model  $\mathcal{M}$  and an agent  $i \in \mathbf{I}$ , an execution tree of  $\mathcal{M}$  for  $i$  is a labeled tree  $\mathcal{T} = \langle N, E, L \rangle$ , where  $\langle N, E \rangle$  is a tree consisting of a nonempty set  $N$  of nodes and a set  $E$  of edges and  $L$  is a label function that labels each node in  $N$  with an  $i$ -belief-state and each edge in  $E$  with an action  $a \in \mathbf{A}_i$ , such that, for all  $(n, m) \in E$ , (1)  $L(n, m)$  is strongly executable on  $L(n)$ ; (2)  $L(n, m) = L(n', m')$  if  $(n', m') \in E$  and  $n = n'$ ; (3) there exists  $t \in L(m)$  such that  $t \in Q(L(n, m))(L(n))$ ; (4) for each  $t \in Q(L(n, m))(L(n))$ , there exists  $k \in N$  such that  $(n, k) \in E$  and  $t \in L(k)$ .*

We use  $r^{\mathcal{T}}$  (or simply  $r$ ) to refer to the root node of the execution tree  $\mathcal{T}$ . The following two propositions show the relations between execution trees and knowledge-based plans.

**Proposition 13** *Given a model  $\mathcal{M}$  and a program  $\pi \in \text{KBP}(i)$ , if  $\pi$  is strongly executable on  $[s]^i$ , then there is a finite execution tree  $\langle N, E, L \rangle$  of  $\mathcal{M}$  for  $i$  such that  $L(r) = [s]^i$  and that for each leaf node  $k$ ,  $L(k) \subseteq Q(\pi)([s]^i)$ .*

**Proposition 14** *Let  $\mathcal{T} = \langle N, E, L \rangle$  be a finite execution tree of  $\mathcal{M}$  for  $i$ . There exists  $\pi \in \text{KBP}(i)$  such that  $\pi$  is strongly executable on  $L(r)$  and that for each  $t \in Q(\pi)(L(r))$ , there exists a leaf node  $k$  such that  $\mathcal{M}, [t]^i \Leftrightarrow \mathcal{M}, L(k)$ .*

## 4 Model checking knowledge-how

Extending the work of [16], we introduce the multi-agent epistemic language of know-how and its semantics and study the model checking complexity in detail.

**Definition 15 (ELKh Language)** Let  $\mathbf{P}$  be a set of variables and  $\mathbf{I}$  be a set of agents, the language of epistemic logic with the know-how operator (ELKh) is defined as follows:

$$\varphi ::= \top \mid p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_i\varphi \mid \text{Kh}_i\varphi$$

where  $p \in \mathbf{P}$  and  $i \in \mathbf{I}$ .

**Definition 16 (Semantics)** The truth conditions for basic propositions, Boolean connectives and the epistemic operator  $K_i$  are as before in the case of EAL.

$\mathcal{M}, s \models \text{Kh}_i\varphi \iff$ there is a plan $\pi \in \text{KBP}(i)$ such that <ol style="list-style-type: none"> <li>1. <math>\pi</math> is strongly executable on <math>[s]^i</math>;</li> <li>2. <math>\mathcal{M}, t \models \varphi</math> for each <math>t \in Q(\pi)([s]^i)</math>.</li> </ol>
---

The readers can go back to Example 1.1 to verify the truth of the formula mentioned there.

We can show that ELKh is invariant under bisimulation defined in Definition 5 which also match labelled transitions (cf. [16]).

**Proposition 17** If  $\mathcal{M}, s \leftrightarrow \mathcal{M}, u$  then  $\mathcal{M}, s \models \varphi$  iff  $\mathcal{M}, u \models \varphi$  for all  $\varphi \in \text{ELKh}$ .

In the remainder of this section, we will propose a model checking algorithm for ELKh. The key part of the model checking algorithm is to check the  $\text{Kh}_i$ -formulas, and the problem of whether  $\mathcal{M}, s \models \text{Kh}_i\varphi$  depends on whether there is a good plan for  $\text{Kh}_i\varphi$ . Our strategy is to reduce the problem of whether there is a good plan for  $\text{Kh}_i\varphi$  to the nondeterministic planning problem in [6], which we will briefly introduce below.

**Definition 18 (Planning problem [6])** Let  $T = \langle D, \mathbf{A}, \{\overset{a}{\rightarrow} \mid a \in \mathbf{A}\} \rangle$  be a labeled transition system, where  $D$  is a finite set,  $\mathbf{A}$  is a set of actions, and  $\overset{a}{\rightarrow} \subseteq D \times D$  is a binary relation. A fully observable nondeterministic planning problem is a tuple  $P = \langle T, s, G \rangle$  where  $s \in D$  is the initial state, and  $G \subseteq D$  is the goal set.

Given a labeled transition system  $T$  (note that there is no epistemic relation), we often write  $(s, t) \in \overset{a}{\rightarrow}$  as  $s \overset{a}{\rightarrow} t$ . The sequence  $s_0 \overset{a_1}{\rightarrow} s_1 \overset{a_2}{\rightarrow} \dots s_n$  is called an execution trace of  $T$  from  $s_0$ . A partial function  $f : D \rightarrow \mathbf{A}$  is called a *policy*. An execution trace  $s_0 \overset{a_1}{\rightarrow} s_1 \overset{a_2}{\rightarrow} \dots s_n$  is induced by  $f$  if  $f(s_k) = a_{k+1}$  for all  $0 \leq k < n$ . It is complete w.r.t.  $f$  if  $s_n$  is not in the domain of  $f$ .

**Definition 19 (Strong plan)** A policy  $f$  is a strong plan for a planning problem  $P = \langle T, s, G \rangle$  if each complete execution trace induced by  $f$  from  $s$  is finite and terminates in  $G$ .

**Proposition 20** Let  $\mathcal{T} = \langle D, \mathbf{A}, \{\overset{a}{\rightarrow} \mid a \in \mathbf{A}\} \rangle$ . The existence of strong plans for a planning problem  $P = \langle T, s, G \rangle$  is in PTIME in the size of  $T$ .

**PROOF** It is shown in [6] that the procedure presented in Algorithm 1 always terminates and is correct for whether  $P$  has a strong plan. Moreover, the loop is executed at most  $|D| + 1$  times. Therefore, the procedure is in PTIME.  $\square$

**Algorithm 1:** A planning procedure

---

```

Procedure StrongPlanExistence ( $T, s, G$ ):
  OldSA := false;
  SA :=  $\emptyset$ ;
  while OldSA  $\neq$  SA and  $s \notin G \cup ST_s(SA)$  do           /*  $ST_s(SA)$  is the state set
     $\{v \in D \mid (v, a) \in SA \text{ for some } a \in \mathbf{A}\}$ . */
    PreI :=  $\{(v, a) \in D \times \mathbf{A} \mid v \notin G \cup ST_s(SA), v \text{ has } a\text{-successors, and all } a\text{-successors of}$ 
       $v \text{ are in } G \cup ST_s(SA)\}$ ;
    OldSA := SA;
    SA := SA  $\cup$  PreI;
  if  $s \in G \cup ST_s(SA)$  then return true;
  return false;

```

---

Now we define the nondeterministic planning problem that corresponds to  $\mathcal{M}, s \models \text{Kh}_i \varphi$ . Since the language of ELKh is a multi-agent one where different agents have different abilities, the corresponding nondeterministic planning problem must take the agent  $i$  into account. This makes the reduction more complex than that of single-agent know-how logics.

Given a model  $\mathcal{M} = \langle W, \{\sim_i \mid i \in \mathbf{I}\}, \{\mathbf{A}_i \mid i \in \mathbf{I}\}, \{Q(a) \mid a \in \mathbf{A}\}, V \rangle$  and an agent  $i$ , a planning problem for  $i$  is  $P(\mathcal{M}, i) = \langle T(\mathcal{M}, i), [w]^i, G \rangle$  where  $[w]^i$  is an  $i$ -belief-state,  $G$  is a set of  $i$ -belief-states, and  $T(\mathcal{M}, i) = \langle D, \mathbf{A}_i, \{\xrightarrow{a} \mid a \in \mathbf{A}_i\} \rangle$  is a labeled transition system where  $D = [W]^i$  and for each  $a \in \mathbf{A}_i$ ,  $[s]^i \xrightarrow{a} [t]^i \iff a$  is strongly executable on  $[s]^i$ , and there exists  $t' \in [t]^i : t' \in Q(a)([s]^i)$ .

We will show that the problem of whether  $\mathcal{M}, s \models \text{Kh}_i \varphi$  indeed can be reduced to a nondeterministic planning problem for  $i$  (Lemma 23). Before that, we need two auxiliary propositions.

**Proposition 21** *Given a model  $\mathcal{M}$  and an agent  $i$ , if a program  $\pi \in \text{KBP}(i)$  is strongly executable on  $[s]^i$ , then the planning problem  $P = \langle T(\mathcal{M}, i), [s]^i, G \rangle$  where  $G = \{[v]^i \mid \mathcal{M}, [v]^i \Leftrightarrow \mathcal{M}, [t]^i \text{ for some } t \in Q(\pi)([s]^i)\}$  has a strong plan.*

**Proposition 22** *Given  $\mathcal{M}$  and  $i$ , if a planning problem for  $i$ ,  $P(\mathcal{M}, i) = \langle T(\mathcal{M}, i), [s]^i, G \rangle$ , has a strong plan, then there exists  $\pi \in \text{KBP}(i)$  such that  $\pi$  is strongly executable on  $[s]^i$  and that for each  $t \in Q(a)([s]^i)$ , there exists  $[v]^i \in G$  such that  $\mathcal{M}, t \Leftrightarrow \mathcal{M}, v$ .*

**Lemma 23** *Given a model  $\mathcal{M}$ , we have that  $\mathcal{M}, s \models \text{Kh}_i \varphi$  iff the planning problem  $P(\mathcal{M}, i) = \langle T(\mathcal{M}, i), [s]^i, G \rangle$  where  $G = \{[t]^i \mid \mathcal{M}, [t]^i \models \text{K}_i \varphi\}$  has a strong plan.*

**PROOF** If  $\mathcal{M}, s \models \text{Kh}_i \varphi$ , it follows that there exists  $\pi \in \text{KBP}(i)$  such that  $\pi$  is strongly executable on  $[s]^i$  and that  $\mathcal{M}, t \models \varphi$  for each  $t \in Q(\pi)([s]^i)$ . By Proposition 10, it follows that  $\mathcal{M}, [t]^i \models \text{K}_i \varphi$  for each  $t \in Q(\pi)([s]^i)$ . Let  $G'$  be the set  $\{[v]^i \mid \mathcal{M}, [v]^i \Leftrightarrow \mathcal{M}, [t]^i \text{ for some } t \in Q(\pi)([s]^i)\}$ . By Proposition 17, we have that  $\mathcal{M}, [v]^i \models \text{K}_i \varphi$  for each  $[v]^i \in G'$ . It follows that  $G' \subseteq G$ . Let  $P'$  be the planning problem  $\langle T(\mathcal{M}, i), [s]^i, G' \rangle$ . By Proposition 21, the planning problem  $P'$  has a strong plan. Since  $G' \subseteq G$ , it follows that the planning problem  $P(\mathcal{M}, i)$  als has a strong plan.

If the planning problem  $P(\mathcal{M}, i)$  has a strong plan, by Proposition 22, it follows that there exists  $\pi \in \text{KBP}(i)$  such that  $\pi$  is strongly executable on  $[s]^i$  and that for each  $t' \in Q(a)([s]^i)$ , there exists  $[t]^i \in G$  such that  $\mathcal{M}, t' \Leftrightarrow \mathcal{M}, t$ . Since  $\mathcal{M}, t \models \varphi$  for each  $[t]^i \in G$ , by Proposition 17, it follows that  $\mathcal{M}, t' \models \varphi$  for each  $t' \in Q(a)([s]^i)$ . We then have that  $\mathcal{M}, s \models \text{Kh}_i \varphi$ .  $\square$

Finally, we are ready to show the upper bound of model checking ELKh.



**Axioms**

TAUT	all axioms of propositional logic
DISTK	$K_i p \wedge K_i(p \rightarrow q) \rightarrow K_i q$
T	$K_i p \rightarrow p$
4	$K_i p \rightarrow K_i K_i p$
5	$\neg K_i p \rightarrow K_i \neg K_i p$
AxKtoKh	$K_i p \rightarrow Kh_i p$
AxKhtoKhK	$Kh_i p \rightarrow Kh_i K_i p$
AxKhtoKKh	$Kh_i p \rightarrow K_i Kh_i p$
AxKhKh	$Kh_i Kh_i p \rightarrow Kh_i p$
AxKhbot	$\neg Kh_i \perp$

**Rules**

MP	$\frac{\varphi, \varphi \rightarrow \psi}{\psi}$	NECK	$\frac{\varphi}{K_i \varphi}$
MONOKh	$\frac{\varphi \xrightarrow{\psi} \psi}{Kh_i \varphi \rightarrow Kh_i \psi}$	SUB	$\frac{K_i \varphi}{\varphi[\psi/p]}$

Table 1: SLKHC

specifically, the validity of Axiom AxKhKh, which was highly non-trivial in [8], follows from the fact that two subsequent execution trees can be combined into one execution tree. (The detailed proof can be found in the following theorem of soundness.) For the completeness, the corresponding execution tree plays an important role in the proof of the truth lemma (Lemma 34).

**Theorem 26 (Soundness)** *The proof system SLKHC is sound over finite models with perfect recall.*

PROOF The validity of axioms T, 4, 5 is due to the standard semantics for  $K_i$ .

The axiom AxKtoKh says if  $p$  is known then you know how to achieve  $p$ . Its validity is guaranteed the empty program. The axiom AxKhtoKhK is valid due to Proposition 10. The axiom AxKhtoKKh is the positive introspection axiom for  $Kh_i$  whose validity is due to the semantics for  $Kh_i$ . The axiom AxKhbot says we cannot guarantee contradiction, which indirectly requires that the plan should terminate at someplace and is guaranteed by the condition (1) of the semantics for  $Kh_i$ .

Finally, we will show the validity of the axiom AxKhKh, that is,  $\mathcal{M}, s \models Kh_i \varphi$  if  $\mathcal{M}, s \models Kh_i Kh_i \varphi$ . If  $\mathcal{M}, s \models Kh_i Kh_i \varphi$ , it follows that there is a program  $\pi \in \text{KBP}(i)$  such that  $\pi$  is strongly executable on  $[s]^i$  and that  $\mathcal{M}, t \models Kh_i \varphi$  for each  $t \in Q(\pi)([s]^i)$ . By Proposition 13, there is a finite execution tree  $\mathcal{T}$  of  $\mathcal{M}$  for  $i$  such that  $L(r) = [s]^i$  and that for each leaf node  $k$ ,  $L(k) \subseteq Q(\pi)([s]^i)$ . It follows that  $\mathcal{M}, L(k) \models Kh_i \varphi$  for each leaf node  $k$ . We then have that for each leaf node  $k$ , there is a program  $\pi_k \in \text{KBP}(i)$  such that  $\pi_k$  is strongly executable on  $L(k)$  and that  $\mathcal{M}, v \models \varphi$  for all  $v \in Q(\pi_k)(L(k))$ . By Proposition 13 again, for each leaf  $k$  of  $\mathcal{T}$ , there is a finite execution tree  $\mathcal{T}_k$  of  $\mathcal{M}$  for  $i$  such that  $L^{\mathcal{T}_k}(r^{\mathcal{T}_k}) = L(k)$  and that for each leaf node  $l$  of  $\mathcal{T}_k$ ,  $L^{\mathcal{T}_k}(l) \subseteq Q(\pi_k)(L(k))$ . Moreover, since  $\mathcal{M}, v \models \varphi$  for all  $v \in Q(\pi_k)(L(k))$ , it follows that  $\mathcal{M}, L^{\mathcal{T}_k}(l) \models K_i \varphi$  for each leaf node  $l$  of  $\mathcal{T}_k$ . We then construct an execution tree  $\mathcal{T}'$  of  $\mathcal{M}$  for  $i$  by extending each leaf node  $k$  of  $\mathcal{T}$  with  $\mathcal{T}_k$ . Since  $\mathcal{T}$  and all  $\mathcal{T}_k$  are finite, it follows that  $\mathcal{T}'$  is finite. Moreover, we have that the root of  $\mathcal{T}'$  is labeled with  $[s]^i$  and  $\mathcal{M}, L^{\mathcal{T}'}(l) \models K_i \varphi$  for each leaf node  $l$  of  $\mathcal{T}'$ . By Proposition 14, there is a program  $\pi' \in \text{KBP}(i)$  such that  $\pi'$  is strongly executable on  $[s]^i$  and that for each  $v \in Q(\pi')([s]^i)$ , there exists a leaf node  $l$  of  $\mathcal{T}'$  such that  $\mathcal{M}, [v]^i \Leftrightarrow \mathcal{M}, L^{\mathcal{T}'}(l)$ . Since  $\mathcal{M}, L^{\mathcal{T}'}(l) \models K_i \varphi$  for each leaf node  $l$  of  $\mathcal{T}'$ , by Proposition 17, we have that  $\mathcal{M}, v \models K_i \varphi$  for all  $v \in Q(\pi')([s]^i)$ . Thus, we have that  $\mathcal{M}, s \models Kh_i \varphi$ .  $\square$

Next, we will show the completeness of  $\text{SLKHC}$ . The key is to show that every consistent formula is satisfiable. We will construct a canonical model that consists of all atoms which are maximal consistent sets with respect to a closure set of formulas and show that every consistent formula is satisfied in the canonical model.

Given a set of formulas  $\Delta$ , let:  $\Delta|_{K_i} = \{K_i\varphi \mid K_i\varphi \in \Delta\}$ ,  $\Delta|_{\neg K_i} = \{\neg K_i\varphi \mid \neg K_i\varphi \in \Delta\}$ ,  $\Delta|_{Kh_i} = \{Kh_i\varphi \mid Kh_i\varphi \in \Delta\}$ ,  $\Delta|_{\neg Kh_i} = \{\neg Kh_i\varphi \mid \neg Kh_i\varphi \in \Delta\}$ . Let  $\Phi$  be a subformula-closed set of  $\text{ELKh}$ -formulas that is finite.

**Definition 27** The closure  $cl(\Phi)$  is defined as:

$$\Phi \cup \{K_i\varphi, \neg K_i\varphi \mid \varphi \in \Phi, i \in \mathbf{I} \text{ occurs in } \Phi\}.$$

If  $\Phi$  is finite, so is  $cl(\Phi)$ . Next, we will use it to build a canonical model with respect to  $\Phi$ .

**Definition 28 (Atom)** We enumerate the formulas in  $cl(\Phi)$  by  $\{\psi_0, \dots, \psi_h\}$  where  $h \in \mathbb{N}$ . The formula set  $\Delta = \{Y_k \mid k \leq h\}$  is an atom of  $cl(\Phi)$  if

- $Y_k = \psi_k$  or  $Y_k = \neg\psi_k$  for each  $\psi_i \in cl(\Phi)$ ;
- $\Delta$  is consistent in  $\text{SLKHC}$ .

The following two propositions are similar with their single-agent versions in [8].

**Proposition 29** Let  $\Delta$  be an atom of  $cl(\Phi)$  and  $K_i\varphi \in cl(\Phi)$ . If  $K_i\varphi \notin \Delta$  then there exists  $\Delta'$  such that  $\Delta'|_{K_i} = \Delta|_{K_i}$  and  $\neg\varphi \in \Delta'$ .

**Proposition 30** Let  $\Delta$  and  $\Delta'$  be two atoms of  $cl(\Phi)$  such that  $\Delta|_{K_i} = \Delta'|_{K_i}$ . We have  $\Delta|_{Kh_i} = \Delta'|_{Kh_i}$ .

**Definition 31 (Canonical model)** Given a subformula-closed set  $\Phi$ , the canonical model  $\mathcal{M}^\Phi = \langle W, \{\sim_i \mid i \in \mathbf{I}\}, \{\mathbf{A}_i \mid i \in \mathbf{I}\}, \{Q(a) \mid a \in \bigcup_{i \in \mathbf{I}} \mathbf{A}_i\}, V \rangle$  is defined as:

- $W = \{\Delta \mid \Delta \text{ is an atom of } cl(\Phi)\}$ ;
- for each  $i \in \mathbf{I}$ ,  $\Delta \sim_i \Gamma \iff \Delta|_{K_i} = \Gamma|_{K_i}$ ;
- for each  $i \in \mathbf{I}$ ,  $\mathbf{A}_i = \{(i, \varphi) \mid Kh_i\varphi \in \Phi\}$ ;
- for each  $(i, \varphi) \in \bigcup \mathbf{A}_i$ , we have that  $(\Delta, \Gamma) \in Q(\varphi) \iff Kh_i\varphi \in \Delta$  and  $K_i\varphi \in \Gamma$ ;
- for each  $p \in \Phi$ ,  $p \in V(\Delta) \iff p \in \Delta$ .

**Proposition 32** If  $\Gamma \in Q(i, \varphi)(\Delta)$  and  $\Gamma' \in [\Gamma]^i$ , then  $\Gamma' \in Q(i, \varphi)([\Delta]^i)$ . In other words, the model  $\mathcal{M}^\Phi$  has perfect recall.

**PROOF** We only need to show that  $\Gamma' \in Q(i, \varphi)(\Delta)$ . Since  $\Gamma \in Q(i, \varphi)(\Delta)$ , it follows that  $Kh_i\varphi \in \Delta$  and  $K_i\varphi \in \Gamma$ . Since  $\Gamma' \in [\Gamma]^i$ , it follows that  $K_i\varphi \in \Gamma'$ . Thus, we have that  $\Gamma' \in Q(i, \varphi)(\Delta)$ .  $\square$

**Proposition 33** Let  $\Delta$  be a state in  $\mathcal{M}^\Phi$  and  $(i, \psi) \in \mathbf{A}_i$  be executable at  $\Delta$ . If  $Kh_i\varphi \in \Delta'$  for all  $\Delta' \in Q(i, \psi)(\Delta)$  then  $Kh_i\varphi \in \Delta$ .

**PROOF** First, we show that  $K_i\psi$  is not consistent with  $\neg Kh_i\varphi$ . Since  $(i, \psi) \in \mathbf{A}_i$  is executable at  $\Delta$ , it follows that  $Kh_i\psi \in \Delta$  and that there is some state in  $Q(i, \psi)(\Delta)$  that contains  $K_i\psi$ . Moreover, it is obvious that  $Kh_i\varphi \in cl(\Phi)$ . Assume that  $K_i\psi$  is consistent with  $\neg Kh_i\varphi$ . By a Lindenbaum argument, there exists an atom  $\Gamma$  of  $cl(\Phi)$  such that  $\{K_i\psi, \neg Kh_i\varphi\} \subseteq \Gamma$ . By the definition of  $\mathcal{M}^\Phi$ , it follows that  $(\Delta, \Gamma) \in Q(i, \psi)$ . Since we know that  $Kh_i\varphi \in \Delta'$  for all  $\Delta' \in Q(i, \psi)(\Delta)$ , it follows

that  $\text{Kh}_i\varphi \in \Gamma$ . It is contradictory with the fact that  $\Gamma$  is consistent. Thus,  $\text{K}_i\psi$  is not consistent with  $\neg\text{Kh}_i\varphi$ . Hence, we have that  $\vdash \text{K}_i\psi \rightarrow \text{Kh}_i\varphi$ .

Since  $\vdash \text{K}_i\psi \rightarrow \text{Kh}_i\varphi$ , it follows by Rule  $\text{MONOKh}$  and Axiom  $\text{AxKhtokhK}$  that  $\vdash \text{Kh}_i\psi \rightarrow \text{Kh}_i\text{Kh}_i\varphi$ . Moreover, it follows by Axiom  $\text{AxKhKh}$  that  $\vdash \text{Kh}_i\psi \rightarrow \text{Kh}_i\varphi$ . Since we have shown that  $\text{Kh}_i\psi \in \Delta$ , we have that  $\text{Kh}_i\varphi \in \Delta$ .  $\square$

**Lemma 34 (Truth Lemma)** *For each  $\varphi \in \text{cl}(\Phi)$ ,  $\mathcal{M}^\Phi, \Delta \models \varphi$  iff  $\varphi \in \Delta$ .*

**PROOF** We prove it by induction on  $\varphi$ . The atomic and Boolean cases are straightforward. The case of  $\text{K}_i\varphi$  can be proved by Proposition 29. Next, we only focus on the case of  $\text{Kh}_i\varphi$ .

**Right to Left:** If  $\text{Kh}_i\varphi \in \Delta$ , we will show  $\mathcal{M}^\Phi, \Delta \models \text{Kh}_i\varphi$ . We first show that  $\text{K}_i\varphi$  is consistent. If not, namely  $\vdash \text{K}_i\varphi \rightarrow \perp$ , it follows by Rule  $\text{MONOKh}$  that  $\vdash \text{Kh}_i\text{K}_i\varphi \rightarrow \text{Kh}_i\perp$ . It follows by Axiom  $\text{AxKhbot}$  that  $\vdash \text{Kh}_i\text{K}_i\varphi \rightarrow \perp$ . Since  $\text{Kh}_i\varphi \in \Delta$ , it follows by Axiom  $\text{AxKhtokhK}$  that  $\Delta \vdash \perp$ , which is in contradiction with the fact that  $\Delta$  is consistent. Therefore,  $\text{K}_i\varphi$  is consistent.

By Lindenbaum's Lemma, there exists an atom  $\Gamma$  such that  $\text{K}_i\varphi \in \Gamma$ . By the definition of  $\mathcal{M}^\Phi$ , it follows that  $(i, \varphi) \in \mathbf{A}_i$  and that  $(\Delta', \Gamma) \in Q(i, \varphi)$  for each  $\Delta' \in [\Delta]^i$ . It means  $(i, \varphi)$  is strongly executable on  $[\Delta]^i$ . For each  $\Theta \in Q(i, \varphi)([\Delta]^i)$ , by the definition of  $\mathcal{M}^\Phi$ , it follows that  $\text{K}_i\varphi \in \Theta$ . By Axiom  $\text{T}$ , it follows that  $\varphi \in \Theta$ . By IH, we then have that  $\mathcal{M}^\Phi, \Theta \models \varphi$  for each  $\Theta \in Q(i, \varphi)([\Delta]^i)$ . Thus, we have that  $\mathcal{M}^\Phi, \Delta \models \text{Kh}_i\varphi$ .

**Left to Right:** Suppose  $\mathcal{M}^\Phi, \Delta \models \text{Kh}_i\varphi$ , we will show  $\text{Kh}_i\varphi \in \Delta$ . Since  $\mathcal{M}^\Phi, \Delta \models \text{Kh}_i\varphi$ , it follows that there exists  $\pi \in \text{KBP}(i)$  such that  $\pi$  is strongly executable on  $[\Delta]^i$  and that  $\mathcal{M}^\Phi, \Gamma \models \varphi$  for all  $\Gamma \in Q(\pi)([\Delta]^i)$ . By IH, we have that  $\varphi \in \Gamma$  for all  $\Gamma \in Q(\pi)([\Delta]^i)$ . Moreover, for each  $\Gamma \in Q(\pi)([\Delta]^i)$ , if  $\Gamma' \in [\Gamma]^i$ , by Proposition 32, it follows that  $\Gamma' \in Q(\pi)([\Delta]^i)$ , and we then have that  $\varphi \in \Gamma'$ . Moreover, with Proposition 29, it is easy to check that  $\text{K}_i\varphi \in \Gamma$  for all  $\Gamma \in Q(\pi)([\Delta]^i)$ .

Since  $\pi$  is strongly executable on  $[\Delta]^i$ , by Proposition 13, there is a finite execution tree  $\mathcal{T}$  of  $\mathcal{M}^\Phi$  for  $i$  such that  $L(r) = [\Delta]^i$  and that  $L(k) \subseteq Q(\pi)([\Delta]^i)$ . Since we have shown that  $\text{K}_i\varphi \in \Gamma$  for all  $\Gamma \in Q(\pi)([\Delta]^i)$ , it follows that if  $k$  is a leaf node then  $\text{K}_i\varphi$  is in all states in  $L(k)$ . Next, we firstly show the following claim:

for each node  $k$  and each  $\Theta \in L(k) : \text{Kh}_i\varphi \in \Theta$ .

Please note that the execution tree  $\mathcal{T}$  is finite. We prove the claim above by induction on the height of nodes. If the height of  $k$  is 0, it means that  $k$  is a leaf node. It follows that  $\text{K}_i\varphi \in \Theta$  for each  $\Theta \in L(k)$ . By Axiom  $\text{AxKtokh}$ , it follows that  $\text{Kh}_i\varphi \in \Theta$  for each  $\Theta \in L(k)$ . With the IH that the claim holds for each node whose height is less than  $h$ , we will show that the claim holds for nodes whose height is  $h \geq 1$ . Given a node  $k$  whose height is  $h$ , since  $h \geq 1$ , it follows that there is a node  $k'$  such that  $(k, k')$  is an edge in  $\mathcal{T}$  and the height of  $k'$  is less than  $h$ , in other words,  $k'$  is a child node of  $k$ . Let  $L(k) = [\Theta]^i$  and  $L(k, k') = (i, \psi)$ . Since  $\mathcal{T}$  is an execution tree, it follows that  $(i, \psi)$  is strongly executable on  $[\Theta]^i$ , and then  $(i, \psi)$  is executable on  $\Theta$ . For each  $\Theta' \in Q(i, \psi)(\Theta) \subseteq Q(L(k, k'))(L(k))$ , by the definition of execution trees, it follows that there is a  $k$ -child  $k''$  such that  $\Theta' \in L(k'')$ . Since  $k''$  is a child of  $k$ , by IH, it follows that  $\text{Kh}_i\varphi \in \Theta'$ . Thus, we have shown that  $\text{Kh}_i\varphi \in \Theta'$  for each  $\Theta' \in Q(i, \psi)(\Theta)$ . By Proposition 33, it follows that  $\text{Kh}_i\varphi \in \Theta$ . Thus, we have shown the claim. Since the root is labeled with  $[\Delta]^i$ , we then have that  $\text{Kh}_i\varphi \in \Delta$ .  $\square$

Let  $\varphi$  be a consistent formula. By Proposition 29, it follows that there is an atom  $\Delta$  of  $\text{cl}(\Phi)$  such that  $\varphi \in \Delta$ , where  $\Phi$  is the set of subformulas of  $\varphi$ . By Proposition 32 and Lemma 34, it follows that  $\varphi$  is satisfiable over finite models with perfect recall. The completeness then follows:

**Theorem 35 (Completeness)** *The proof system  $\text{SLKHIC}$  is complete over finite models with perfect recall.*

From the construction of the canonical model  $\mathcal{M}^\Phi$  in Definition 31, we can see that both the state set  $W$  and the action set  $A$  are bounded. This implies that the logic has a small model property. Moreover, by Theorem 24, the decidability then follows:

**Theorem 36** *The logic  $\text{ELKh}$  is decidable.*

## 6 Conclusion

In this paper, we propose the notion of higher-order epistemic planning by using an epistemic logic of knowing how. The planning problems can be encoded using model checking problems in the framework, which can be computed in PTIME in the size of the the model. We also axiomatize the logic over finite models with perfect recall.

As for future work, besides many theoretical questions about the knowing how logic as it is, we may extend its expressive power to capture conditional knowledge-how, which is very useful in reasoning about planning problems. For example, one may say I know how to achieve  $\varphi$  given  $\psi$ . Note that it is very different from  $\text{Kh}_i(\psi \rightarrow \varphi)$  or  $\text{K}_i\psi \rightarrow \text{Kh}_i\varphi$ . The important difference is that such conditional knowledge-how is global, compared to the current local semantics of  $\text{Kh}_i$ . This is similar to the binary global  $\text{Kh}_i$  operator introduced in [25]. It would be interesting to combine the two notations of know-how in the same framework. On the practical side, we can consider the model checking problems over compact representations of the actions and states using the techniques in [15, 16] connecting the explicit-state models with the compact representations.

## References

- [1] Thomas Ågotnes, Valentin Goranko & Wojciech Jamroga (2007): *Alternating-time temporal logics with irrevocable strategies*. In Dov Samet, editor: *Proceedings of (TARK 2007)*, pp. 15–24, doi:10.1145/1324249.1324256.
- [2] Patrick Blackburn, Maarten de Rijke & Yde Venema (2002): *Modal Logic*. Cambridge University Press, doi:10.1017/CBO9781107050884.
- [3] Thomas Bolander & Mikkel Birkegaard Andersen (2011): *Epistemic planning for single and multi-agent systems*. *Journal of Applied Non-Classical Logics* 21(1), pp. 9–34, doi:10.3166/jancl.21.9-34.
- [4] Thomas Bolander, Tristan Charrier, Sophie Pinchinat & François Schwarzenrüber (2020): *DEL-based epistemic planning: Decidability and complexity*. *Artificial Intelligence* 287, p. 103304, doi:10.1016/j.artint.2020.103304.
- [5] Thomas Bolander, Martin Holm Jensen & François Schwarzenrüber (2015): *Complexity Results in Epistemic Planning*. In: *Proceedings of IJCAI '15*, pp. 2791–2797. Available at <http://ijcai.org/Abstract/15/395>.
- [6] A. Cimatti, M. Pistore, M. Roveri & P. Traverso (2003): *Weak, Strong, and Strong Cyclic Planning via Symbolic Model Checking*. *Artificial Intelligence* 147(1–2), p. 35–84, doi:10.1016/S0004-3702(02)00374-0.
- [7] Thorsten Engesser, Robert Mattmüller, Bernhard Nebel & Michael Thielscher (2021): *Game description language and dynamic epistemic logic compared*. *Artificial Intelligence* 292, p. 103433, doi:10.1016/j.artint.2020.103433.

- [8] Raul Fervari, Andreas Herzig, Yanjun Li & Yanjing Wang (2017): *Strategically knowing how*. In: *Proceedings of IJCAI '17*, pp. 1031–1038, doi:10.24963/ijcai.2017/143.
- [9] Malik Ghallab, Dana Nau & Paolo Traverso (2004): *Automated planning: theory and practice*. Elsevier.
- [10] Fausto Giunchiglia & Paolo Traverso (2000): *Planning as Model Checking*, doi:10.1007/10720246\_1.
- [11] Andreas Herzig (2015): *Logics of knowledge and action: critical analysis and challenges*. *Autonomous Agents and Multi-Agent Systems* 29(5), pp. 719–753, doi:10.1007/s10458-014-9267-z.
- [12] Wojciech Jamroga & Thomas Ågotnes (2007): *Constructive knowledge: what agents can achieve under imperfect information*. *Journal of Applied Non-Classical Logics* 17(4), pp. 423–475, doi:10.3166/jancl.17.423-475.
- [13] Jérôme Lang & Bruno Zanuttini (2012): *Knowledge-Based Programs as Plans - The Complexity of Plan Verification*. In: *ECAI '12*, pp. 504–509, doi:10.3233/978-1-61499-098-7-504.
- [14] Jérôme Lang & Bruno Zanuttini (2013): *Knowledge-Based Programs as Plans : Succinctness and the Complexity of Plan Existence (Extended Abstract)*. In: *TARK '13*. Available at <https://arxiv.org/abs/1310.6429>.
- [15] Yanjun Li & Yanjing Wang (2019): *Multi-agent Knowing How via Multi-step Plans: A Dynamic Epistemic Planning Based Approach*. In: *Proceedings of LORI VII*, pp. 126–139, doi:10.1007/978-3-662-60292-8\_10.
- [16] Yanjun Li & Yanjing Wang (2021): *Planning-Based Knowing How: A Unified Approach*. *Artificial Intelligence* 296, p. 103487, doi:10.1016/j.artint.2021.103487.
- [17] Bastien Maubert, Aniello Murano, Sophie Pinchinat, François Schwarzenrüber & Silvia Stranieri (2020): *Dynamic Epistemic Logic Games with Epistemic Temporal Goals*. In: *Proceedings of ECAI 2020*, IOS Press, pp. 155–162, doi:10.3233/FAIA200088.
- [18] Bastien Maubert, Sophie Pinchinat, François Schwarzenrüber & Silvia Stranieri (2020): *Concurrent Games in Dynamic Epistemic Logic*. In Christian Bessiere, editor: *Proceedings of IJCAI 2020*, ijcai.org, pp. 1877–1883, doi:10.24963/ijcai.2020/260.
- [19] Pavel Naumov & Jia Tao (2018): *Second-Order Know-How Strategies*. In: *Proceedings of AAMAS '18*, pp. 390–398, doi:10.5555/3237383.3237444.
- [20] Pavel Naumov & Jia Tao (2018): *Together we know how to achieve: An epistemic logic of know-how*. *Artificial Intelligence* 262, pp. 279–300, doi:10.1016/j.artint.2018.06.007.
- [21] Anantha Padmanabha, R. Ramanujam & Yanjing Wang (2018): *Bundled fragments of first-order modal logic: (un)decidability*. In: *Proceedings of FSTTCS '18*, doi:10.4230/LIPICs.FSTTCS.2018.43.
- [22] Pierre-Yves Schobbens (2004): *Alternating-time logic with imperfect recall*. *Electron. Notes Theor. Comput. Sci.* 85(2), pp. 82–93, doi:10.1016/S1571-0661(05)82604-0.
- [23] Jason Stanley & Timothy Williamson (2001): *Knowing how*. *The Journal of Philosophy*, pp. 411–444, doi:10.2307/2678403.
- [24] Yanjing Wang (2017): *A New Modal Framework for Epistemic Logic*. In Jérôme Lang, editor: *Proceedings of TARK '17, EPTCS* 251, pp. 515–534, doi:10.4204/EPTCS.251.38.
- [25] Yanjing Wang (2018): *A logic of goal-directed knowing how*. *Synthese* 10, pp. 4419–4439, doi:10.1007/s11229-016-1272-0.
- [26] Yanjing Wang (2018): *Beyond knowing that: a new generation of epistemic logics*. In: *Jaakko Hintikka on knowledge and game theoretical semantics*, Springer, Cham, pp. 499–533, doi:10.1007/978-3-319-62864-6\_21.
- [27] Bruno Zanuttini, Jérôme Lang, Abdallah Saffidine & François Schwarzenrüber (2020): *Knowledge-based programs as succinct policies for partially observable domains*. *Artificial Intelligence* 288, p. 103365, doi:10.1016/j.artint.2020.103365.



# Probabilistic Stability and Statistical Learning

Krzysztof Mierzewski

Carnegie Mellon University

kmierzew@andrew.cmu.edu

A canonical way to bridge the probabilistic, gradational notion of belief studied by Bayesian probability theory with the more mundane, all-or-nothing concept of qualitative belief is in terms of *acceptance rules* [Kelly and Lin, 2012]: maps that specify which propositions a rational agent accepts in light of their numerical credences (given by a probability model). Among the various acceptance rules proposed in the literature, an especially prominent one is Leitgeb’s *stability rule* [Leitgeb, 2013, 2014, 2017; Rott, 2017], based on the notion of *probabilistically stable* hypotheses: that is, hypotheses that maintain sufficiently high probability under conditioning on new information.

When applied to discrete probability spaces, the stability rule for acceptance guarantees logically closed and consistent belief sets, and it suggests a promising account of the relationship between subjective probabilities and qualitative belief. Yet, most natural inductive problems—particularly those commonly occurring in statistical inference—are best modelled with continuous probability distributions and statistical models with a richer internal structure. This paper explores the possibility of extending Leitgeb’s stability rule to more realistic learning scenarios and general probability spaces. This is done by considering a generalised notion of probabilistic stability, in which acceptance depends not only on the underlying probability space, but also on a *learning problem*—namely, a probability space equipped with a distinguished family of events capturing the relevant evidence (e.g., the observable data) in the given learning scenario. This view of acceptance as being relative to an *evidence context* is congenial to (topological approaches to) formal learning theory and hypothesis testing in statistics (where one typically distinguishes the hypotheses being considered from observable sample data), as well as logics of evidence-relative belief [van Benthem and Pacuit, 2011].

Here we consider the case of statistical learning. We show that, in the context of standard (parametric) Bayesian learning models, the stability rule yields a notion of acceptance that is either trivial (only hypotheses with probability 1 are accepted) or fails to be conjunctive (accepted hypotheses are not closed under conjunctions). The first problem chiefly affects statistical hypotheses, while the second one chiefly affects predictive hypotheses about future outcomes. The failure of conjunctivity for the stability rule is particularly salient, as it affects a wide class of consistent Bayesian priors and learning models with exchangeable random variables. In particular, the results presented here apply to many distributions commonly used in statistical inference, as well as to every method in Carnap’s continuum of inductive logics [Carnap, 1980; Skyrms, 1996]. These results highlight a serious tension between (1) being responsive to evidence and (2) having conjunctive beliefs induced by the stability rule. In the statistical context, certain properties of priors that are conducive to inductive learning—open-mindedness, as well as certain symmetries in the agent’s probability assignments—act against conjunctive belief. Thus, the main selling points of the stability account of belief—its good logical behaviour and its close connection to the Lockean thesis—do not survive the passage to richer probability models, such as canonical statistical models for i.i.d learning. We conclude by discussing the consequences the results bear on Leitgeb’s *Humean Thesis* on belief [Leitgeb, 2017].

## References

- J. van Benthem and E. Pacuit. Dynamic Logics of Evidence-Based Beliefs. *Studia Logica*, 99(1): 61–92, 2011. doi:[10.1007/s11225-011-9347-x](https://doi.org/10.1007/s11225-011-9347-x).
- R. Carnap. A Basic System of Inductive Logic. in R.C. Jeffrey (ed.), *Studies in Inductive Logic and Probability*, vol. 2, Berkeley: University of California Press., 1980. doi:[10.1525/9780520334250-003](https://doi.org/10.1525/9780520334250-003)
- K. T. Kelly and H. Lin. A geo-logical solution to the lottery-paradox, with applications to conditional logic. *Synthese*, 186(2): 531–575, 2012. doi:[10.1007/s11229-011-9998-1](https://doi.org/10.1007/s11229-011-9998-1).
- H. Leitgeb. Reducing belief simpliciter to degrees of belief. *Annals of Pure and Applied Logic*, 164: 1338–1389, 2013. doi:[10.1016/j.apal.2013.06.015](https://doi.org/10.1016/j.apal.2013.06.015).
- H. Leitgeb. The Stability Theory of Belief. *Philosophical Review*, 123(2): 131–171, 2014. doi:[10.1215/00318108-2400575](https://doi.org/10.1215/00318108-2400575).
- H. Leitgeb. *The Stability of Belief*. Oxford University Press, Oxford, 2017. doi:[10.1093/acprof:oso/9780198732631.001.0001](https://doi.org/10.1093/acprof:oso/9780198732631.001.0001).
- H. Rott. Stability and Scepticism in the Modelling of Doxastic States: Probabilities and Plain Beliefs. *Minds and Machines*, 27(1): 167–197, 2017. doi:[10.1007/s11023-016-9415-0](https://doi.org/10.1007/s11023-016-9415-0).
- B. Skyrms. Carnapian inductive logic and Bayesian statistics. In Ferguson, T. S., Shapley, L. S. and MacQueen, J. B., editors, *Statistics, probability and game theory: Papers in honor of David Blackwell*, Hayward, CA, Institute of Mathematical Statistics, pages 321–336, 1996. doi:[10.1214/lnms/1215453580](https://doi.org/10.1214/lnms/1215453580).

# Attainable Knowledge and Omniscience

Pavel Naumov

King's College  
Pennsylvania, United States  
pgn2@cornell.edu

Jia Tao

Lafayette College  
Pennsylvania, United States  
taoj@lafayette.edu

The paper investigates an evidence-based semantics for epistemic logics. It is shown that the properties of knowledge obtained from a potentially infinite body of evidence are described by modal logic S5. At the same time, the properties of knowledge obtained from only a finite subset of this body are described by modal logic S4. The main technical result is a sound and complete bi-modal logical system that describes properties of these two modalities and their interplay.

## 1 Introduction

The distinction between *potential* and *actual* infinity goes back to Zeno's paradoxes. The former refers to constructions involving a potentially unlimited, but finite, number of steps while the latter refers to an infinite set as a whole.

There has been a long tradition of distinguishing these two types of infinity in logic. We say that a set is decidable if the membership in the set can be verified through a potentially unlimited but finite number of steps of a Turing machine. Most people believe that Peano Arithmetic is consistent and, thus, its consistency could be verified by observing that none of the countably many finite-length derivations ends with a contradiction. Yet, Gödel's second incompleteness theorem claims that the consistency cannot be established by a single finite-length derivation or a finite set of such derivations. More generally, logicians usually accept admissible inference rules with finitely many assumptions and reject  $\omega$ -rules that deduce statements from infinitely many assumptions. A compactness property for a logical system states that if a statement semantically follows from an infinite set of assumptions, then it semantically follows from a potentially unlimited but finite subset of the assumptions.

In this paper we investigate the difference between the knowledge that can be derived from a potentially unlimited, but finite, subset of the body of evidence and the knowledge that can be inferred from the whole infinite body of evidence. We refer to the first form of knowledge as *attainable knowledge* and to the second as *omniscience*. We show that the properties of the attainable knowledge are captured by modal logic S4, while those of the omniscience are given by modal logic S5.

### 1.1 Epistemic Logics S4 and S5

Epistemic logic S5 is an extension of logic S4 by the Negative Introspection axiom:  $\neg K\phi \rightarrow K\neg K\phi$ , where K is the knowledge modality. Informally, this axiom expresses an agent awareness of the limits of her knowledge: if an agent does not know something, then she must know that she does not know it. Even without taking into account the size of the body of supporting evidence, the validity of this axiom has long been a subject for philosophical discussions. Hintikka [8] rejects this axiom. Stalnaker [15] cites a belief-based argument against it. The argument relies on three standard assumptions about knowledge and beliefs: (i) beliefs of an agent are consistent, (ii) an agent believes in anything she knows, and (iii) anything an agent knows is true. We interpret Stalnaker's argument in terms of a mathematician's belief

about her knowledge. Imagine a mathematician who spent last several years working on a conjecture  $\varphi$ . One day she announces “I believe I now know that the conjecture is true”<sup>1</sup>. We write this as  $BK\varphi$ . Because her beliefs are consistent, we must conclude that she does not believe in the opposite:  $\neg B\neg K\varphi$ . Hence,  $\neg K\neg K\varphi$  by the contraposition of assumption (ii). Thus, by the contrapositive of the Negative Introspection axiom, we have  $K\varphi$ . Then,  $\varphi$  is true by assumption (iii). Following Stalnaker, we have just shown that if a mathematician believes that she knows that conjecture  $\varphi$  is true, then conjecture  $\varphi$  must be true. However, intuitively, this cannot be true because her belief alone should not make the conjecture true.

While the Negative Introspection axiom might not be true as the most general epistemic axiom, Fagin, Halpern, Moses, and Vardi point out that it is “appropriate for many multi-agent systems applications”, such as the analysis of the muddy children puzzle, knowledge bases, games of imperfect information, and message-passing systems [6, p.18]. In spite of its limitations, S5 has been accepted as the standard epistemic logic in the modern literature [1, 2, 5, 11]. There also have been suggestions to consider modal systems in between S4 and S5 [9], [10, p.82].

The purpose of this paper is not to settle the discussion about the merits of S4 vs. S5, but to highlight the difference between these two systems from the viewpoint of an evidence-based semantics. The main technical result of our work is a sound and complete logical system with two modalities that capture individual properties of the attainable knowledge and of the omniscience as well as the properties that connect these two types of knowledge.

## 1.2 Grand Hotel Example

Consider the famous Hilbert Grand Hotel that has infinitely many rooms. By a single observation in this setting we mean examining a room in order to establish whether it is empty. If the hotel has vacancies, then at least one room is empty. By opening just a single door of that specific room one can learn that the hotel has vacancy. We write  $A$ (“the hotel has vacancy”) to express this fact. The modality  $A$  denotes *attainable knowledge* that can be formed from finitely many observations. In fact, in this case a single observation suffices. On the other hand, if the hotel has no vacancy, then the knowledge about this cannot be obtained from examining any finite subset of the rooms:  $\neg A$ (“the hotel has no vacancy”). At the same time, one can learn that the hotel is full by examining all rooms in the hotel. We write this as  $O$ (“the hotel has no vacancy”), where modality  $O$  denotes the *omniscience* that can be formed from a potentially infinite body of evidence.

Let us first show that attainable knowledge does not satisfy the Negative Introspection axiom:

$$\neg A\varphi \rightarrow A\neg A\varphi. \quad (1)$$

Consider the epistemic world  $w$  in which the hotel is full and let  $\varphi$  be statement “the hotel has vacancy”. Thus,  $w \not\models \varphi$ . Hence, one cannot know that  $\varphi$  is true from examining (finitely many) hotel rooms:  $w \not\models A\varphi$ . Hence,  $w \models \neg A\varphi$ . At the same time, informally, the only reason why  $w \models \neg A\varphi$  is true is because the hotel is full. Since the latter cannot be established through finitely many observations,  $w \models \neg A\neg A\varphi$ . More formally, to prove  $w \models \neg A\neg A\varphi$  we need to show that no matter which finite set of evidence is examined, there still will be an epistemic world  $u$  indistinguishable from  $w$  such that  $u \models A\varphi$ . Indeed, suppose that we have chosen to examine rooms  $r_1, \dots, r_n$ . Let  $r$  be any room different from rooms  $r_1, \dots, r_n$ . Let  $u$  be an epistemic world in which all rooms except for room  $r$  are occupied. Thus, epistemic worlds  $w$  and  $u$  are indistinguishable through the chosen finite set of examinations. Yet,  $u \models A\varphi$

<sup>1</sup>In Stalnaker’s words, “agent . . . take[s] herself to know” that the conjecture is true [15, p.177].

because in epistemic world  $u$ , it is enough to examine a single room  $r$  to learn that the hotel has vacancy. This concludes the counterexample for the Negative Introspection axiom (1).

From the classical (non-evidence-based) Kripke semantics point of view, the Negative Introspection axiom captures the fact that the indistinguishability relation on epistemic states is symmetric. In our evidence-based semantics, the indistinguishability by a single evidence is also symmetric. That is, if one can distinguish one state of the hotel from another state by examining just a single room, then one can also distinguish the latter state from the former by examining the same room. The reason why the Negative Introspection axiom fails, under the evidence-based semantics, is because *the indistinguishability relation between a world and a set of worlds by a potentially unlimited, but finite body of evidence is not symmetric*. Indeed, consider the epistemic world in which the hotel is empty and the set of all worlds in which the hotel is not empty. By examining a right single room in a non-empty hotel, one can distinguish it from an empty hotel. Yet, after examining any finite subset of rooms in an empty hotel one cannot distinguish the current world from the set of worlds in which the hotel is not empty.

Let us now consider a weaker form of the Negative Introspection axiom used in logic S4.4 [9]:

$$\varphi \rightarrow (\neg A\varphi \rightarrow A\neg A\varphi). \quad (2)$$

The above counterexample for formula (1) does not work as a counterexample for formula (2) because in that setting  $w \not\models \varphi$ . Nevertheless, principle (2) is not valid in general. Indeed, let us modify the hotel setting by assuming that some rooms in the hotel might have bedbugs. The presence of the bedbugs can be tested when a room is examined. Furthermore, let us assume that once a single room in the hotel becomes infected with bedbugs all guests immediately leave the hotel. Thus, the hotel might have either (a) visitors and no bedbugs, or (b) no visitors and no bedbugs, or (c) bedbugs and no visitors. Consider an epistemic world  $w$  in which the hotel has no visitors and no bugs. Let  $\varphi$  be the statement “the hotel has no visitors”. Thus,  $w \models \varphi$ . Note that there are two ways to verify that the hotel has no visitors: either by examining all rooms to observe that they are all vacant or by examining a single room that contains bedbugs. Since in the epistemic world  $w$  the hotel is not infected with bugs, the only way to verify that the hotel is empty in this world is to examine all rooms. Thus,  $w \models \neg A\varphi$ . To finish the counterexample for formula (2), it suffices to show that  $w \not\models A\neg A\varphi$ . In other words, we need to prove that, after examining a finite set of rooms  $r_1, \dots, r_n$ , we will not be able to distinguish epistemic state  $w$  from such an epistemic state  $u$  that  $u \models A\varphi$ . Indeed, let  $r$  be any room different from rooms  $r_1, \dots, r_n$  and let  $u$  be the epistemic world in which room  $r$  is infected with bedbugs. Note that  $u \models A\varphi$  because in epistemic state  $u$  it is enough to examine the bug-infected room  $r$  to conclude that the hotel is empty.

### 1.3 Formal Semantics of Evidence

To make our Grand Hotel example more formal, we introduce the formal semantics of evidence-based knowledge. Since we view evidence as a way to distinguish epistemic worlds, in this paper we interpret the pieces of evidence as equivalence relations on the worlds. When an agent takes into account several pieces of evidence, equivalence relations corresponding to these pieces intersect to form the equivalence relation of the agent. In the second Grand Hotel example above, each room is in one of the three states: occupied, vacant without bedbugs, and vacant with bedbugs. An epistemic world can be described by specifying the state of each room. The observation that consists of examining room  $r$  can distinguish two epistemic worlds in which room  $r$  is in different states. It cannot distinguish two epistemic worlds in which room  $r$  is in the same state.

In other words, we assemble an agent’s knowledge from several pieces of evidence in a similar way as how distributed knowledge [6] of a group is assembled from the individual observations of the members

of the group. Thus, the logical system developed in this paper could also be used to describe two different forms of group knowledge by an infinite group of agents: modality  $O$  represents the standard distributed knowledge by the whole group and modality  $A$  represents the distributed knowledge by a finite subgroup of the whole group.

The formal evidence-based semantics described above is similar to the one for the budget-constrained knowledge proposed by us earlier [12, 13]. It is different from the neighbourhood semantics of evidence investigated by van Benthem and Pacuit [4] and the probabilistic approach of Halpern and Pucella [7]. It is not clear how and if the results in the current paper could be applied to these alternative semantics.

## 1.4 Logical System

In this paper we propose a sound and complete logical system that describes the universal properties of omniscience modality  $O$  and attainable knowledge modality  $A$ . The axioms involving only modality  $O$  are exactly those forming modal logic S5:

1. Truth:  $O\varphi \rightarrow \varphi$ ,
2. Positive Introspection:  $O\varphi \rightarrow OO\varphi$ ,
3. Negative Introspection:  $\neg O\varphi \rightarrow O\neg O\varphi$ ,
4. Distributivity:  $O(\varphi \rightarrow \psi) \rightarrow (O\varphi \rightarrow O\psi)$ .

Later we do not list the Positive Introspection axiom among axioms of our system because, just like in the case of logic S5, the Positive Introspection axiom is derivable from the other axioms. We prove this in Lemma 2. The axioms involving only modality  $A$  are exactly those forming modal logic S4:

1. Truth:  $A\varphi \rightarrow \varphi$ ,
2. Positive Introspection:  $A\varphi \rightarrow AA\varphi$ ,
3. Distributivity:  $A(\varphi \rightarrow \psi) \rightarrow (A\varphi \rightarrow A\psi)$ .

Finally, there appears to be two independent axioms that capture the interplay of the two modalities:

1. Monotonicity:  $A\varphi \rightarrow O\varphi$ ,
2. Mixed Negative Introspection:  $\neg A\varphi \rightarrow O\neg A\varphi$ .

The first of these axioms states that any knowledge that can be formed from a finite subset of observations can also be formed on the bases of the whole set of observations. The second axiom is significantly more interesting. We have seen in our first Grand Hotel example that the Negative Introspection axiom is not true for attainable knowledge. Namely an agent in a fully occupied hotel cannot learn that statement  $A$  (“the hotel has vacancies”) is false by examining only finitely many rooms. It can learn this, however, by examining all room in the hotel and this is exactly what the Mixed Negative Introspection axiom claims. Surprisingly, the Mixed Negative Introspection axiom is provable from other axioms of our logical system. We show this in Lemma 3. As a result, Mixed Negative Introspection is not included as an axiom of our system. Additionally, it is easy to see that the Truth axiom for modality  $A$  follows from the Truth axiom for modality  $O$  and the Monotonicity axiom. For this reason, we do not list the Truth axiom for modality  $A$  as one of our axioms either. Finally, although one can state two forms of the Necessitation inference rule: one for modality  $O$  and another for modality  $A$ , the former follows from the latter and the Monotonicity axiom. Thus, our system, in addition to Modus Ponens, only includes the Necessitation rule for modality  $A$ .

## 1.5 Outline

The rest of the paper is organized as following. Section 2 defines the syntax and an evidence-based semantics of our logical system. Section 3 lists the axioms and the inference rules of this system and proves the Mixed Negative Introspection axiom mentioned above. In Section 4, we prove the soundness of our logical system with respect to the evidence-based semantics. In Section 5, we state the completeness theorem and define the canonical model used in its proof. The full proof of the completeness can be found in the appendix. Section 7 concludes.

## 2 Syntax and Semantics

In this section, we describe the formal syntax and the formal evidence-based semantics of our logical system. Throughout the rest of the paper we assume a fixed infinite set of propositional variables.

**Definition 1** *The set  $\Phi$  of all formulae  $\varphi$  is defined by grammar  $\varphi := p \mid \neg\varphi \mid \varphi \rightarrow \varphi \mid A\varphi \mid O\varphi$ , where  $p$  represents propositional variables.*

**Definition 2** *An evidence model is  $\langle W, E, \{\sim_e\}_{e \in E}, \pi \rangle$ , where*

1.  $W$  is a (possibly empty) set of “epistemic worlds”,
2.  $E$  is an arbitrary (possibly empty) “evidence” set,
3.  $\sim_e$  is an “indistinguishability” equivalence relation on  $W$  for each  $e \in E$ ,
4.  $\pi$  is a function that maps propositional variables into subsets of  $W$ .

For instance, in our first Grand Hotel example, an epistemic world is a function  $\mathbb{N} \rightarrow \{\text{vacant}, \text{occupied}\}$  that assigns a state to each room in the hotel. The set of evidence is  $\mathbb{N}$ , where evidence with number  $r \in \mathbb{N}$  corresponds to examining room with number  $r$  in this hotel. Epistemic worlds  $w_1$  and  $w_2$  are  $\sim_r$ -equivalent if room number  $r$  has the same state in both of the worlds. In other words,  $w_1 \sim_r w_2$  if  $w_1(r) = w_2(r)$ . Finally a function  $\pi$  may, for example, map propositional variable  $p$  into the set of all epistemic worlds representing a nonempty hotel:  $\pi(p) = \{w \in W \mid \exists r \in \mathbb{N} \text{ such that } w(r) = \text{occupied}\}$ .

In this paper, we write  $w_1 \sim_F w_2$  if  $w_1 \sim_e w_2$  for each  $e \in F$ . In particular,  $w_1 \sim_\emptyset w_2$  for any two epistemic worlds  $w_1, w_2 \in W$ .

**Definition 3** *For any formula  $\varphi \in \Phi$  and any epistemic world  $w \in W$  of an evidence model  $\langle W, E, \{\sim_e\}, \pi \rangle$ , let the satisfaction relation  $w \Vdash \varphi$  be defined as follows,*

1.  $w \Vdash p$ , if  $w \in \pi(p)$ ,
2.  $w \Vdash \neg\varphi$ , if  $w \not\Vdash \varphi$ ,
3.  $w \Vdash \varphi \rightarrow \psi$ , if  $w \not\Vdash \varphi$  or  $w \Vdash \psi$ ,
4.  $w \Vdash A\varphi$ , if there is a finite  $F \subseteq E$  such that  $w \sim_F u$  implies  $u \Vdash \varphi$  for each  $u \in W$ ,
5.  $w \Vdash O\varphi$ , if  $u \Vdash \varphi$  for each  $u \in W$  such that  $w \sim_E u$ .

## 3 Axioms

In addition to propositional tautologies in language  $\Phi$ , our system consists of the following axioms:

1. Truth:  $O\varphi \rightarrow \varphi$ ,

2. Positive Introspection:  $A\phi \rightarrow AA\phi$ ,
3. Negative Introspection:  $\neg O\phi \rightarrow O\neg O\phi$ ,
4. Distributivity:  $A(\phi \rightarrow \psi) \rightarrow (A\phi \rightarrow A\psi)$  and  $O(\phi \rightarrow \psi) \rightarrow (O\phi \rightarrow O\psi)$ ,
5. Monotonicity:  $A\phi \rightarrow O\phi$ .

We say that a formula  $\phi$  is a theorem in our logical system and write  $\vdash \phi$  if formula  $\phi$  is derivable from the axioms of our systems using the Modus Ponens and the Necessitation inference rules:

$$\frac{\phi, \phi \rightarrow \psi}{\psi} \quad \frac{\phi}{A\phi}.$$

We write  $X \vdash \phi$  if formula  $\phi$  is derivable from the theorems of our logical systems and an additional set of axioms  $X$  using *only* the Modus Ponens inference rule. We say that set  $X \subseteq \Phi$  is inconsistent if there is a formula  $\phi \in \Phi$  such that  $X \vdash \phi$  and  $X \vdash \neg\phi$ .

**Lemma 1** *If  $X$  is a finite consistent set of formulae, and  $p$  is a propositional variable that does not occur in  $X$ , then sets  $X \cup \{p\}$  and  $X \cup \{\neg p\}$  are both consistent.*

PROOF. Suppose that either set  $X \cup \{p\}$  or set  $X \cup \{\neg p\}$  is inconsistent. Thus, either  $X \vdash \neg p$  or  $X \vdash p$ .

First, we show that  $X \vdash \neg p$  in both cases. Indeed, suppose that  $X \vdash p$ . Let  $\phi_1, \dots, \phi_n$  be the proof of formula  $p$  from the set of the assumption  $X$ . Let  $\phi'_i$  be the result of the replacement of propositional variable  $p$  by formula  $\neg p$  in formula  $\phi_i$ . Note that variable  $p$  does not occur in set  $X$  by the assumption of the lemma. Then,  $\phi'_1, \dots, \phi'_n$  is the proof of formula  $\neg p$  from the set of the assumption  $X$ . Therefore,  $X \vdash \neg p$ .

Next, we show that  $X \vdash \neg\neg p$ . Let  $\psi_1, \dots, \psi_m$  be the proof of formula  $\neg p$  from the set of the assumption  $X$ . Consider now sequence  $\psi'_1, \dots, \psi'_m$  obtained by replacing each occurrence of variable  $p$  with  $\neg p$  in proof  $\psi_1, \dots, \psi_m$ . Note again that variable  $p$  does not occur in set  $X$  by the assumption of the lemma. Then, sequence  $\psi'_1, \dots, \psi'_m$  is a proof of formula  $\neg\neg p$  from the set of the assumption  $X$ . Therefore,  $X \vdash \neg\neg p$ .

Statements  $X \vdash \neg p$  and  $X \vdash \neg\neg p$  imply that set  $X$  is not consistent. \(\square\)

**Lemma 2**  $\vdash O\phi \rightarrow OO\phi$ .

PROOF. Note that formula  $O\neg O\phi \rightarrow \neg O\phi$  is an instance of the Truth axiom. Thus,  $\vdash O\phi \rightarrow \neg O\neg O\phi$  by the law of contrapositive in the propositional logic. Hence, taking into account the following instance of the Negative Introspection axiom  $\neg O\neg O\phi \rightarrow O\neg O\neg O\phi$ , one can conclude that

$$\vdash O\phi \rightarrow O\neg O\neg O\phi. \quad (3)$$

At the same time,  $\neg O\phi \rightarrow O\neg O\phi$  is an instance of the Negative Introspection axiom. Thus,  $\vdash \neg O\neg O\phi \rightarrow O\phi$  by contraposition. Hence, by the Necessitation inference rule,  $\vdash A(\neg O\neg O\phi \rightarrow O\phi)$ . Thus,  $\vdash O(\neg O\neg O\phi \rightarrow O\phi)$  by the Monotonicity axiom and the Modus Ponens inference rule. Thus, by the Distributivity axiom and the Modus Ponens inference rule,  $\vdash O\neg O\neg O\phi \rightarrow OO\phi$ . The last statement, together with statement (3), implies the statement of the lemma by the laws of propositional reasoning. \(\square\)

We conclude this section with the proof of the Mixed Negative Introspection axiom mentioned in Section 1.4 of the introduction.

**Lemma 3**  $\vdash \neg A\varphi \rightarrow O\neg A\varphi$ .

PROOF. By the Truth axiom,  $\vdash OA\varphi \rightarrow A\varphi$ . Thus, by the law of contrapositive in the propositional logic,  $\vdash \neg A\varphi \rightarrow \neg OA\varphi$ . At the same time, by the Negative Introspection axiom,  $\vdash \neg OA\varphi \rightarrow O\neg OA\varphi$ . Hence, by the laws of propositional reasoning,

$$\vdash \neg A\varphi \rightarrow O\neg OA\varphi. \quad (4)$$

Note that  $\vdash A\varphi \rightarrow AA\varphi$  by the Positive Introspection axiom and  $\vdash AA\varphi \rightarrow OA\varphi$  by the Monotonicity axiom. Thus, by the laws of propositional reasoning,  $\vdash A\varphi \rightarrow OA\varphi$ . Hence, by the law of contrapositive in the propositional logic,  $\vdash \neg OA\varphi \rightarrow \neg A\varphi$ . Then,  $\vdash A(\neg OA\varphi \rightarrow \neg A\varphi)$  by the Necessitation inference rule. Thus,  $\vdash O(\neg OA\varphi \rightarrow \neg A\varphi)$  by the Monotonicity axiom and the Modus Ponens inference rule. Hence,  $\vdash O\neg OA\varphi \rightarrow O\neg A\varphi$  by the Distributivity axiom and the Modus Ponens inference rule. Therefore,  $\vdash \neg A\varphi \rightarrow O\neg A\varphi$ , by the laws of propositional reasoning taking into account statement (4).  $\square$

## 4 Soundness

**Theorem 1 (strong soundness)** *For any world  $w \in W$  of each evidence model  $\langle W, E, \{\sim_e\}_{e \in E}, \pi \rangle$ , any set of formulae  $X \subseteq \Phi$ , and any formula  $\varphi \in \Phi$ , if  $w \Vdash \chi$  for each formula  $\chi \in X$  and  $X \vdash \varphi$ , then  $w \Vdash \varphi$ .*

The soundness of propositional tautologies and Modus Ponens inference rule is straightforward. Below we prove the soundness of each of the remaining axioms and the Necessitation inference rule as separate lemmas.

**Lemma 4** *If  $w \Vdash O\varphi$ , then  $w \Vdash \varphi$ .*

PROOF. By Definition 3, assumption  $w \Vdash O\varphi$  implies that  $u \Vdash \varphi$  for all  $u \in W$  such that  $w \sim_E u$ . Note that  $w \sim_e w$  for all  $e \in E$  because  $\sim_e$  is an equivalence relation. Hence,  $w \sim_E w$ . Therefore,  $w \Vdash \varphi$ .  $\square$

**Lemma 5** *If  $w \Vdash A\varphi$ , then  $w \Vdash AA\varphi$ .*

PROOF. By Definition 3, the assumption  $w \Vdash A\varphi$  implies that there is a finite set  $F \subseteq E$  such that  $u \Vdash \varphi$  for each  $u \in W$  where  $w \sim_F u$ .

Again by Definition 3, it suffices to show that  $v \Vdash A\varphi$  for all  $v \in W$  such that  $w \sim_F v$ . To establish this, it is enough to prove that  $u \Vdash \varphi$  for all  $u \in W$  such that  $v \sim_F u$ . Note that  $w \sim_F v \sim_F u$ . Thus,  $w \sim_F u$  because  $\sim_f$  is an equivalence relation for each element  $f \in F$ . Then,  $u \Vdash \varphi$  by the choice of set  $F$ .  $\square$

**Lemma 6** *If  $w \Vdash \neg O\varphi$ , then  $w \Vdash O\neg O\varphi$ .*

PROOF. By Definition 3, assumption  $w \Vdash \neg O\varphi$  implies that there is  $u \in W$  such that  $w \sim_E u$  and  $u \not\Vdash \varphi$ .

Consider any  $v \in W$  such that  $w \sim_E v$ . By Definition 3, to prove  $w \Vdash O\neg O\varphi$ , it suffices to show that  $v \Vdash \neg O\varphi$ . Note that  $w \sim_E u$  and  $w \sim_E v$ . Thus,  $v \sim_E u$  due to  $\sim_e$  being an equivalence relation for each  $e \in E$ . Recall that  $u \not\Vdash \varphi$ . Hence,  $v \not\Vdash O\varphi$  by Definition 3. Therefore,  $v \Vdash \neg O\varphi$  again by Definition 3.  $\square$

**Lemma 7** *If  $w \Vdash A(\varphi \rightarrow \psi)$  and  $w \Vdash A\varphi$ , then  $w \Vdash A\psi$ .*

PROOF. By Definition 3, the assumption  $w \Vdash A(\varphi \rightarrow \psi)$  implies that there is a finite set  $F_1 \subseteq E$  such that  $u \Vdash \varphi \rightarrow \psi$  for each  $u \in W$  such that  $w \sim_{F_1} u$ . Similarly, the assumption  $w \Vdash A\varphi$  implies that there is a finite set  $F_2 \subseteq E$  such that  $u \Vdash \varphi$  for each  $u \in W$  such that  $w \sim_{F_2} u$ .

Let  $F = F_1 \cup F_2$ . It suffices to show that  $u \Vdash \psi$  for each  $u \in W$  such that  $w \sim_F u$ . Indeed, statement  $w \sim_F u$  implies that  $w \sim_{F_1} u$  and  $w \sim_{F_2} u$ . Hence,  $u \Vdash \varphi \rightarrow \psi$  and  $u \Vdash \varphi$  due to the choice of sets  $F_1$  and  $F_2$ . Therefore,  $u \Vdash \psi$  by Definition 3.  $\square$

**Lemma 8** *If  $w \Vdash O(\varphi \rightarrow \psi)$  and  $w \Vdash O\varphi$ , then  $w \Vdash O\psi$ .*

PROOF. Consider any  $u \in W$  such that  $w \sim_E u$ . By Definition 3, it suffices to show that  $u \Vdash \psi$ . Indeed, by Definition 3, the assumptions  $w \Vdash O(\varphi \rightarrow \psi)$  and  $w \Vdash O\varphi$  imply that  $u \Vdash \varphi \rightarrow \psi$  and  $u \Vdash \varphi$ . Therefore,  $u \Vdash \psi$ , again by Definition 3.  $\square$

**Lemma 9** *If  $w \Vdash A\varphi$ , then  $w \Vdash O\varphi$ .*

PROOF. Consider any  $u \in W$  such that  $w \sim_E u$ . By Definition 3, it suffices to prove that  $u \Vdash \varphi$ . By the same definition, the assumption  $w \Vdash A\varphi$  implies that there is a finite set  $F \subseteq E$  such that  $v \Vdash \varphi$  for each  $v \in W$  such that  $w \sim_F v$ . Note that statement  $w \sim_E u$  implies that  $w \sim_F u$ . Therefore,  $u \Vdash \varphi$ .  $\square$

**Lemma 10** *If  $w \Vdash \varphi$  for each epistemic world  $w \in W$  of each evidence model  $\langle W, E, \{\sim_e\}_{e \in E}, \pi \rangle$ , then  $w \Vdash A\varphi$  for each epistemic world  $w \in W$  of each evidence model  $\langle W, E, \{\sim_e\}_{e \in E}, \pi \rangle$ .*

PROOF. Consider any epistemic world  $w \in W$  of an arbitrary evidence model  $\langle W, E, \{\sim_e\}_{e \in E}, \pi \rangle$ . By Definition 3, it suffices to show that there is  $F \subseteq E$  such that  $u \Vdash \varphi$  for each  $u \in W$  where  $w \sim_F u$ . Indeed, let  $F = \emptyset$ . Note that  $u \Vdash \varphi$  for each  $u \in W$  due to the assumption of the lemma.  $\square$

This concludes the proof of the soundness of our logical system.

## 5 Canonical Model

**Theorem 2 (strong completeness)** *For any set of formulae  $X \subseteq \Phi$  and any formula  $\varphi \in \Phi$ , if  $X \not\vdash \varphi$ , then there is an evidence model  $\langle W, E, \{\sim_e\}_{e \in E}, \pi \rangle$  and an epistemic world  $w \in W$  such that  $w \Vdash \chi$  for each  $\chi \in X$  and  $w \not\vdash \varphi$ .*

As usual, the proof of a completeness theorem is using a canonical model construction. We define the canonical model in this section and give the full proof of the completeness in the appendix.

The standard proof of completeness for modal logics S4 and S5 defines worlds of a canonical model as maximal consistent sets of formulae. In our case, to specify a canonical model we need to define not only the set of worlds, but also the evidence set. In the construction that we propose, these two sets are defined to be the same set  $W_\infty$ , which is a set of nodes in a certain infinite forest. As a result, the proof of the completeness is significantly more involved than the completeness proofs for logics S4 and S5.

Throughout the section we use two operations on sequences. If  $w$  is a sequence  $(x_1, x_2, \dots, x_n)$  and  $u$  is a sequence  $(y_1, y_2, \dots, y_m)$ , then by concatenation  $w :: u$  of these two sequences we mean sequence  $(x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m)$ . By head  $hd(w)$  of a nonempty sequence  $w = (x_1, x_2, \dots, x_n)$  we mean element  $x_n$ . For example,  $(a, b) :: (c) = (a, b, c)$  and  $hd(a, b, c) = c$ .

We now define the “canonical” evidence model  $\langle W_\infty, W_\infty, \{\sim_e\}_{e \in W_\infty}, \pi \rangle$ . The set of epistemic worlds in the canonical model is identical to the evidence set. One can think that all pieces of evidence related in some sense to a given epistemic world are combined together into a single evidence associated with this world. Because of this, the evidence associated with world  $w$  is simply referred to as evidence  $w$ .

Informally, set  $W_\infty$  consists of sequences of the form  $(X_0, \varepsilon_1, X_1, \dots, \varepsilon_n, X_n)$  where  $X_0, \dots, X_n$  are maximal consistent sets of formulae and each of  $\varepsilon_1, \dots, \varepsilon_n$  is either a finite subset of  $W_\infty$  or symbol  $*$ . In what follows, the case when  $\varepsilon$  is a finite subset of  $W_\infty$  will form an A-accessibility relation and the case  $\varepsilon = *$  will form both A-accessibility and O-accessibility relations. Such sequences can be visualised, see Figure 1, as paths in an infinite collection of infinite trees whose vertices are maximal consistent sets of formulae and whose edges are labeled with the  $\varepsilon$ 's described above.

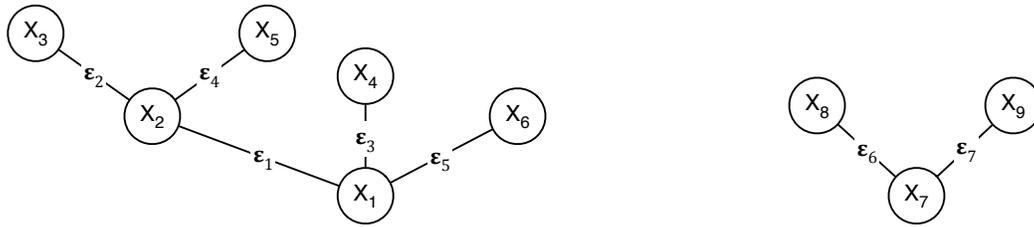


Figure 1: Fragments of the trees of sequences.

Formally, set  $W_\infty$  is specified as the union of a recursively defined infinite sequence of sets  $W_0, W_1, \dots$ . These sets represent different stages of building the infinite forest of infinite trees partially depicted in Figure 1. Note, however, that stages do not correspond to the levels of the trees. Generally speaking, vertices at the same level are not created at the same stage. After a new epistemic world is added, this world can be used as a part of evidence later in the construction.

**Definition 4** A sequence of sets  $W_0, W_1, \dots$  is defined recursively as follows.

1.  $W_0$  is the set of all single-element sequences  $(X_0)$ , where  $X_0$  is a maximal consistent subset of  $\Phi$ ,
2.  $W_{n+1}$  contains all sequences of the form  $w :: (\varepsilon, Y)$  such that
  - (a)  $w \in \bigcup_{i \leq n} W_i$ ,
  - (b)  $\varepsilon$  is either symbol  $*$  or a finite subset of  $\bigcup_{i \leq n} W_i$ ,
  - (c)  $Y$  is a maximal consistent subset of  $\Phi$ ,
  - (d) if  $\varepsilon = *$ , then  $\{\varphi \mid O\varphi \in hd(w)\} \subseteq Y$ ,
  - (e) if  $\varepsilon \subseteq \bigcup_{i \leq n} W_i$ , then  $\{\varphi \mid A\varphi \in hd(w)\} \subseteq Y$ .

**Definition 5**  $W_\infty = \bigcup_i W_i$ .

**Lemma 11** Set  $W_\infty$  is infinite.

PROOF. By Theorem 1, all axioms of our logical system are satisfied, in particular, in the world of any single-world model with the empty evidence set. Thus, our logical system is consistent. Hence, the empty set of formulae is consistent. Since our language contains infinitely many propositional variables, the empty set could be extended to a maximal consistent set in an infinitely many ways by Lemma 1. Hence, set  $W_0$  is infinite by Definition 4. Therefore, set  $W_\infty$  is infinite by Definition 5.  $\square$

Informally, two sequences are  $\sim_e$ -equivalent if they start with the same prefix and once they deviate all subsequent  $\varepsilon$ 's either are equal to  $*$  or contain element  $e$ , see Figure 2. The formal definition is below.

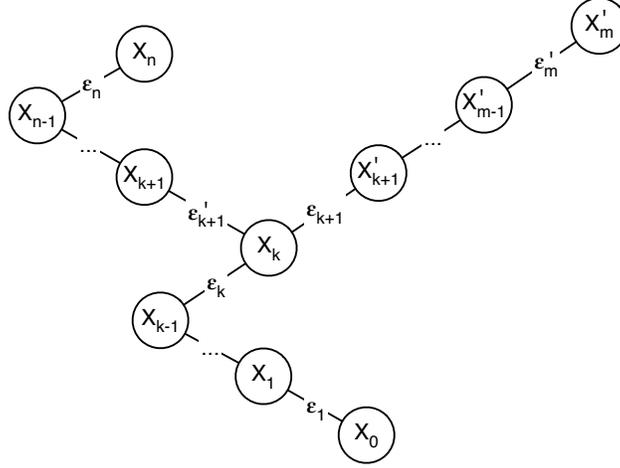


Figure 2: Illustration for Definition 6, where  $\varepsilon_i$  and  $\varepsilon'_i$  either are equal to  $*$  or contain  $e$  for each  $i \geq k$ .

**Definition 6** For any world  $w = (X_0, \varepsilon_1, X_1, \dots, \varepsilon_n, X_n)$ , any world  $u = (X'_0, \varepsilon'_1, X'_1, \dots, \varepsilon'_m, X'_m)$ , and any  $e \in W_\infty$ , let  $w \sim_e u$  if there is  $k$  such that

1.  $0 \leq k \leq \min\{n, m\}$ ,
2.  $X_i = X'_i$  for all  $i$  such that  $0 \leq i \leq k$ ,
3.  $\varepsilon_i = \varepsilon'_i$  for all  $i$  such that  $0 < i \leq k$ ,
4. for all  $i$ , if  $k < i \leq n$ , then either  $e \in \varepsilon_i$  or  $\varepsilon_i = *$ ,
5. for all  $i$ , if  $k < i \leq m$ , then either  $e \in \varepsilon'_i$  or  $\varepsilon'_i = *$ .

**Definition 7**  $\pi(p) = \{w \in W_\infty \mid p \in hd(w)\}$ .

The canonical evidence model  $\langle W_\infty, W_\infty, \{\sim_e\}_{e \in W_\infty}, \pi \rangle$  is now fully defined.

## 6 Completeness

We start by establishing several properties of the canonical model that are necessary for our proof of the completeness. First, we show that if set  $hd(w)$  contains a formula  $A\varphi$ , then so does set  $hd(u)$  for each descendant  $u$  of vertex  $w$ .

**Lemma 12** For any  $k \geq 0$ , any  $n \geq 0$ , and any  $(X_0, \varepsilon_1, X_1, \dots, \varepsilon_k, X_k, \varepsilon_{k+1}, X_{k+1}, \dots, \varepsilon_{k+n}, X_{k+n}) \in W_\infty$ , if  $A\varphi \in X_k$ , then  $A\varphi \in X_{k+n}$ .

PROOF. We prove the lemma by induction on  $n$ , see Figure 3 (a). If  $n = 0$ , then assumption  $A\varphi \in X_k$  implies that  $A\varphi \in X_{k+n}$ . If  $n > 0$ , then  $A\varphi \in X_{k+n-1}$  by the induction hypothesis. Hence,  $X_{k+n-1} \vdash AA\varphi$  by the Positive Introspection axiom.

Case I:  $\varepsilon_{k+n} = *$ . Note that  $X_{k+n-1} \vdash AA\varphi$  implies  $X_{k+n-1} \vdash OA\varphi$  by the Monotonicity axiom. Hence,  $OA\varphi \in X_{k+n-1}$  due to the maximality of set  $X_{k+n-1}$ . Then,  $A\varphi \in X_{k+n}$  by Definition 4 and due to the assumption  $\varepsilon_{k+n} = *$ .

Case II:  $\varepsilon_{k+n} \neq *$ . Statement  $X_{k+n-1} \vdash AA\varphi$  implies that  $AA\varphi \in X_{k+n-1}$  due to the maximality of set  $X_{k+n-1}$ . Thus,  $A\varphi \in X_{k+n}$  by Definition 4.  $\square$

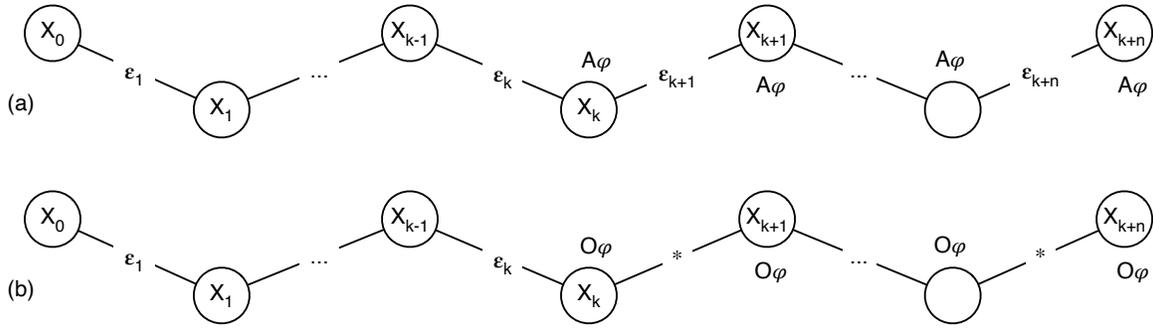


Figure 3: Illustrations: (a) for Lemma 12, and (b) for Lemma 13 and Lemma 14.

Next, we show that if  $hd(w)$  contains a formula  $O\phi$ , then so does  $hd(u)$  for each descendant  $u$  of vertex  $w$  reachable through edges all of which are labeled by symbol  $*$ .

**Lemma 13** For any  $k \geq 0$  and  $n \geq 0$ , and any

$$(X_0, \epsilon_1, X_1, \dots, \epsilon_k, X_k, \epsilon_{k+1}, X_{k+1}, \dots, \epsilon_{k+n}, X_{k+n}) \in W_\infty,$$

if  $O\phi \in X_k$  and  $\epsilon_i = *$  for all  $i$  such that  $k < i \leq k+n$ , then  $O\phi \in X_{k+n}$ .

PROOF. We prove the lemma by induction on  $n$ , see Figure 3 (b). If  $n = 0$ , then assumption  $O\phi \in X_k$  implies that  $O\phi \in X_{k+n}$ . If  $n > 0$ , then  $O\phi \in X_{k+n-1}$  by the induction hypothesis. Hence,  $X_{k+n-1} \vdash OO\phi$  by Lemma 2. Thus,  $O\phi \in X_{k+n}$  by Definition 4 and due to the assumption  $\epsilon_{k+n} = *$ .  $\square$

The next lemma is a converse of Lemma 13. It shows that if  $hd(u)$  contains a formula  $O\phi$ , where  $u$  is a descendant of vertex  $w$  reachable through edges all of which are labeled by symbol  $*$ , then so does  $hd(w)$ .

**Lemma 14** For any  $k \geq 0$ , any  $n \geq 0$ , and any

$$(X_0, \epsilon_1, X_1, \dots, \epsilon_k, X_k, \epsilon_{k+1}, X_{k+1}, \dots, \epsilon_{k+n}, X_{k+n}) \in W_\infty,$$

if  $O\phi \in X_{k+n}$  and  $\epsilon_i = *$  for all  $i$  such that  $k < i \leq k+n$ , then  $O\phi \in X_k$ .

PROOF. We prove the statement of the lemma by induction on  $n$ , see again Figure 3 (b). If  $n = 0$ , then assumption  $O\phi \in X_{k+n}$  implies that  $O\phi \in X_k$ . In what follows we assume that  $n > 0$ .

Case I:  $O\phi \notin X_{k+n-1}$ . Thus,  $\neg O\phi \in X_{k+n-1}$  due to the maximality of the set  $X_{k+n-1}$ . Hence,  $X_{k+n-1} \vdash O\neg O\phi$  by the Negative Introspection axiom. Then,  $O\neg O\phi \in X_{k+n-1}$  due to the maximality of the set  $X_{k+n-1}$ . Hence,  $\neg O\phi \in X_{k+n}$  by Definition 4 and because  $\epsilon_{k+n} = *$ . Thus,  $O\phi \notin X_{k+n}$  due to the consistency of the set  $X_{k+n}$ , which is a contradiction with the assumption of the lemma.

Case II:  $O\phi \in X_{k+n-1}$ . Thus,  $O\phi \in X_k$  by the induction hypothesis.  $\square$

The next two lemmas are relatively standard lemmas for a completeness proof of a modal logic. Their proofs show how a sequence representing an epistemic world can be extended to produce different types of child nodes on the trees in Figure 1.

**Lemma 15** For any  $w \in W_\infty$ , any  $\neg A\phi \in hd(w)$ , and any finite  $F \subseteq W_\infty$ , there is  $u \in W_\infty$  such that  $w \sim_F u$  and  $\neg\phi \in hd(u)$ .

PROOF. We first show that the following set is consistent:  $Y_0 = \{\neg\varphi\} \cup \{\psi \mid A\psi \in hd(w)\}$ . Assume the opposite. Thus, there must exist  $A\psi_1, \dots, A\psi_n \in hd(w)$  such that  $\psi_1, \dots, \psi_n \vdash \varphi$ . Hence, by the deduction theorem for propositional logic,  $\vdash \psi_1 \rightarrow (\psi_2 \rightarrow \dots (\psi_n \rightarrow \varphi) \dots)$ . Then, by the Necessitation inference rule,  $\vdash A(\psi_1 \rightarrow (\psi_2 \rightarrow \dots (\psi_n \rightarrow \varphi) \dots))$ . By the Distributivity axiom and the Modus Ponens inference rule,  $\vdash A\psi_1 \rightarrow A(\psi_2 \rightarrow \dots (\psi_n \rightarrow \varphi) \dots)$ . By the Modus Ponens inference rule,  $A\psi_1 \vdash A(\psi_2 \rightarrow \dots (\psi_n \rightarrow \varphi) \dots)$ . By repeating the last two steps  $n - 1$  times,  $A\psi_1, \dots, A\psi_n \vdash A\varphi$ . Hence,  $hd(w) \vdash A\varphi$  by the choice of formulae  $\psi_1, \dots, \psi_n$ . Thus,  $\neg A\varphi \notin hd(w)$  due to the consistency of the set  $hd(w)$ , which contradicts the assumption of the lemma. Therefore, set  $Y_0$  is consistent.

Let  $Y$  be any maximal consistent extension of set  $Y_0$  and let  $u$  be the sequence  $w :: (F, Y)$ . We next show that  $u \in W_\infty$ . Indeed, since  $w \in W_\infty$ , by Definition 5, there must exist  $n_1 \geq 0$  such that  $w \in W_{n_1}$ . At the same time, since set  $F$  is a finite subset of  $W_\infty$ , by Definition 5, there must exist  $n_2 \geq 0$  such that  $F \subseteq \bigcup_{i \leq n_2} W_i$ . Let  $n = \max\{n_1, n_2\}$ . Thus,  $w \in \bigcup_{i \leq n} W_i$  and  $F \subseteq \bigcup_{i \leq n} W_i$ . Hence,  $w :: (F, Y) \in W_{n+1}$  by Definition 4. Therefore,  $u = w :: (F, Y) \in W_\infty$  by Definition 5. Finally,  $w \sim_F u$  by Definition 6. To finish the proof of the lemma, note that  $\neg\varphi \in Y_0 \subseteq Y = hd(u)$ .  $\square$

**Lemma 16** For any  $w \in W_\infty$  and any  $\neg O\varphi \in hd(w)$ , there is  $u \in W_\infty$  such that  $w \sim_{W_\infty} u$  and  $\neg\varphi \in hd(u)$ .

PROOF. We first show that the following set is consistent:  $Y_0 = \{\neg\varphi\} \cup \{\psi \mid O\psi \in hd(w)\}$ . Assume the opposite. Thus, there must exist formulae  $O\psi_1, \dots, O\psi_n \in hd(w)$  and  $\psi_1, \dots, \psi_n \vdash \varphi$ . Hence, by the deduction theorem for the propositional logic,  $\vdash \psi_1 \rightarrow (\psi_2 \rightarrow \dots (\psi_n \rightarrow \varphi) \dots)$ . Then, by the Necessitation inference rule,  $\vdash A(\psi_1 \rightarrow (\psi_2 \rightarrow \dots (\psi_n \rightarrow \varphi) \dots))$ . By the Monotonicity axiom and the Modus Ponens Inference rule,  $\vdash O(\psi_1 \rightarrow (\psi_2 \rightarrow \dots (\psi_n \rightarrow \varphi) \dots))$ . By the Distributivity axiom and the Modus Ponens inference rule,  $\vdash O\psi_1 \rightarrow O(\psi_2 \rightarrow \dots (\psi_n \rightarrow \varphi) \dots)$ . By the Modus Ponens inference rule,  $O\psi_1 \vdash O(\psi_2 \rightarrow \dots (\psi_n \rightarrow \varphi) \dots)$ . By repeating the last two steps  $n - 1$  times,  $O\psi_1, \dots, O\psi_n \vdash O\varphi$ . Hence,  $hd(w) \vdash O\varphi$  by the choice of formulae  $\psi_1, \dots, \psi_n$ . Thus,  $\neg O\varphi \notin hd(w)$  due to the consistency of the set  $hd(w)$ , which contradicts the assumption of the lemma. Therefore, set  $Y_0$  is consistent.

Let  $Y$  be any maximal consistent extension of set  $Y_0$  and let  $u$  be sequence  $w :: (*, Y)$ . We next show that  $u \in W_\infty$ . Indeed, since  $w \in W_\infty$ , by Definition 5, there must exist  $n \geq 0$  such that  $w \in W_n$ . Hence,  $w :: (*, Y) \in W_{n+1}$  by Definition 4. Therefore,  $u = w :: (*, Y) \in W_\infty$  by Definition 5.

Finally, let us observe that  $w \sim_e u$  for each  $e \in W_\infty$  by Definition 6. Thus,  $w \sim_{W_\infty} u$ . To finish the proof of the lemma, note that  $\neg\varphi \in Y_0 \subseteq Y = hd(u)$ .  $\square$

By Definition 4, elements of set  $W_\infty$  are sequences of the form  $(X_0, \varepsilon_1, X_1, \dots, \varepsilon_k, X_k)$ , where  $\varepsilon_i$  is either symbol  $*$  or a subset of  $W_\infty$ . One might wonder if elements of  $W_\infty$  are wellfounded. In other words, is it possible for an element  $w \in W_\infty$  to be a member of one of its own  $\varepsilon_i$ ? Lemma 18 shows that such elements do not exist. This is a very important observation for our proof of completeness. Lemma 17 is essentially a different form of Lemma 18 which is easier to prove by induction.

**Lemma 17** For any  $k \geq 0$ , any  $n \geq 0$ , any  $t \geq 0$ , and any  $w = (X_0, \varepsilon_1, X_1, \dots, \varepsilon_k, X_k, \varepsilon_{k+1}, X_{k+1}, \dots, \varepsilon_{k+n}, X_{k+n}) \in W_t$ , if  $\varepsilon_k \neq *$ , then  $\varepsilon_k \subseteq \bigcup_{i=0}^{t-1} W_i$ .

PROOF. We prove this statement by induction on  $n$ . First, let  $n = 0$ . Thus,  $w = (X_0, \varepsilon_1, X_1, \dots, \varepsilon_k, X_k)$ . Hence, By Definition 4, assumptions  $w \in W_t$  and  $\varepsilon_k \neq *$  imply that  $\varepsilon_k \subseteq \bigcup_{i=0}^{t-1} W_i$ .

Suppose now that  $n > 0$ . By Definition 4, the assumption  $w \in W_t$  implies that

$$(X_0, \varepsilon_1, X_1, \dots, \varepsilon_k, X_k, \varepsilon_{k+1}, X_{k+1}, \dots, \varepsilon_{k+n-1}, X_{k+n-1}) \in W_{t-1}.$$

Thus, by the induction hypothesis,  $\varepsilon_k \subseteq \bigcup_{i=0}^{t-2} W_i$ . Therefore,  $\varepsilon_k \subseteq \bigcup_{i=0}^{t-1} W_i$ .  $\square$

**Lemma 18**  $w \notin \varepsilon_k$  for each  $w = (X_0, \varepsilon_1, X_1, \dots, \varepsilon_n, X_n) \in W_\infty$  and each  $k \leq n$ .

PROOF. By Definition 5, assumption  $w \in W_\infty$  implies that there is  $t \geq 0$  such that  $w \in W_t$ . Let  $m$  be the smallest  $m$  such that  $w \in W_m$ . Thus,  $w \notin \bigcup_{i=0}^{m-1} W_i$ . At the same time,  $\varepsilon_k \subseteq \bigcup_{i=0}^{m-1} W_i$  by Lemma 17. Therefore,  $w \notin \varepsilon_k$ .  $\square$

The next lemma puts together the pieces of the proof that we have developed. It connects the satisfaction of a formula in an epistemic world of the canonical evidence model with the maximal consistent sets out of which the world is constructed.

**Lemma 19**  $w \Vdash \varphi$  iff  $\varphi \in hd(w)$ , for each epistemic world  $w \in W_\infty$  and each formula  $\varphi \in \Phi$ .

PROOF. We prove the lemma by induction on the structural complexity of formula  $\varphi$ . If formula  $\varphi$  is an atomic proposition, then the required follows from Definition 7 and Definition 3. If formula  $\varphi$  is a negation or an implication, then the required follows from Definition 3 and the maximality and the consistency of the set  $hd(w)$  in the standard way.

Suppose now that formula  $\varphi$  has the form  $A\psi$ .

( $\Rightarrow$ ) If  $A\psi \notin hd(w)$ , then  $\neg A\psi \in hd(w)$  due to the maximality of the set  $hd(w)$ . To prove  $w \not\Vdash A\psi$ , by Definition 3, we need to show that for any finite set  $F \subseteq W_\infty$  there is  $u \in W_\infty$  such that  $w \sim_F u$  and  $u \not\Vdash \psi$ . Indeed, by Lemma 15, there is  $u \in W_\infty$  such that  $w \sim_F u$  and  $\neg\psi \in hd(u)$ . Thus,  $\psi \notin hd(u)$  due to the consistency of the set  $hd(u)$ . Therefore,  $u \not\Vdash \psi$  by the induction hypothesis.

( $\Leftarrow$ ) Suppose that  $A\psi \in hd(w)$ . Thus,  $hd(w) \vdash AA\psi$  by the Positive Introspection axiom. Hence,  $hd(w) \vdash OA\psi$  by the Monotonicity axiom. Then,

$$OA\psi \in hd(w) \tag{5}$$

due to the maximality of the set  $hd(w)$ .

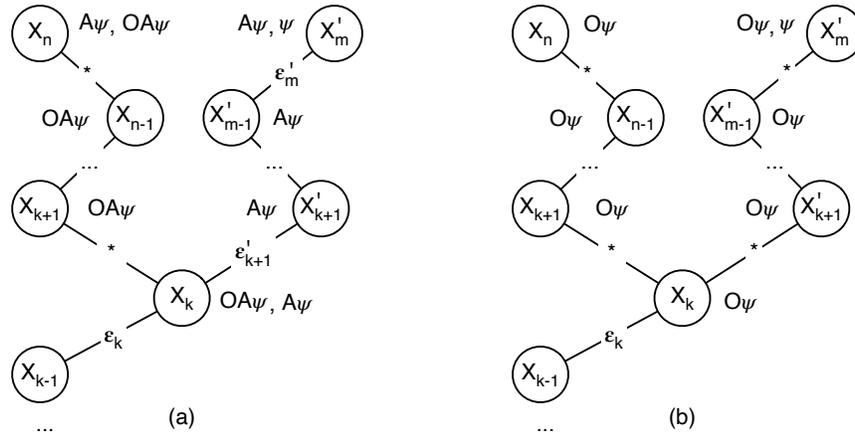


Figure 4: Illustrations for the proof of Lemma 19.

By Definition 3, it suffices to show that  $u \Vdash \psi$  for all  $u \in W_\infty$  such that  $w \sim_w u$ . By Definition 6, assumption  $w \sim_w u$  implies that if  $w = (X_0, \varepsilon_1, X_1, \dots, \varepsilon_n, X_n)$  and  $u = (X'_0, \varepsilon'_1, X'_1, \dots, \varepsilon'_m, X'_m)$ , then there is  $k$  such that,

1.  $0 \leq k \leq \min\{n, m\}$ ,
2.  $X_i = X'_i$  for all  $i$  such that  $0 \leq i \leq k$ ,

3.  $\varepsilon_i = \varepsilon'_i$  for all  $i$  such that  $0 < i \leq k$ ,
4. for all  $i$ , if  $k < i \leq n$ , then either  $w \in \varepsilon_i$  or  $\varepsilon_i = *$ ,
5. for all  $i$ , if  $k < i \leq m$ , then either  $w \in \varepsilon'_i$  or  $\varepsilon'_i = *$ .

By Lemma 18,  $w \notin \varepsilon_i$  for each  $i \leq n$ . Thus, from the condition 4 above,  $\varepsilon_i = *$  for each  $i$  such  $k < i \leq n$ , see Figure 4 (a). Hence, by Lemma 14, statement (5) implies that  $\text{OA}\psi \in X_k$ . Thus,  $X_k \vdash \text{A}\psi$  by the Truth axiom. Then,  $\text{A}\psi \in X_k$  due to the maximality of set  $X_k$ . Therefore,  $\psi \in X'_m = \text{hd}(u)$  by Lemma 12 and condition 5 above. Therefore,  $u \Vdash \psi$  by the induction hypothesis.

Finally, let formula  $\varphi$  have the form  $\text{O}\psi$ .

( $\Rightarrow$ ) If  $\text{O}\psi \notin \text{hd}(w)$ , then  $\neg\text{O}\psi \in \text{hd}(w)$  due to the maximality of the set  $\text{hd}(w)$ . To prove  $w \not\Vdash \text{O}\psi$ , by Definition 3, we need to show that  $u \not\Vdash \psi$  for some  $u \in W_\infty$  such that  $w \sim_{W_\infty} u$ . Indeed, by Lemma 16, there is  $u \in W_\infty$  such that  $w \sim_{W_\infty} u$  and  $\neg\psi \in \text{hd}(u)$ . Thus,  $\psi \notin \text{hd}(u)$  due to the consistency of the set  $\text{hd}(u)$ . Therefore,  $u \not\Vdash \psi$  by the induction hypothesis.

( $\Leftarrow$ ) Suppose that  $\text{O}\psi \in \text{hd}(w)$ . By Definition 3, it suffices to show that  $u \Vdash \psi$  for all  $u \in W_\infty$  such that  $w \sim_{W_\infty} u$ . Indeed, let  $w = (X_0, \varepsilon_1, X_1, \dots, \varepsilon_n, X_n)$  and  $u = (X'_0, \varepsilon'_1, X'_1, \dots, \varepsilon'_m, X'_m)$ . By Definition 6, assumption  $w \sim_{W_\infty} u$  implies that for each  $e \in W_\infty$  there is integer  $k_e$  such that

1.  $0 \leq k_e \leq \min\{n, m\}$ ,
2.  $X_i = X'_i$  for all  $i$  such that  $0 \leq i \leq k_e$ ,
3.  $\varepsilon_i = \varepsilon'_i$  for all  $i$  such that  $0 < i \leq k_e$ ,
4. for all  $i$ , if  $k_e < i \leq n$ , then either  $w \in \varepsilon_i$  or  $\varepsilon_i = *$ ,
5. for all  $i$ , if  $k_e < i \leq m$ , then either  $w \in \varepsilon'_i$  or  $\varepsilon'_i = *$ .

By Lemma 11, set  $W_\infty$  is infinite and, thus, it is nonempty. Hence, set  $\{k_e \mid e \in W_\infty\}$  is nonempty. Set  $\{k_e \mid e \in W_\infty\}$  is finite because  $0 \leq k_e \leq \min\{n, m\}$  for each  $e \in W_\infty$ . Let  $k$  be the maximal element of the set  $\{k_e \mid e \in W_\infty\}$ . Hence,

1.  $0 \leq k \leq \min\{n, m\}$ ,
2.  $X_i = X'_i$  for all  $i$  such that  $0 \leq i \leq k$ ,
3.  $\varepsilon_i = \varepsilon'_i$  for all  $i$  such that  $0 < i \leq k$ ,
4. for all  $i$ , if  $k < i \leq n$ , then either  $W_\infty \subseteq \varepsilon_i$  or  $\varepsilon_i = *$ ,
5. for all  $i$ , if  $k < i \leq m$ , then either  $W_\infty \subseteq \varepsilon'_i$  or  $\varepsilon'_i = *$ .

By Lemma 11, set  $W_\infty$  is infinite yet sets  $\varepsilon'_i$  and  $\varepsilon_i$  are finite for each  $i$  by Definition 4. Hence,  $W_\infty \not\subseteq \varepsilon_i$  and  $W_\infty \not\subseteq \varepsilon'_i$  for each  $i$ . Thus, conditions 4 and 5 above imply that  $\varepsilon_i = *$  for all  $i$  such that  $k < i \leq n$  and  $\varepsilon'_i = *$  for all  $i$  such that  $k < i \leq m$ , see Figure 4 (b). Thus, by Lemma 14, assumption  $\text{O}\psi \in \text{hd}(w) = X_n$  implies that  $\text{O}\psi \in X_k = X'_k$ . Therefore,  $\psi \in X'_m = \text{hd}(u)$ , by Lemma 13.  $\square$

We are now ready to finish the proof of Theorem 2. Set  $X \cup \{\neg\varphi\}$  is consistent by the assumption  $X \not\vdash \varphi$  of the theorem. Consider any maximal consistent extension  $X_0$  of set  $X \cup \{\neg\varphi\}$ . Let  $w_0$  be the single-element sequence  $(X_0)$ . Thus,  $w_0 \Vdash \chi$  for each formula  $\chi \in X$  and  $w_0 \not\Vdash \varphi$  by Lemma 19. This concludes the proof of the theorem.

## 7 Conclusion

In this paper we have shown that knowledge obtained from infinitely many pieces of evidence has properties captured by modal logic S5 and knowledge obtained from finitely many pieces of evidence has properties described by modal logic S4. The main technical result is a sound and complete propositional bi-modal logic that captures properties of both of these types of knowledge and their interplay.

A natural next step is to consider first-order logic with the same two modalities. Note that the knowledge modality  $O$  satisfies Barcan Formula [3]  $\forall x O\varphi \rightarrow O\forall x\varphi$  because this formula is derivable from S5 axioms stated in the first-order modal language [14]. At the same time, attainable knowledge modality  $A$  does not satisfy Barcan Formula  $\forall x A\varphi \rightarrow A\forall x\varphi$ . Indeed, if variable  $x$  ranges over an infinite domain and each value of  $x$  in this domain has a distinct single evidence  $e_x$  that justifies  $\varphi(x)$ , then  $\forall x A\varphi$  is true, but  $A\forall x\varphi$  is not. A complete axiomatization of the interplay of these two modalities in the first-order language remains an open problem.

## References

- [1] Thomas Ågotnes & Natasha Alechina (2019): *Coalition Logic with Individual, Distributed and Common Knowledge*. *Journal of Logic and Computation* 29, pp. 1041–1069, doi:10.1093/logcom/exv085.
- [2] Thomas Ågotnes, Philippe Balbiani, Hans van Ditmarsch & Pablo Seban (2010): *Group announcement logic*. *Journal of Applied Logic* 8(1), pp. 62 – 81, doi:10.1016/j.jal.2008.12.002.
- [3] Ruth C Barcan (1946): *A functional calculus of first order based on strict implication*. *The Journal of Symbolic Logic* 11(01), pp. 1–16, doi:10.2307/2269159.
- [4] Johan van Benthem & Eric Pacuit (2011): *Dynamic logics of evidence-based beliefs*. *Studia Logica* 99(1-3), pp. 61–92, doi:10.1007/s11225-011-9347-x.
- [5] Hans van Ditmarsch, Wiebe van der Hoek & Barteld Kooi (2007): *Dynamic Epistemic Logic*. Springer, doi:10.1007/978-1-4020-5839-4.
- [6] Ronald Fagin, Joseph Y. Halpern, Yoram Moses & Moshe Y. Vardi (1995): *Reasoning about knowledge*. MIT Press, Cambridge, MA, doi:10.7551/mitpress/5803.001.0001.
- [7] Joseph Y. Halpern & Riccardo Pucella (2006): *A Logic for Reasoning About Evidence*. *J. Artif. Int. Res.* 26(1), pp. 1–34, doi:10.1613/jair.1838.
- [8] Jaakko Hintikka (1962): *Knowledge and Belief - An Introduction to the Logic of the Two Notions*. Contemporary philosophy, Cornell University Press, Ithaca, NY, doi:10.2307/2217611.
- [9] F von Kutschera (1976): *Einführung in die intensional Semantik*.
- [10] Wolfgang Lenzen (1978): *Recent Work in Epistemic Logic*. *Acta Philosophica Fennica* 30(1), pp. 1–219.
- [11] John-Jules Ch. Meyer & Wiebe van der Hoek (2004): *Epistemic Logic for AI and Computer Science*. Cambridge University Press.
- [12] Pavel Naumov & Jia Tao (2015): *Budget-Constrained Knowledge in Multiagent Systems*. In: *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, International Foundation for Autonomous Agents and Multiagent Systems, pp. 219–226.
- [13] Pavel Naumov & Jia Tao (2017): *Price of privacy*. *Journal of Applied Logic* 20, pp. 32–48, doi:10.1016/j.jal.2016.11.035.
- [14] Arthur N Prior (1956): *Modality and quantification in S5*. *The Journal of Symbolic Logic* 21(01), pp. 60–62, doi:10.2307/2268488.
- [15] Robert Stalnaker (2006): *On logics of knowledge and belief*. *Philosophical studies* 128(1), pp. 169–199, doi:10.1007/s11098-005-4062-y.



# Failures of Contingent Thinking

Evan Piermont

Royal Holloway, University of London  
Department of Economics  
United Kingdom  
evan.piermont@rhul.ac.uk

Peio Zuazo-Garin

Higher School of Economics  
International College of Economics and Finance  
Russia  
p.zuazogarin@hse.ru

In this paper, we provide a theoretical framework to analyze an agent who misinterprets or misperceives the true decision problem she faces. Within this framework, we show that a wide range of behavior observed in experimental settings manifest as failures to *perceive implications*, in other words, to properly account for the logical relationships between various payoff relevant contingencies. We present behavioral characterizations corresponding to several benchmarks of logical sophistication and show how it is possible to identify *which* implications the agent fails to perceive. Thus, our framework delivers both a methodology for assessing an agent's level of contingent thinking and a strategy for identifying her beliefs in the absence full rationality.

KEYWORDS: Bounded rationality, contingent thinking, subjective beliefs, monotonicity.



# Reasoning about Emergence of Collective Memory

R. Ramanujam

Institute of Mathematical Sciences

Homi Bhabha National Institute

Chennai, India

jam@imsc.res.in

We offer a very simple model of how collective memory may form. Agents keep signalling within neighbourhoods, and depending on how many support each signal, some signals “win” in that neighbourhood. By agents interacting between different neighbourhoods, ‘influence’ spreads and sometimes, a collective signal emerges. We propose a logic in which we can reason about such emergence of memory and present preliminary technical results on the logic.

## 1 Introduction

*Strictly speaking, there is no such thing as collective memory – part of the same family of spurious notions as collective guilt. But there is collective instruction ... All memory is individual, unreproducible; it dies with each person. What is called collective memory is not a remembering but a stipulating: that this is important, and this is the story about how it happened, with the pictures that lock the story in our minds.*

Susan Sontag ([24])

Any discussion on individual values and social values visits the question, *How are we to act?* at some point. The question, of course, is, who is this **we** referred to here? Clearly this **we** is a social construction, one that depends on the very social norms and social values that we wish to reason about ([27]). Integral to such social construction of a collective is the memory ascribed to that collective. Group identity is constructed structurally by ascribing memory to the group, and in turn, such identity shapes its memory. Remembrance has a crucial impact on preferences and values, influences action.

It is here that Susan Sontag’s quote above assumes significance. Sontag calls collective memory a process of stipulation. Somehow the collective ascribes importance to an item of memory, authenticates it and symbolizes it; then on, the symbolism “locks” the memory item, in Sontag’s account.

Note that this is a significant departure from the structural conception of memory, that visualises memory as a notebook, and remembering as looking it up. Wittgenstein ([28]) strongly attacked such a conception of memory.

I saw this man years ago: now I have seen him again, I recognize him, I remember his name. And why does there have to be a cause of this remembering in my nervous system? Why must something or other, whatever it may be, be stored up there in any form? Why must a trace have been left behind? Why should there not be a psychological regularity to which no physiological regularity corresponds?

If this upsets our concept of causality then it is high time it was upset.

Scholars like Sutton ([25]) have discussed this issue at length. For Wittgenstein, social acts were important in shaping memory, and based on this, scholars like Rusu ([23]) even talk of *social time*, and modern theories of connectionism and distributed memory build on many such notions.

For us, these remarks are relevant from two viewpoints. The 1950's saw the development of automata theory as a study of *memory structures*, and in theory of computation, automata provide a model of memory that Wittgenstein might have approved of. In this view, memory is not a table to be looked up, but is constituted by states of being of the automaton. Observations cause changes in state, some states remember (some of the past) and some forget. Thus, remembering and forgetting are built into system structure. Such a view is important for seeing memory and reasoning as *interdependent* rather than as separate (as psychologists used to consider). Logicians are used to equating automata and logics, as in the case of monadic second order logics of order or in the case of Pressburger arithmetic. (Wittgenstein would have approved.)

The other viewpoint relates to *distributed memory*, where interacting agents rely on memory external to them. Computer science has evolved impressive models of highly flexible interaction and memory that has literally changed the everyday life of much of humanity in the last few decades. Today, memory storage on the "cloud" has become indispensable for many, and people voluntarily 'post' personal information to make it socially available in an attempt to write personal information into social memory.

In social theory, the notion of collective memory is influential. *Maurice Halbwachs* ([14]) talked of how an individual's understanding of the past is strongly linked to a group consciousness, which in turn is a form of *group memory* that lives beyond the memories of individuals that form the group.

For the logician, these notions pose an interesting challenge: what are the logical properties of collective remembering? What is the rationale followed by a group in ascribing / stipulating collective importance to events and their remembering? Why is a particular idealisation chosen? These are difficult questions to answer, but a more modest reformulation of such questions offers an approach to solutions. If the memory of an automaton is describable via logic, we can perhaps build a model of group and individual memory based on automata whose interactions lead to collective memory, which in turn influences behaviour of individual automata.

Why should one bother? In ([19]) Rohit Parikh speaks of *cultural structures* providing an infrastructure to social algorithms (much as data structures do for computational algorithms). Epistemic reasoning is an essential component of social algorithms, as persuasively argued by Parikh. We can then see collective memory as an essential gradient of its infrastructure creating the 'common ground' in which social objectives and communications are interpreted ([10]). Moreover, social algorithms such as elections need social memory if they are to achieve their democratic purpose.

Moreover, there has been extensive research in recent years on notions of collective agency ([26]), collective action ([22]), collective belief ([13]) and many more. However, the memory required for collective action, belief and agency is largely assumed rather than explicitly discussed. While social theorists extensively discuss the role of society's needs for remembrance (as a way of acknowledging the past and taking responsibility for it) the notion of memory as an infrastructural need for such functions is often glossed over. Moreover, the notions of memory that inform these discussions are embodied in text, icons and physical tokens, much like the notebook that Wittgenstein refers to (and objects to). Contemporary reality, with the extension of human memory using technology, suggests that more dynamic, behavioural models of memory, and the re-inforcing behaviour that leads to stipulation as referred to by Sontag, may be relevant.

What follows is a very simple, perhaps very simplistic, attempt at formalization of this notion, inspired by the study of *population protocols* in distributed computing ([4]) and large anonymous games [11]. We offer this formal model tentatively, as an initial step of a (hopefully) detailed research programme. As it stands, the model has no agency or epistemic attitudes or social aggregation. Instead it focusses only on how local attempts at signalling "importance" of an observation can spread and lead to some stable phenomenon that can be meaningfully construed as collective memory.

The crucial element here is that individuals perceive events differently, based on social and cultural background. The word *partition* would evoke the image of equivalence classes to a combinatorist in general, but very likely, that of a terrible tragedy first to an Indian combinatorist.

## 1.1 Related work

The literature on memory studies principally consists of two strands, one on individual memory, substantially incorporated into psychology, and the other on group memory, primarily on social representations of history ([7]). A major question of interest is whether such groups manifest *emergent*, robustly collective forms of memory. In general, while social context is seen to influence remembering, the act of remembering is held to be individual. However the literature on collective intentionality ([27]) considers collective memory as a form of collective attention to the past. The concepts underlying the model we discuss below are greatly inspired by the “*we-mode*”, discussed by the philosopher *Raimo Tuomela*, though we place an additional emphasis on patterns of social interaction (rather than the interactions themselves).

If we admit the existence of emergent collective memory, the major question is whether the processes of social collective memory resemble in any way the processes of individual memory or that of small groups (like families) studied extensively by psychologists. In an astonishing thesis from an interdisciplinary collaboration of neurobiology, medicine, cognitive science and anthropology, Anastasio et al ([2]) assert that these two processes are in fact the same. The model we present below, taking automata as memory representations, attempts to build a ‘social automaton’ (roughly speaking) in this spirit, showing a correspondence of processes.

The mechanism that we use in the model is closely related to that of *majority dynamics* ([15]) and *spread of influence* ([18]) studied in the analysis of social networks. There are also influential logical studies of diffusion in social networks ([5]). In other papers, we have studied similar phenomena in the context of *large games* ([20], [21]). While there is a correspondence at an intuitive level between memory systems and large games, a formal correspondence would lead to many potential applications.

The logic we propose here is a simple linear time temporal logic (with past) with variables and bounded quantification. There is extensive logical literature on the role of memory in multi-agent epistemic reasoning. For instance, the notion of bounded-recall plans has been studied in ([12], [8], [1]), to mention some recent works. However, this literature is principally on the interactions between individual memories; our point of departure is in studying emergent collective memory as an entity in itself. Moreover, the logic we discuss here is extremely poor in logical machinery, in comparison: there is no agency and no ascription of epistemic attitudes. Rather, we discuss only the temporal evolution of signalling interaction and the stability of signals. On the other hand, logical studies of social networks such as ([5]) and [17]) are closer in spirit to the automaton model we present, though the logics themselves are different.

## 2 A model

Let  $N$  denote a fixed finite set of agent names. Let  $\mathcal{C} \subseteq 2^N$  be a nonempty set of nonempty subsets of  $N$ , referred to as *neighbourhoods* over  $N$ . We assume  $|N| > 2$ .

For presenting the model we will make some simplifying assumptions. We fix a finite **signal alphabet**  $\Gamma$  common to all agents. Let  $\Gamma = \{\gamma_1, \dots, \gamma_m\}$ .

Let  $I \in \mathcal{C}$  and  $|I| = k$ . A **distribution** over  $I$  is an  $m$  tuple of integers  $\mathbf{y} = (y_1, \dots, y_m)$  such that

$y_j \geq 0$  and  $\sum_{j=1}^m y_j = k$ ,  $1 \leq j \leq m$ . That is, the  $j$ th component of  $\mathbf{y}$  gives the number of agents in the neighbourhood  $I$  who give signal  $\gamma_j$ . Let  $\mathbf{Y}[I]$  denote the set of all signal distributions of a neighbourhood  $I$  and let  $\mathbf{Y} = \bigcup_{I \in \mathcal{C}} \mathbf{Y}[I]$ .

The main idea is this. All agents initially receive an external input and assume some state. At each state, an agent produces a signal. Interactions occur in neighbourhoods nondeterministically, and an agent who is a member of many, could be interacting in different neighbourhoods (though every interaction at any instant is confined to one neighbourhood). Each interaction induces a state transition that is determined only by the distribution of signals: it does not depend on who is signalling what, but how many are producing each signal. Such interactions keep occurring repeatedly until a stable configuration is reached.

Below, for  $I \in \mathcal{C}$ , we use the notation  $\Gamma^I$  for a vector of signals, one signal for each of the agents in  $I$ . Note that every such vector induces a distribution over  $\Gamma$  in  $\mathbf{Y}[I]$ .

We will consider systems of agents below where the set of possible states  $Q$  is uniform for all agents. Thus a state transition in  $Q \times Q$  is also possible uniformly for all agents. By a triple  $(\gamma, q, q')$  we mean that an agent who is signalling  $\gamma$  changes state from  $q$  to  $q'$ .  $\sigma : \Gamma \rightarrow (Q \times Q)$  specifies such a signal based change of state. Let  $\Sigma$  denote the collection of such maps.

**Definition 1** A memory system over  $\mathcal{C}$  is a tuple  $M = (Q, \delta, \iota, \omega)$ , where

- $Q$  is a finite set of memory states,
- $\iota : N \rightarrow Q$  is the initial state,
- $\omega : Q \rightarrow \Gamma$  is the signalling function, and
- $\delta$  is a finite family of transition relations  $\delta_I \subseteq (\mathbf{Y}[I] \times \Sigma)$ , where  $I \in \mathcal{C}$ .

## 2.1 Dynamics

A configuration  $\chi$  is an element of  $Q^N$ . Let  $\omega(\chi)$  denote the vector in  $\Gamma^N$  induced by  $\omega$ . For  $I \in \mathcal{C}$  and let  $\mathbf{y}_I$  be a distribution of signals induced by the vector  $\omega(\chi)$  restricted to  $I$ .

**Definition 2** We say that an  $I$ -interaction is enabled at  $\chi$  if there is a transition  $(\mathbf{y}, \sigma)$  in  $\delta_I$  where  $\mathbf{y}$  is the distribution induced by  $\omega(\chi)$  and  $\sigma \in \Sigma$  is a signal-based change of state.

The effect of the transition is determined by  $\sigma$  and the new configuration  $\chi'$  is given by:

$$\chi'(j) = q', \text{ where } j \in I, \sigma(\omega(\chi(j))) = (\chi(j), q')$$

and  $\chi'(j) = \chi(j)$ , otherwise. Thus, we have a transition  $(\chi, I, \chi')$  on configurations labelled by neighbourhoods.

The dynamics of  $M$  is then given by a configuration graph  $G_M$  whose vertices are configurations and edges are labelled by neighbourhoods: an edge  $(\chi, I, \chi')$  is present if an  $I$ -interaction is enabled at  $\chi$  by a transition in  $\delta$  with resulting configuration  $\chi'$  as above. Note that  $\iota$  specifies an initial configuration  $\chi_0$ . A history  $\rho$  is any finite or infinite path in  $G_M$  starting from  $\chi_0$ . When  $\rho$  is finite,  $|\rho|$  denotes its length. Let  $\mathcal{H}_M$  denote the set of all maximal histories of  $M$ .

## 2.2 Collective memory in $M$

Consider a history  $\rho = \chi_0\chi_1\dots$  of system  $M$ . We say that signal  $\gamma$  is eventually stable for neighbourhood  $I$  in  $\rho$  if there exists  $k$  such that for all  $\ell \geq k$ , (when  $\rho$  is finite, for all  $\ell$  such that  $k \leq \ell \leq |\rho|$ ), and for all  $j \in I$ ,  $\omega(\chi_\ell(j)) = \gamma$ .

We say that  $\gamma$  is in collective memory in  $\rho$  if it is eventually stable for  $N$  in  $\rho$ : that is, no matter what interactions take place, all agents remain in states that emit signal  $\gamma$ .

When we have stable configurations, we see them as formation of collective memory. For this it is of course essential that a history allows for signalling to spread across neighbourhoods; if some neighbourhoods never interact, then signalling can remain confined within pockets. So we consider spanning histories where we impose the condition that every interaction that is infinitely often enabled (according to the transition rule) in an infinite history takes place infinitely often. This is a typical fairness condition used in the theory of computation, but weaker, or different, conditions may be sufficient for many systems. For instance, we might ask that the union of neighbourhoods in a history span all of  $N$ ; this merely says that all the agents have interacted at least once. Note that this depends on  $\delta$ : the distributions specified may already disable some agents or neighbourhoods from ever interacting.

**Definition 3** We say that the system  $M$  supports emergence of collective memory if, for every spanning history  $\rho$  of  $M$ , there exists a signal  $\gamma_\rho$  that is in collective memory in  $\rho$ .

## 2.3 An example

Consider a system with two signals  $\{g, b\}$ , standing for “good” and “bad”. There are only two states:  $G$  and  $B$ , signalling  $g$  and  $b$  respectively. Initially every agent perceives some global event as good or bad. Thus  $\iota$  is an arbitrary distribution of signals in the system.

The transition rule is simple: for any neighbourhood  $I$ , if more than half in  $I$  signal  $x$ , then all agents in  $I$  signal  $x$  in the new state. If they are exactly even, they continue evenly matched. Let  $I$  be a neighbourhood with  $k$  agents.  $\delta_I = \{(m, n), (g, q, G), (b, q, G) | m > n, m + n = k\} \cup \{(m, n), (g, q, B), (b, q, B) | m < n, m + n = k\} \cup \{(m, m), (s, q, q) | k = 2m, s \in \{g, b\}\}$ .

Now we can see that whether either of the signals becomes stable in a history depends on both the initial perception as well as which  $I$ -interactions are enabled. If odd-sized neighbourhoods can interact, a signal will begin to dominate. When the initial distribution is exactly even, and the interacting neighbourhoods are always split evenly, neither signal dominates the other.

On the other hand, suppose that a large fraction of the population receives the signal  $g$ , but  $\delta$  only enables neighbourhoods with the majority signalling  $b$  and a minority signalling  $g$  to interact. Then the signal  $b$  emerges as a stable signal, despite the initial distribution. This can be seen as the influence of social structures on collective memory.

Such behavioural analysis is common in the study of *runaway phenomena* ([6]) and so-called *informational cascades* ([9]).

## 2.4 Some subclasses

Consider a memory system in which all interactions are constrained to be pairwise. In such a system, the distribution profiles are entirely irrelevant, since given a pair of agents, there are four possible pairs of signals, determined by their states, and we only need to specify the resulting pair of states. Thus  $\delta \subseteq Q^4$ . Such systems have been studied as **population protocols** ([4]), for which a number of technical results are known.

We also have other interesting subclasses of systems: for instance, those where the transition relation is presented as  $\delta_k$  where  $1 < k \leq N$ . That is, only the size of a neighbourhood determines whether it can interact and not the identities of agents in it. (Note that even this restriction does not rule out the predatory dominance of a signal, as illustrated in the example above.)

Of great interest to social systems is where  $\mathcal{C}$  represents a hierarchy: we have an ordering  $<$  on  $N$  and impose the condition that all  $I \in \mathcal{C}$  are downward-closed with respect to the ordering relation.

However, note that these subclasses only constrain interaction structure and not the memory updates. This may be consonant with the discussion on collective memory in social theory, whereby social structures (and belief systems) constrain opportunity and influence, but what persists in social memory may well be impervious to social structures.

## 2.5 Computational power

Note that when  $|\Gamma| = d$ , every distribution over  $\Gamma$  is a  $d$ -dimensional vector in  $\mathbf{N}^d$ . The initial state  $\iota$  specifies such a distribution. Now consider a spanning history that is stably signalling  $\gamma$ ; we can consider this the output of the memory system for that history. Viewed thus, every system that supports emergence of collective memory can be said to *compute* a function from  $\mathbf{Y} \rightarrow \Gamma$ . Alternatively, we can consider such a function to be a predicate over  $\mathbf{N}^d \times \{1, \dots, d\}$ . Thus we can speak of arithmetical predicates computable by memory systems.

An important theorem in the study of population protocols guides us to the study of what memory systems can compute. In the study of population protocols, systems come with an *output function*, mapping to the two element output alphabet  $\{0, 1\}$  (without loss of generality). In this case we can consider the population protocol to be computing a predicate over  $\mathbf{N}^d$ .

Recall that a *semi-linear set* is a subset of  $\mathbf{N}^d$  that is a finite union of *linear* sets of the form  $\{\mathbf{b} + k_1 \mathbf{a}_1 + \dots + k_m \mathbf{a}_m \mid k_1, \dots, k_m \in \mathbf{N}\}$ , where  $\mathbf{b} \in \mathbf{N}^d$  and  $\mathbf{a}_1 \dots \mathbf{a}_m$  are  $d$ -dimensional basis vectors.

**Theorem 1 [3]:** *A predicate is computable by a population protocol iff it is semi-linear.*

An alternative characterization of these predicates is that they can be expressed in first-order Presburger arithmetic, which is first order arithmetic on the natural numbers with addition but not multiplication.

**Theorem 2** *Given a memory system  $M$ , checking whether  $M$  supports emergence of collective memory is decidable. Moreover the class of predicates computable by memory systems is exactly that of population protocols.*

There are two parts to the proof. We construct a Parikh automaton ([16]) that represents the configuration space of the memory system and reduce the check for stability of signals in the system to the nonemptiness problem for the associated Parikh automaton. This gives us the required decision procedure.

In the process we show that every predicate computed by a memory system is semi-linear. By the earlier theorem, such a predicate can be computed by a population protocol. Conversely, since population protocols are a subclass of memory systems, the predicates computed by the former are computable by memory systems as well. (There are some details related to the output function, and the restriction to spanning histories, which complicate the construction a little.)

To get an intuitive idea of the construction, we define Parikh automaton below, which in turn needs the definition of Presburger arithmetic.

Firstly, let  $\Delta = \{a_1, \dots, a_m\}$  be any finite alphabet, and  $w \in \Delta^*$ . The Parikh image of  $w$  counts the number of occurrences of each letter of the alphabet in  $w$ . Formally, we have the map  $\pi : \Delta^* \rightarrow \mathbf{N}^m$  given by:  $\pi(a_i) = e_i$  and  $\pi(uv) = \pi(u) + \pi(v)$ , where  $e_i$  is the unit vector of length  $m$  where the  $i^{\text{th}}$  coordinate is 1.

Clearly the map  $\pi$  can be lifted to alphabets of the form  $\Delta \times D$  where  $D \subseteq \mathbf{N}^d$ :  $\pi(a_i, \hat{k}) = \hat{k}$  and  $\pi(uv) = \pi(u) + \pi(v)$ . We can also consider the alphabetic projection into  $D^*$ :  $\lambda(a_i, \hat{k}) = a_i$  and  $\lambda(uv) = \lambda(u)\lambda(v)$ .

Presburger arithmetic is first order logic with the only atomic formulas of the form  $t \text{ rel } t'$  where  $t$  and  $t'$  are terms, and  $\text{rel} \in \{>, <, \geq, \leq\}$ . Terms are built from two constants 0 and 1, and variables, using addition and  $n \cdot t$  where  $n \in \mathbf{N}$ . Formulas are interpreted over the structure  $\mathcal{N} = (\mathbf{N}, +, \cdot, 0, 1)$ . When  $\phi(\hat{x})$  is a formula with free variables  $\hat{x}$  the notion  $\mathcal{N}, \hat{k} \models \phi(\hat{x})$  is defined in the standard fashion.

Given  $D \subseteq \mathbf{N}^d$  and  $L \subseteq (\Delta \times D)^*$ , and a formula  $\phi$  of Presburger arithmetic, we define  $L \upharpoonright_{\phi} = \{\lambda(w) \mid \mathcal{N}, \pi(w) \models \phi(\hat{x})\}$ .

**Definition 4** A **Parikh automaton** of dimension  $d > 0$  is a pair  $(A, \phi)$  where  $\phi(x_1, \dots, x_d)$  is a formula of Presburger arithmetic over  $d$  variables, and  $A$  is a finite word automaton with the finite alphabet  $\Delta \times D$  where  $D \subseteq \mathbf{N}^d$ . We say that  $(A, \phi)$  recognizes  $L(A, \phi) = L(A) \upharpoonright_{\phi}$ , where  $L(A)$  is the language recognized by the automaton  $A$ .

For the construction we need, there are some points to note. Configurations of memory systems carry states, from which we compute signal distributions which cause state changes. For the Parikh automaton, transitions are labelled by distributions. More importantly we need to carry the neighbourhood based signalling in the transitions of the Parikh automaton. While these are matters of detail, the harder part of the proof is the definition of the formula of Presburger arithmetic, for which we closely follow the proof method for population protocols ([3]).

### 3 A logic for the rationale

We now turn our attention to our principal logical interest, namely, the rationale by which agents decide what signals are chosen. This is surely complex, depending on the systems being modelled. Here we propose a minimal logic, in which agents evaluate signals based on their own evaluation of the current state and their evaluation of signals from the neighbourhood, depending on the signal distributions.

Let  $V$  be a countable set of variables. Let the terms of the logic be defined as

$$\tau ::= i \mid x, i \in N, x \in V$$

That is, a term is either an agent name or a variable (which takes agents as its values). Let  $\mathcal{P}$  denote a countable set of atomic propositional symbols.

The formulas of the logic are built using the following syntax:

$$\begin{aligned} \Phi ::= & \tau_1 = \tau_2 \mid \tau \in I \mid p@ \tau, p \in \mathcal{P} \mid \gamma@ \tau, \gamma \in \Gamma \mid \neg \phi \mid \\ & \phi_1 \vee \phi_2 \mid \ominus \phi \mid \bigcirc \phi \mid \diamond \phi \mid \heartsuit \phi \mid \#x \cdot \phi(x) \text{ op } k \end{aligned}$$

where  $\tau_1$  and  $\tau_2$  are terms,  $I \in \mathcal{C}$ ,  $\text{op} \in \{=, \neq, <, \leq, >, \geq\}$  and  $k \in \{0, \dots, |N|\}$ .

The formula  $\tau \in I$  asserts that the agent denoted by the term  $\tau$  is a member of the neighbourhood  $I$ .  $p@ \tau$  asserts that the condition  $p$  holds for agent  $\tau$ , and  $\gamma@ \tau$  specifies that  $\tau$  is signalling  $\gamma$  at that instant.

$\sharp$  is a counting quantifier, and  $\sharp x \cdot \phi(x) \geq k$  (for example) says that at least  $k$  agents support the assertion  $\phi$  at the instant. The modalities  $\ominus$  and  $\bigcirc$  denote the predecessor and successor instants, whereas  $\diamond$  and  $\diamond$  denote some time in the past and some time in the future, respectively. Their dual modalities are denoted  $\boxplus$  and  $\boxminus$  respectively. We talk of free and bound occurrences of variable  $x$  in formula  $\phi$  in the standard manner.  $\phi$  is said to be a sentence if it has no free occurrences of variables.

The existential and universal quantifiers are defined easily:  $\exists x \cdot \phi(x) = \sharp x \cdot \phi(x) > 0$  and  $\forall x \cdot \phi(x) = \neg \exists x \neg \phi(x)$ . We use the abbreviation  $\gamma@I = \forall x \cdot (x \in I \supset \gamma@x)$  to denote that all agents in the neighbourhood signal  $\gamma$ .

The formula  $\boxplus \forall x \cdot \gamma@x$  is denoted  $stable(\gamma)$ . The formula  $\bigvee_{\gamma \in \Gamma} stable(\gamma)$  is special and is called **emergence**.

The semantics is defined on histories. A model is a tuple  $(M, val, \eta)$ , where  $M$  is a memory system,  $\eta : V \rightarrow N$  denote an assignment of agent variables to agent names and  $val : \mathcal{Q} \rightarrow 2^{\mathcal{P}}$  is the propositional valuation map.  $val$  is lifted to configurations by the map  $\hat{val} : \mathcal{Q}^N \rightarrow (N \rightarrow 2^{\mathcal{P}})$  by:  $\hat{val}(\chi)(i) = val(\chi(i))$ .

Let  $\rho \in \mathcal{H}_M$  be a finite or infinite history  $\rho = \chi_0 \chi_1 \dots$ . The notion that  $\rho, k \models \phi$  is defined in the standard fashion, for  $k \geq 0$ .

- $\rho, k \models \tau_1 = \tau_2$  iff  $\eta(\tau_1) = \eta(\tau_2)$ .
- $\rho, k \models \tau \in I$  iff  $\eta(\tau) \in I$ .
- $\rho, k \models p@ \tau$  iff  $p \in \hat{val}(\rho_k)(\eta(\tau))$ .
- $\rho, k \models \gamma@ \tau$  iff  $\omega(\rho_k(\tau)) = \gamma$ .
- $\rho, k \models \neg \phi$  iff  $\rho, k \not\models \phi$ .
- $\rho, k \models \phi_1 \vee \phi_2$  iff  $\rho, k \models \phi_1$  or  $\rho, k \models \phi_2$ .
- $\rho, k \models \ominus \phi$  iff  $k > 0$  and  $\rho, k-1 \models \phi$ .
- $\rho, k \models \bigcirc \phi$  iff there exists a successor instant in the history and  $\rho, k+1 \models \phi$ .
- $\rho, k \models \diamond \phi$  iff there exists  $\ell \leq k$  such that  $\rho, \ell \models \phi$ .
- $\rho, k \models \diamond \phi$  iff there exists  $\ell \geq k$  such that  $\rho, \ell \models \phi$ .
- $\rho, k \models \sharp x \cdot \phi(x) \text{ op } k$  iff  $|\{j \mid \rho, k \models \phi[j/x]\}| \text{ op } k$ .

The notions of satisfiability and validity are standard. Given a model  $(M, val)$  and a sentence  $\phi$  by  $M \models \phi$  we denote that for all histories  $\rho$  in  $\mathcal{H}_M$ ,  $\rho, 0 \models \phi$ .

### 3.1 Examples

It is easily seen that specific distributions can be described in the logic. For instance consider the distribution where there are 100 agents, 30 of whom signal  $a$  and the rest signal  $b$ . We can specify this as:

$$\sharp x \cdot \gamma_a@x = 30 \wedge \sharp x \cdot \gamma_b@x = 70$$

With inequalities, classes of distributions that lead to the same signalling' behaviour can be specified.

Further the structure of interaction in histories can be constrained in the logic:

$$\exists x \cdot (x \in I \supset \boxplus(x \in I))$$

This asserts that once an agent participates in an interaction in the history, it continues to participate in every interaction in the history. (In social theory, such persistent actors in the structure are associated with memorials that keep reminding everyone of a memory token.)

The state transition structure of memory systems can be described only to a limited extent since the logic is first order and the state information which may include modular counting of signals cannot be expressed in it. However, propositional updates based on signal distributions can be specified.

The logic can describe various kinds of signalling schemes by agents.

- Signal  $a$  and  $b$  alternatively:

$$\Box[(\neg\gamma_a@i \supset \gamma_b@i) \wedge (\neg\gamma_b@i \supset \gamma_a@i)]$$

- If more than 5 agents in my neighbourhood previously signalled  $a$  then I signal  $a$ :

$$(i \in I \wedge \#x.(x \in I \wedge \neg\gamma_a@x) > 5) \supset \gamma_a@i$$

- $\gamma$  is collective memory:

$$\Diamond\Box(\forall x.\gamma@x)$$

- $\gamma$  is collective amnesia:

$$\Diamond\Box(\forall x.\neg\gamma@x)$$

However, in terms of validities, the logic has little structure to force validities beyond that of linear time temporal logic. The interest of the logic is mainly in its role as a specification language for requirements on memory systems, and hence we are more interested in checking whether a specific memory system satisfies such a specification.

### 3.2 Model checking

In general, first order temporal logics are highly undecidable and non-axiomatizable. However, what we have here is bounded quantification, since the set of agents over which we quantify, is fixed and finite. So we can effectively eliminate quantifiers and translate formulas into propositional linear time temporal logic. Thus satisfiability is decidable via automaton construction, but yet, extracting a memory system from the formula automaton has some interesting details.

The model checking problem for the logic asks, given a model  $(M, val)$  and a sentence  $\varphi$ , whether  $M \models \varphi$ .

**Theorem 3** *The model checking problem for the logic is decidable in time linear in  $M$  and singly exponential in  $\varphi$ .*

In this case, we construct a Parikh automaton that represents the configuration space of the memory system, and take its product with the formula automaton associated with the given sentence  $\varphi$ . The construction is straightforward, though not entirely trivial.

## 4 Discussion

We began with the intention of reasoning about collective memory. How do systems of signalling in neighbourhoods embody reasoning?

Firstly, it should be clear that the history model and the logical language are rich enough to talk about the “remembrance of things past”. However, the model assumes perfect observability for agents within neighbourhoods within a fixed interaction structure. Both of these assumptions need to be relaxed, leading to epistemic logics.

Such a logical exercise is not sufficient in itself to uncover the process of *stipulation* mentioned by Sontag, or the *interdependence* between memory and reasoning demanded by Wittgenstein. However, in our opinion, the model holds considerable promise. For achieving the richness required, we hold two features to be essential: reinforcement of memory that comes through repeated interactions inside local neighbourhoods, but not confined to those neighbourhoods; complex social rules that determine influence in signalling. Elements of both are present in this model.

Further, while we have presented simple stability as the basic notion of collective memory which is persistent, we can symmetrically study notions like *collective amnesia* in the model whereby signals predominant in the system lose out in interactions and ultimately vanish. Incorporating both, to study collective memory of some events while at the same time forgetting others presents no logical difficulties but makes the model considerably complex to reason about, requiring a structural insight for simplification.

Moreover, we have taken collective memory to mean the entire system. We can instead parameterize such memory by a subset of agents to get group notions of memory; this is relevant in the study of notions such as common ground for communication ([10]).

It is interesting to consider dynamic memory systems, whereby endogenous changes in signalling behaviour can lead to altering interaction structures leading to new update rules. In particular, neighbourhoods need not be static, but may expand and contract. This is analogous to dynamic form games as we have studied elsewhere ([20]).

Social choice theory offers a variety of methods for aggregation of information from individuals for collectives, but these are static rules. Automata models such as the ones studied here can offer dynamic methods of aggregation whereby we formulate ‘local’ aggregation rules which are applied repeatedly. Whether this can lead to meaningful insight for social theories remains to be seen, offering interesting technical questions for study in the meanwhile.

Interactions in the model are nondeterministic. Stochastic models of interactions may be more appropriate for social behaviour, relevant to social network studies such as ([15], [18]). However, logicising the rationale of such interaction would perhaps need a different approach.

## 5 Acknowledgement

We thank Oliver Roy and Wang Yi for discussions on the theme, and the TARK reviewers for thoughtful remarks and helpful comments.

## References

- [1] Natasha Alechina & Brian Logan (2010): *Belief ascription under bounded resources*. *Synth.* 173(2), pp. 179–197, doi:10.1007/s11229-009-9706-6.

- [2] Thomas J Anastasio, Kristen Ann Ehrenberger, Patrick Watson & Wenyi Zhang (2012): *Individual and collective memory consolidation: Analogous processes on different levels*. MIT Press, doi:10.7551/mitpress/9173.001.0001.
- [3] Dana Angluin, James Aspnes, David Eisenstat & Erik Ruppert (2007): *The computational power of population protocols*. *Distributed Computing* 20(4), pp. 279–304, doi:10.1007/s00446-007-0040-2.
- [4] James Aspnes & Eric Ruppert (2009): *An Introduction to Population Protocols*, pp. 97–120. Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-540-89707-1\_5.
- [5] Alexandru Baltag, Zoé Christoff, Rasmus K. Rendsvig & Sonja Smets (2019): *Dynamic Epistemic Logics of Diffusion and Prediction in Social Networks*. *Studia Logica* 107(3), pp. 489–531, doi:10.1007/s11225-018-9804-x.
- [6] Abhijit V. Banerjee (1992): *A Simple Model of Herd Behaviour*. *The Quarterly Journal of Economics* 107(3), pp. 797–817, doi:10.2307/2118364.
- [7] Jeffrey Andrew Barash (2016): University of Chicago Press, doi:10.7208/9780226399294-toc.
- [8] Francesco Belardinelli, Alessio Lomuscio & Emily Yu (2020): *Model Checking Temporal Epistemic Logic under Bounded Recall*. In: *Proceedings AAAI 2020*, AAAI Press, pp. 7071–7078, doi:10.1609/aaai.v34i05.6193.
- [9] Zoé Christoff & Jens Ulrik Hansen (2015): *A logic for diffusion in social networks*. *J. Appl. Log.* 13(1), pp. 48–77, doi:10.1016/j.jal.2014.11.011.
- [10] Herbert H Clark & Susan E Brennan (1991): *Grounding in communication*. American Psychology Association, doi:10.1037/10096-006.
- [11] C. Daskalakis & C. H. Papadimitriou (2007): *Computing equilibria in anonymous games*. In: *Proceedings of the 48th symposium on Foundations of Computer Science (FOCS)*, IEEE Computer Society Press, pp. 83–93, doi:10.1109/FOCS.2007.24.
- [12] Kaya Deuser & Pavel Naumov (2020): *On composition of bounded-recall plans*. *Artif. Intell.* 289, p. 103399, doi:10.1016/j.artint.2020.103399.
- [13] Margaret Gilbert (1987): *Modelling Collective Belief*. *Synthese* 73(1), pp. 185–204, doi:10.1007/BF00485446.
- [14] Maurice Halbwachs (1950): *The Collective Memory*. Harper and Row, New York.
- [15] David Kempe, Jon Kleinberg & Éva Tardos (2015): *Maximizing the Spread of Influence through a Social Network*. *Theory of Computing* 11(4), pp. 105–147, doi:10.4086/toc.2015.v011a004.
- [16] Felix Klaedtke & Harald Rueß (2003): *Monadic Second-Order Logics with Cardinalities*. In Jos C. M. Baeten, Jan Karel Lenstra, Joachim Parrow & Gerhard J. Woeginger, editors: *Automata, Languages and Programming*, Springer Berlin Heidelberg, pp. 681–696, doi:10.1007/3-540-45061-0\_54.
- [17] Fenrong Liu, Jeremy Seligman & Patrick Girard (2014): *Logical dynamics of belief change in the community*. *Synthese* 191(11), pp. 2403–2431, doi:10.1007/s11229-014-0432-3.
- [18] Elchanan Mossel, Joe Neeman & Omer Tamuz (2013): *Majority dynamics and aggregation of information in social networks*. *Autonomous Agents and Multi-Agent Systems* 28(3), p. 408–429, doi:10.1007/s10458-013-9230-4.
- [19] Rohit Parikh (2012): *What Is Social Software?*, pp. 3–13. LNCS 7010, Springer, doi:10.4204/EPTCS.
- [20] Soumya Paul & R. Ramanujam (2013): *Dynamics of Choice restriction in Large Games*. *IGTR* 15(4), doi:10.1142/S0219198913400318.
- [21] Soumya Paul & R. Ramanujam (2014): *Subgames within Large Games and the Heuristic of Imitation*. *Studia Logica* 102(2), pp. 361–388, doi:10.1007/s11225-014-9549-0.
- [22] William G. Roy & Rachel Parker-Gwin (1999): *How Many Logics of Collective Action? Theory and Society* 28(2), pp. 203–237, doi:10.1023/A:1006946310119.

- [23] Mihai Rusu (2017): *Transitional Politics of Memory: Political Strategies of Managing the Past in Post-communist Romania*. *Europe-Asia Studies*, pp. 260–282, doi:10.1080/09668136.2017.1380783.
- [24] Susan Sontag (2003): *Regarding the pain of others*. Picador, doi:10.3917/dio.201.0127.
- [25] John Sutton (2014): *Remembering as Public Practice: Wittgenstein, memory, and distributed cognitive ecologies*. doi:10.1515/9783110378795.409.
- [26] Allard Tamminga, Hein Duijf & Frederik Van De Putte (2020): *Expressivity results for deontic logics of collective agency*. *Synthese*, pp. 260–282, doi:10.1007/s11229-020-02597-0. Forthcoming.
- [27] Raimo Tuomela (2013): *Social Ontology: Collective Intentionality and Group Agents*. OUP USA, doi:10.1515/jso-2014-0040.
- [28] Ludwig Wittgenstein (1967): *Zettel*. Univ of California Press.

# A Deontic Stit Logic Based on Beliefs and Expected Utility

Aldo Iván Ramírez Abarca

Utrecht University  
Utrecht, The Netherlands

boiangaleano@hotmail.com

Jan Broersen

Utrecht University  
Utrecht, The Netherlands

J.M.Broersen@uu.nl

The formalization of action and obligation using logic languages is a topic of increasing relevance in the field of ethics for AI. Having an expressive syntactic and semantic framework to reason about agents' decisions in moral situations allows for unequivocal representations of components of behavior that are relevant when assigning blame (or praise) of outcomes to said agents. Two very important components of behavior in this respect are belief and belief-based action. In this work we present a logic of doxastic oughts by extending epistemic deontic stit theory with beliefs. On one hand, the semantics for formulas involving belief operators is based on probability measures. On the other, the semantics for doxastic oughts relies on a notion of optimality, and the underlying choice rule is maximization of expected utility. We introduce an axiom system for the resulting logic, and we refer to its soundness, completeness, and decidability results. These results are significant in the line of research that intends to use proof systems of epistemic, doxastic, and deontic logics to help in the testing of ethical behavior of AI through theorem-proving and model-checking.

## 1 Introduction

It has been argued that an appropriate theory of agency and obligation should take into account what agents know—and what they know how to do—both before and at the moment of acting ([41], [22], [21], [13]). Following considerations from epistemic game theory (EGT)—which has clear conceptual and technical connections to stit theory (see [16, Chapter 1] and [37])—we put forward that agents' beliefs also play an important role in the relation between agency and obligation. In recent years we have seen a growing interest in adding *knowledge* modalities to stit theory, but there are relatively few extensions of this logic by means of *belief* operators (notably those in [39] and [12]). The novelty of the present approach lies in its intention to develop a link between beliefs and ought-to-do. This is a natural step to take in the line of both Horty's formalization of *act utilitarian ought-to-do* ([23]) and its extension with epistemic notions ([21], [13]).

According to Horty ([23]), act utilitarian ought-to-do stems from a measure of *optimality* of actions. The consequences of optimal actions are taken to be the conditions that agents ought to have brought about in the world. Horty's idea of *optimality* is undeniably inspired by solution concepts from game theory, particularly by dominance of strategies. Following this idea, the recent works [21] and [1] introduce *epistemic* (resp. *subjective*) ought-to-do's to account for the relation between knowledge and obligation. To be more precise, in both these works the optimal actions once again underlie what agents epistemically (resp. subjectively) ought to do, but the measure of optimality now takes into consideration agents' epistemic states. The resulting formalization can deal with complex scenarios for which the initial—non-epistemic—ought-to-do fell short of very intuitive standards in the context of responsibility attribution. Since stit theory can—at least in principle—incorporate most game theoretic ideas into multi-agent action-settings, we find it worthy to extend the theory of ought-to-do with a notion of belief. Our long-term goal is to achieve a nuanced formalization of obligation and responsibility, where agents' hierarchies of belief would serve as explanations of the actions that these agents perform interactively.

Consider the following example, inspired by the famous film *The Verdict*, of 1982. Suppose that a patient is admitted to the hospital in urgent need of surgery. The nurses draw up a chart with important background information for the surgeons, but unfortunately the figure regarding how long it has been since the patient last ate has a mistake. Anesthetics for this surgery should only be supplied if the patient has had an empty stomach for at least eight hours, and they are deadly otherwise. Because of the mistake in the chart, the anesthesiologist never comes to know that the patient had had a full meal just one hour before admittance. Therefore, she gives the anesthetics and the patient dies. It is clear that the doxastic state of the anesthesiologist plays a key role in determining whether she is morally responsible for the patient's death. On the causal level, the anesthesiologist is responsible for it. However, it seems natural that she should not be held culpable, because she acted upon the false—but justified—belief that the patient had an empty stomach before admittance. Moreover, the anesthesiologist is justified in considering the action of supplying the anesthetics as something that she ought to have done, given the circumstances.

There are several options of conceptual backgrounds for incorporating beliefs into deontic stit logic (see [39], ([4], and [20] for some alternatives). In this work we adopt a *quantitative* version of agentive belief, and we use probability measures (on the domain of stit structures) to represent doxastic states. An agent's subjective belief in a proposition is taken to depend on the probability that the agent assigns to the indices at which the propositions holds. The reason for choosing a probabilistic semantics of belief is that it allows us to base a notion of *doxastic ought* on a key concept in decision theory: *maximization of expected utility*. Thus, we propose that at a given index of evaluation an agent had the doxastic obligation to see to it that  $\varphi$  if  $\varphi$  is a consequence of all the actions that maximized expected (deontic) utility at such index, where this utility is identified with the deontic value of the index just as in [23]. The basic aspects of the logic that we develop here, then, can be summarized in this way: we extend the basic stit language with (a) epistemic and doxastic operators, (b) objective and subjective ought-to-do operators (whose logic was developed in [1]), and (c) a doxastic ought-to-do operator, meant to build up formulas to characterize the effects of those actions that agents' belief-systems render as optimal—i.e., the effects of rational *best responses* (see [8]).

The paper is structured as follows. In Section 2 we introduce the syntax and semantics of the epistemic deontic logic that we use as basis for our theory of doxastic oughts. In Section 3 we present the probabilistic conceptualization of belief that we intend to incorporate into the logic of Section 2, and we elaborate on the reasons for choosing such a notion of belief. In Section 4 we introduce the syntax and semantics for formulas involving the doxastic ought-to-do operator, and we discuss an example to illustrate its reach within a stit-theoretic analysis of responsibility and excusability. In Section 5 we develop an axiomatic system for the resulting logic and address its soundness, completeness, and decidability results, after which we conclude.

## 2 Epistemic Deontic Stit

As discussed in [21], [13], and [1], an adequate theory of ought-to-do should account for agents' epistemic states before and at the moments of action. This is all the more relevant in contexts of responsibility attribution, and more specifically for excusability (see, for instance, [27]). In principle, if an agent does not know how to fulfill an obligation, it should be excused for not having done so. The mentioned [21], [13], and [1] all extend Horty's stit theory of act utilitarian ought-to-do with knowledge operators and explore the relation between uncertainty and obligation. In turn, here we will be extending the logics of

[13] and [1] with modalities for belief.<sup>1</sup>

**Definition 1 (Language)** Given a finite set  $\text{Ags}$  of agent names, a countable set of propositions  $P$  such that  $p \in P$  and  $\alpha \in \text{Ags}$ , the grammar for the formal language  $\mathcal{L}_{\text{K0}}$  is given by:

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \psi \mid \Box\varphi \mid [\alpha]\varphi \mid K_\alpha\varphi \mid \odot[\alpha]\varphi \mid \odot_{\mathcal{S}}[\alpha]\varphi .$$

$\Box\varphi$  is meant to express the ‘historical necessity’ of  $\varphi$  ( $\Diamond\varphi$  abbreviates  $\neg\Box\neg\varphi$ ).  $[\alpha]\varphi$  stands for ‘agent  $\alpha$  has seen to it that  $\varphi$ .’  $K_\alpha$  is the epistemic operator for  $\alpha$ , so that  $K_\alpha\varphi$  stands for ‘agent  $\alpha$  knows that  $\varphi$  holds.’  $\odot[\alpha]\varphi$  is meant to express that  $\alpha$  objectively ought to have seen to it that  $\varphi$ . Finally,  $\odot_{\mathcal{S}}[\alpha]\varphi$  is meant to express that  $\alpha$  subjectively ought to have seen to it that  $\varphi$ .

As for the semantics, the structures on which we evaluate formulas of the language  $\mathcal{L}_{\text{K0}}$  are based on what we call *epistemic act-utilitarian branching-time frames*.

**Definition 2 (Epistemic act-utilitarian branching-time frames)** A *finite epistemic act-utilitarian branching-time frame* (eaubt-frame for short) is a tuple  $\langle M, \sqsubset, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha \in \text{Ags}}, \mathbf{Value} \rangle$  such that:

- $M$  is a non-empty finite set of moments and  $\sqsubset$  is a strict partial ordering on  $M$  satisfying ‘no backward branching.’ Each maximal  $\sqsubset$ -chain is called a *history*, which represents a way in which time might evolve.  $H$  denotes the set of all histories, and for each  $m \in M$ ,  $H_m := \{h \in H; m \in h\}$ . Tuples  $\langle m, h \rangle$  are called *indices* iff  $m \in M$ ,  $h \in H$ , and  $m \in h$ . **Choice** is a function that maps each agent  $\alpha$  and moment  $m$  to a partition  $\mathbf{Choice}_\alpha^m$  of  $H_m$ , where the cells of such a partition represent  $\alpha$ ’s available actions at  $m$ . For  $m \in M$  and  $h \in H_m$ , we denote the equivalence class of  $h$  in  $\mathbf{Choice}_\alpha^m$  by  $\mathbf{Choice}_\alpha^m(h)$ . **Choice** satisfies two constraints: (NC) No choice between undivided histories: For all  $h, h' \in H_m$ , if  $m' \in h \cap h'$  for some  $m' \sqsupset m$ , then  $h \in L$  iff  $h' \in L$  for every  $L \in \mathbf{Choice}_\alpha^m$ . (IA) Independence of agency: A function  $s$  on  $\text{Ags}$  is called a *selection function* at  $m$  if it assigns to each  $\alpha$  a member of  $\mathbf{Choice}_\alpha^m$ . If we denote by  $\mathbf{Select}^m$  the set of all selection functions at  $m$ , then we have that for every  $m \in M$  and  $s \in \mathbf{Select}^m$ ,  $\bigcap_{\alpha \in \text{Ags}} s(\alpha) \neq \emptyset$  (see [7] for a discussion of the property).
- For  $\alpha \in \text{Ags}$ ,  $\sim_\alpha$  is the *epistemic indistinguishability equivalence relation* for agent  $\alpha$ , which satisfies the following constraints: (OAC) Own action condition: For every index  $\langle m_*, h_* \rangle$ , if  $\langle m_*, h_* \rangle \sim_\alpha \langle m, h \rangle$  for some  $\langle m, h \rangle$ , then  $\langle m_*, h_* \rangle \sim_\alpha \langle m, h \rangle$  for every  $h'_* \in \mathbf{Choice}_\alpha^{m_*}(h_*)$ . We refer to this constraint as the ‘own action condition’ because it implies that agents do not know more than what they perform. (Unif – H) Uniformity of historical possibility: For every index  $\langle m_*, h_* \rangle$ , if  $\langle m_*, h_* \rangle \sim_\alpha \langle m, h \rangle$  for some  $\langle m, h \rangle$ , then for every  $h'_* \in H_{m_*}$  there exists  $h' \in H_m$  such that  $\langle m_*, h'_* \rangle \sim_\alpha \langle m, h' \rangle$ . Combined with (OAC), this constraint is meant to capture a notion of uniformity of strategies, where epistemically indistinguishable indices should have the same available actions for the agent to choose upon.

For each index  $\langle m, h \rangle$  and  $\alpha \in \text{Ags}$ , we define  $\alpha$ ’s *information set* at  $\langle m, h \rangle$  as  $\pi_\alpha[\langle m, h \rangle] := \{\langle m', h' \rangle; \langle m, h \rangle \sim_\alpha \langle m', h' \rangle\}$ .

- **Value** is a *deontic function* that assigns to each history  $h \in H$  a real number, representing the utility of  $h$ .

As for the deontic dimension, *objective* ought-to-do’s come from the optimal actions for an agent: to have seen to it that  $\varphi$  is taken to be an objective obligation of an agent at a given index iff  $\varphi$  is an

<sup>1</sup>The models that we use are simpler and at the same time more general than the ones of [21]. The same [13] and [1] discuss some advantages of their models over the ones developed in [21].

effect of all the optimal actions for that agent at that index. The optimality of such actions is relative to a dominance ordering, and this ordering depends on the value of the histories in those actions (provided by **Value**). In order to present the semantics for formulas involving the ought-to-do operator, we therefore need some previous definitions.

For  $m \in M$  and  $\beta \in \text{Ags}$ , we define  $\mathbf{State}_\beta^m = \left\{ S \subseteq H_m; S = \bigcap_{\alpha \in \text{Ags} - \{\beta\}} s(\alpha), \text{ where } s \in \mathbf{Select}^m \right\}$ . For  $\alpha \in \text{Ags}$  and  $m_* \in M$ , we first define a general ordering  $\leq$  on  $\mathcal{P}(H_{m_*})$  such that for  $X, Y \subseteq H_{m_*}$ ,  $X \leq Y$  iff  $\mathbf{Value}(h) \leq \mathbf{Value}(h')$  for every  $h \in X, h' \in Y$ . The objective dominance ordering  $\preceq$  is defined such that for  $L, L' \in \mathbf{Choice}_\alpha^{m_*}$ ,  $L \preceq L'$  iff for each  $S \in \mathbf{State}_\alpha^{m_*}$ ,  $L \cap S \leq L' \cap S$ . The optimal set of actions is the set  $\mathbf{Optimal}_\alpha^{m_*} := \{L \in \mathbf{Choice}_\alpha^{m_*}; \text{there is no } L' \in \mathbf{Choice}_\alpha^{m_*} \text{ such that } L \prec L'\}$ .

As for *subjective* ought-to-do's, they involve a dominance ordering as well, but one different to the one for objective ought-to-do's. To define this subjective dominance ordering, [13] introduces a new semantic concept known as *epistemic clusters*, which are nothing more than a given action's epistemic equivalents in indices that are indistinguishable to the one of evaluation. Formally, we have that for  $\alpha \in \text{Ags}$ ,  $m_*, m \in M$ , and  $L \subseteq H_{m_*}$ ,  $L$ 's *epistemic cluster* at  $m$  is the set  $[L]_\alpha^m := \{h \in H_m; \exists h_* \in L \text{ s.t. } \langle m_*, h_* \rangle \sim_\alpha \langle m, h \rangle\}$ . As a convention, we write  $m \sim_\alpha m'$  if there exist  $h \in H_m, h' \in H_{m'}$  such that  $\langle m, h \rangle \sim_\alpha \langle m', h' \rangle$ . A subjective dominance ordering  $\preceq_s$  is then defined on  $\mathbf{Choice}_\alpha^{m_*}$  by the following rule: for  $L, L' \subseteq H_{m_*}$ ,  $L \preceq_s L'$  iff for each  $m$  such that  $m_* \sim_\alpha m$ , for each  $S \in \mathbf{State}_\alpha^m$ ,  $[L]_\alpha^m \cap S \leq [L']_\alpha^m \cap S$ . Just as in the case of objective ought-to-do's, this ordering allows us to think about a subjectively optimal set of actions  $\mathbf{S-optimal}_\alpha^{m_*} := \{L \in \mathbf{Choice}_\alpha^{m_*}; \text{there is no } L' \in \mathbf{Choice}_\alpha^{m_*} \text{ s.t. } L \prec_s L'\}$ , where we write  $L \prec_s L'$  iff  $L \preceq_s L'$  and  $L' \not\preceq_s L$ . Analogous to what we mentioned regarding objective ought-to-do's, the idea is that to have seen to it that  $\varphi$  is a subjective obligation of an agent at a given index iff it is an effect of all the subjectively optimal actions—and their epistemic equivalents—for that agent at that index.

As is customary, the models and the semantics for the formulas are defined by adding a valuation function to the frames of Definition 2:

**Definition 3** An *eaubt-model*  $\mathcal{M}$  consists of the tuple that results from adding a valuation function  $\mathcal{V}$  to a *eaubt-frame*, where  $\mathcal{V} : P \rightarrow 2^{M \times H}$  assigns to each atomic proposition a set of moment-history pairs. Relative to a model  $\mathcal{M}$ , the semantics for the formulas of  $\mathcal{L}_{\text{KOBDO}}$  is defined recursively by the following truth conditions, evaluated at a given index  $\langle m, h \rangle$ :

$\mathcal{M}, \langle m, h \rangle \models p$	iff	$\mathcal{M}, \langle m, h \rangle \in \mathcal{V}(p)$
$\mathcal{M}, \langle m, h \rangle \models \neg \varphi$	iff	$\mathcal{M}, \langle m, h \rangle \not\models \varphi$
$\mathcal{M}, \langle m, h \rangle \models \varphi \wedge \psi$	iff	$\mathcal{M}, \langle m, h \rangle \models \varphi$ and $\mathcal{M}, \langle m, h \rangle \models \psi$
$\mathcal{M}, \langle m, h \rangle \models \Box \varphi$	iff	for each $h' \in H_m, \mathcal{M}, \langle m, h' \rangle \models \varphi$
$\mathcal{M}, \langle m, h \rangle \models [\alpha] \varphi$	iff	for each $h' \in \mathbf{Choice}_\alpha^m(h), \mathcal{M}, \langle m, h' \rangle \models \varphi$
$\mathcal{M}, \langle m, h \rangle \models K_\alpha \varphi$	iff	for each $\langle m', h' \rangle$ s.t. $\langle m, h \rangle \sim_\alpha \langle m', h' \rangle, \mathcal{M}, \langle m', h' \rangle \models \varphi$
$\mathcal{M}, \langle m, h \rangle \models \odot[\alpha] \varphi$	iff	for each $L \in \mathbf{Optimal}_\alpha^m, h' \in L$ implies that $\mathcal{M}, \langle m, h' \rangle \models \varphi$
$\mathcal{M}, \langle m, h \rangle \models \odot_s[\alpha] \varphi$	iff	for each $L \in \mathbf{S-optimal}_\alpha^m, \text{ for each } m' \text{ s.t. } m \sim_\alpha m', h' \in [L]_\alpha^{m'} \text{ implies that } \mathcal{M}, \langle m, h' \rangle \models \varphi.$

*Satisfiability, validity on a frame, and general validity are defined as usual. We write  $\|\varphi\|$  to refer to the set  $\{\langle m, h \rangle \in M \times H; \mathcal{M}, \langle m, h \rangle \models \varphi\}$ .*

### 3 Introducing Beliefs

The logic presented in the previous section offers many benefits for addressing complex interactive situations. The common thread among these situations is that agentic knowledge is taken into consideration when deciding whether an agent is responsible for having brought about some circumstance (see Horty's

coin-flip puzzles in [21] and [1]). However, we want to enhance the analysis by accounting for agents' belief-systems. As mentioned in the introduction, the beliefs that an agent has at a given index serve as justifications or explanations for a particular choice of action of said agent at said index.

In this work we adapt the arguments of [6] and formalize a notion of *full belief*. To clarify, an agent's full belief in the truth of a proposition means that the agent assigns probability 1 to the set of indices where the proposition is true. However, the typical logic of probabilistic full belief does not involve classical probability. The reason is that it is well known that classical measures yield problems for conditional beliefs and for belief revision. In classical-probability settings, conditioning on events with measure 0 is not defined and therefore "it is unclear how to proceed if an agent learns something to which she initially assigned probability 0" ([18], see also [17], [38], and [10]). Since we want to incorporate to stit theory a paradigm of belief that allows for revision, we follow the method of [6] and use *conditional probability* as primitive.<sup>2</sup> Therefore, we build conditional-probability spaces upon the branching-time structures from Definition 3. To simplify the terminology—and just as done in [6]—we focus on finite discrete structures, for which every subset is measurable with respect to special two-place probability functions mapping pairs of subsets to values in  $[0, 1]$ . These functions underlie the semantics for formulas of conditional belief  $B_\alpha^\Psi \varphi$ , which are meant to be read as 'after learning  $\Psi$ , agent  $\alpha$  believes that  $\varphi$  was the case (before the learning).'

**Definition 4 (Syntax with conditional-belief)** *The grammar for the formal language  $\mathcal{L}_{KOB}$  (an extension of  $\mathcal{L}_{KO}$ ) is given by:  $\varphi := p \mid \neg\varphi \mid \varphi \wedge \psi \mid \Box\varphi \mid [\alpha]\varphi \mid K_\alpha\varphi \mid B_\alpha^\Psi\varphi \mid \odot[\alpha]\varphi \mid \odot_{\mathcal{S}}[\alpha]\varphi$ .*

**Definition 5 (Epistemic act-utilitarian discrete-conditional-probability branching-time frames)** *A finite epistemic act-utilitarian discrete-conditional-probability branching-time frame (dcpbt-frame for short) is a tuple  $\langle M, \sqsubset, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha \in \text{Ags}}, \{\mu_\alpha\}_{\alpha \in \text{Ags}}, \mathbf{Value} \rangle$  such that  $M, \sqsubset, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha \in \text{Ags}}$ , and  $\mathbf{Value}$  are like in Definition 2, and for every  $\alpha \in \text{Ags}$ ,  $\mu_\alpha : \mathcal{P}(M \times H) \times \mathcal{P}(M \times H) \rightarrow [0, 1]$  is such that (a) for each  $B \subseteq M \times H$ ,  $\mu_\alpha^B := \mu_\alpha(\cdot|B)$  is either a constant function with value 1 or a classical-probability function on  $M \times H$ , and (b) for every  $A, B, C \subseteq M \times H$ ,  $\mu_\alpha(A \cap B|C) = \mu_\alpha(A|B \cap C) \cdot \mu_\alpha(B|C)$ .<sup>3</sup>*

A dcpbt-model  $\mathcal{M}$  results from adding a valuation function  $\mathcal{V}$  to a dcpbt-frame, and the semantics for the formulas of  $\mathcal{L}_{KOB}$  over such a model is defined recursively as in Definition 3, with the following additional clause:  $\mathcal{M}, \langle m, h \rangle \models B_\alpha^\Psi \varphi$  iff  $\mu_\alpha(\|\varphi\| \mid \|\Psi\| \cap \pi_\alpha[\langle m, h \rangle]) = 1$ .

**Remark 1** *In what follows, we will refer to any function  $q : \mathcal{P}(M \times H) \times \mathcal{P}(M \times H) \rightarrow [0, 1]$  that meets the requirements (a) and (b) of Definition 5 as a vf-function. Therefore, if  $p : M \times H \rightarrow [0, 1]$  is a classical-probability function, then the function  $p_c : (M \times H) \times (M \times H) \rightarrow [0, 1]$ , defined by the rules  $p_c(A|B) = \frac{p(A \cap B)}{p(B)}$  if  $p(B) > 0$  and  $p_c(A|B) = 1$  if  $p(B) = 0$ , is a vf-function (see [38] and [17, Chapter 3]). This means that probability theory's typical definition of conditional probability in terms of a classical-probability function is a special instance of a vf-function.*

<sup>2</sup>There have been various attempts to deal with the problem of conditioning on events of measure 0. The best known methods involve (1) conditional-probability spaces ([31], [32], [14], [38], [6], [17]), (2) nonstandard probability spaces ([33], [19], [26]), where events with infinitesimally small probability may still be learned or observed, and (3) lexicographic probability systems ([18]), which use sequences of probability measures with a descending order of importance.

<sup>3</sup>This means that  $\mu_\alpha$  is a two-place probability function that meets the following three axioms: (1)  $\mu_\alpha(A|A) = 1$  for every  $A \subseteq M \times H$ , (2) if  $A \cap B = \emptyset$ ,  $C \neq \emptyset$ , and  $\mu_\alpha^C$  is not constant 1, then  $\mu_\alpha(A \cup B|C) = \mu_\alpha(A|C) + \mu_\alpha(B|C)$ , and (3)  $\mu_\alpha(A \cap B|C) = \mu_\alpha(A|B \cap C) \cdot \mu_\alpha(B|C)$ . These three axioms are called the Popper-Rényi axioms for conditional-probability spaces, and they were first introduced in [32]. Observe that if  $\mu_\alpha(B|M \times H) = 0$ , then this condition does not prevent  $\mu_\alpha(\cdot|B)$  from being defined, and this is the quality of the theory of conditional-probability spaces that allows for conditioning on events with measure 0. We illustrate this property and its implications for conditional belief with the example included in Section 4.

Endowed with semantics for formulas involving conditional belief, we take plain (unconditional) *belief* to be represented by beliefs conditional on a tautology. Therefore, in what follows we write  $B_\alpha\varphi$  to denote  $B_\alpha^\top\varphi$ . With these formulas and the models they are evaluated on, we have extended the epistemic stit logic of Section 2 into a system that deals with both *knowledge* and *belief*. As pointed out in [5], [4], and [6], the truth conditions in Definition 5 yield a logic for which the knowledge operators validate the **S5** schemata, the conditional-belief operators validate the **K** schema, and the following interaction schemata are valid:  $K_\alpha\varphi \rightarrow B_\alpha^\psi\varphi$  (*Persistence of knowledge*);  $B_\alpha^\psi\varphi \rightarrow K_\alpha B_\alpha^\psi\varphi$  and  $\neg B_\alpha^\psi\varphi \rightarrow K_\alpha\neg B_\alpha^\psi\varphi$  (*Full introspection of belief*). Additionally, the following axioms that regard revision are also valid:  $B_\alpha^\varphi\varphi$  (*Hypotheses are accepted*);  $B_\alpha^\psi\varphi \rightarrow (B_\alpha^{\varphi\wedge\psi}\theta \leftrightarrow B_\alpha^\psi\theta)$  and  $\neg B_\alpha^\psi\neg\varphi \rightarrow (B_\alpha^{\varphi\wedge\psi}\theta \leftrightarrow B_\alpha^\psi(\varphi \rightarrow \theta))$  (*Minimality of revision*);  $\varphi \rightarrow \neg B_\alpha^\varphi\perp$  (*Weak consistency of belief*).<sup>4</sup>

Why account for belief revision? Well, in [18] Halpern argues that belief revision plays a critical role in the analysis of strategic reasoning in extensive-form games, and since the stit semantics for agency over branching-time structures can adopt most ideas of the theory of extensive-form games (see [23, Chapter 7], [15], and [2]), we believe that stit theory should also benefit from an account of conditional belief and of belief revision. There is a common idea that if a formalization of the temporal evolution of a game incorporates the assumption that players can change their beliefs about the game as it progresses due to some flow of information—for instance by drawing conclusions from opponent’s moves—then the analysis of interactive situations becomes much richer ([9]).<sup>5</sup>

In stit theory, branching-time structures represent an exhaustive set of possibilities for temporal evolution according to multi-agent interaction, and a treatment of belief change allows for a description of how agents could have constrained those possibilities if they had learned information regarding past choices about which they are uncertain. For instance, in the example that we mentioned in the introduction, the doctor did not know that the patient had eaten before being admitted to the hospital, and moreover the doctor *learned*—from the mistaken chart—that the patient had not eaten. In principle, if the doctor had learned that the patient had in fact eaten, then her beliefs should have been different. In other words, she would *revise* her beliefs after learning that the patient had eaten before being admitted to the hospital. By incorporating conditional-belief modalities into stit theory, we can analyze both syntactically and semantically the different ways in which doxastic states can change according to the learning of information. In this way, we can formalize further the justification of choices of action that

<sup>4</sup>Observe that the semantics for conditional belief implies that an agent’s belief of  $\varphi$  given  $\psi$  is relative to the agent’s epistemic state. In other words, we take it that conditional beliefs depend on the *information* available to the agent.

<sup>5</sup>One can easily associate a belief-revision Kripke-structure to an extensive-form game to evaluate formulas that reflect agents’ belief changes as the game “progresses” (see [9] for an illustration of such an association in the analysis of *backward induction* in extensive-form games). The typical way to do so is to think of full strategy profiles—functions that associate each node to a player’s move (or strategy) at that node—as possible worlds, so that the formulas evaluated on these worlds “[...] describe the way the game is actually played, and they provide a set of counterfactuals for evaluating the payoffs if the action taken at any node deviates from the specified action.” The evaluation of formulas at full strategy profiles is reminiscent of stit theory’s use of histories as part of the indices of evaluation, but in extensive-form games one also considers profiles that cannot be realized in the structure of the game: regardless of whether the nodes can be reached in a *play* or not, the functions that serve as possible worlds map each node to a move (or strategy) that can be played at that node. It is this feature of the belief-revision structure associated to an extensive-form game that makes conditional-belief modalities very useful. With the syntax of conditional-belief, we can represent the beliefs of an agent at a node that was actually reached in a play, by conditioning on formulas that ensure that such a node was reached. Although this feature of the language and the semantics of conditional belief could be very well exploited in the context of *strategic* stit theory ([23, Chapter 7]), here we propose that it is also relevant even in instantaneous-stit theory, namely due to agents’ uncertainty across the set of indices. If an agent had uncertainty about the index of evaluation it found itself at (because of some lack of information), then the conditional-belief modalities allow us to reason syntactically about the counterfactual situation in which the agent’s learning of such information would change its state of uncertainty.

have moral consequences.

## 4 Introducing Doxastic Oughts

In keeping with EGT, we argue that agents can be seen as having a doxastic sense of what they ought to do in interactive situations, and that this sense can be traced back both to their beliefs regarding the index at which they are and to the utilities of the histories in their available actions. Rather than delving into the analyses of rationality and rational choice in terms of best responses (see [36] and [29] for surveys of such analyses), we use the concepts of *utility* and *belief* in order to characterize a doxastic sense of ought-to-do in stit theory. Here, utility and belief underlie the process by which the beliefs of an agent explain whether that agent was justified in making a particular choice of action. Again, this extension of deontic stit is important to treat the kind of problems in excusability and responsibility attribution that stit theory deals with ([27]), as we will illustrate by analyzing the example presented in the introduction.

Our interpretation of belief rests upon probabilities assigned to the alternative histories of branching-time structures. Inspired by the customary treatment of decision rules under *uncertainty* (or *risk*) from decision theory,<sup>6</sup> we identify an agent’s doxastic sense of ought-to-do with the effects of actions that maximize *expected (deontic) utility*.

Expected utility theory has somewhat settled interpretations for the components of interactive decision making (see [34], [24], and [25]). Typically, the utilities of outcomes quantify agents’ *preferences*, and the probabilities assigned are seen either as *objective* measures for the frequency with which sets of outcomes—the *events*—ensue or as *subjective* representations of agentive belief (see Footnote 6). As mentioned before, in the context of deontic stit based on act utilitarianism we interpret the value of a history—**Value**(*h*)—as its deontic utility for the whole group of agents, with no specific interpretation for the word ‘utility.’ This means that we do not identify the deontic utility of a history with agentive *preference* but rather allow for the assignment of values to “accommodate a variety of different approaches” (see Section 2.2 [23, Chapter 3]). The notion of deontic utility itself is taken as primitive in act utilitarian stit, and it is a general notion that applies to the whole set of agents—not only to individual ones. Therefore, it may be thought of—but not necessarily so—as the “total utility of the set of agents in that history, their average utility, or perhaps some distribution-sensitive aggregation of the utilities of these individual agents” ([23], p. 38). As for the probabilities, it must be clear by now that we interpret them as a measure of the agents’ individual doxastic state.

**Definition 6** Let  $\mathcal{M}$  be a dcpbt-model. Let  $m \in M$ ,  $h \in H_m$ , and  $\alpha \in \text{Ags}$ . Let  $L \in \text{Choice}_\alpha^m$ . We define  $\alpha$ ’s expected deontic utility of  $L$  at  $\langle m, h \rangle$ —denoted by  $EU_\alpha^{(m,h)}(L)$ —as the value given by the following formula:  $EU_\alpha^{(m,h)}(L) := \sum_{m' \sim_\alpha m, h' \in [L]_\alpha^{m'}} \mu_\alpha(\{h'\} \mid \pi_\alpha[\langle m', h' \rangle]) \cdot \text{Value}(h')$ .

**Remark 2** This means that we calculate  $\alpha$ ’s expected deontic utility for one of its available actions  $L$  at  $m$  by summing the utilities of all the histories lying in the epistemic clusters of  $L$ , weighted by the probabilities that  $\alpha$  assigns to those histories, conditional on  $\alpha$ ’s information set at the index where  $\alpha$

<sup>6</sup>It is convenient to remark about subtle differences in definitions for overlapping concepts. Decision theorists typically distinguish between *choice under uncertainty*, for which decision makers do not know the outcome of a decision they engage in, and *choice under risk*, for which decision makers have probabilistic information regarding the outcomes. In decision theory, when agents have such probabilistic information for choices under risk, it means that they have *objective* information regarding the outcomes. In other words, the probabilistic information is not meant to embody an agent’s subjective beliefs regarding the outcomes. When *subjective* probabilities are introduced, like the ones we deal with here, the general view is to treat these cases as involving choice under uncertainty (see [30], [35]).

is. Observe that for every  $m \in M$  and  $L \in \mathbf{Choice}_\alpha^m$ , we have that  $EU_\alpha^{\langle m, h \rangle}(L) = EU_\alpha^{\langle m, h' \rangle}(L)$  for every  $h, h' \in H_m$ .

Our notion of expected deontic utility can be seen as a stit version of EGT's interpretation of an agent's expected utility for a given strategy, conditional on that agent's information. According to [29], in EGT an agent's expected utility for one of its strategies is calculated with respect to so-called *conjectures*. An agent's conjecture at a given world is a probability distribution on the set of all strategy profiles involving the *other* agents. Such a distribution is typically based on probabilities conditional either on the agent's strategy-choice at the world of evaluation or on the agent's information set at said world ([29]). Our framework differs from EGT's in three essential points: (1) we do not have individual utilities for each outcome; (2) in our structures, each *possible world* (which we refer to as an *index*) has a deontic utility, whereas in EGT only outcomes—or full strategy profiles—have utilities; and (3) while EGT regards information sets as subsets of the available strategies, we do not impose this condition—in fact, the semantic condition (OAC) that we adopt yields that information sets are unions of choice cells.

Since *dcpbt*-models are finite, we have that for every  $m \in M$  and  $h \in H_m$ , the set  $\{EU_\alpha^{\langle m, h \rangle}(L); L \in \mathbf{Choice}_\alpha^m\}$  has a maximum. Therefore, there are actions that maximize  $\alpha$ 's expected deontic utility at every index, namely the ones whose expected deontic utility is the same as said maximum. We denote by  $\mathbf{EU}_\alpha^{\langle m, h \rangle}$  the set of actions that maximize  $\alpha$ 's expected deontic utility at  $\langle m, h \rangle$ .<sup>7</sup>

**Definition 7 (Full syntax)** *The grammar for the formal language  $\mathcal{L}_{\text{KOBO}}$  (an extension of  $\mathcal{L}_{\text{KOB}}$ ) is given by:  $\varphi := p \mid \neg\varphi \mid \varphi \wedge \psi \mid \Box\varphi \mid [\alpha]\varphi \mid K_\alpha\varphi \mid B_\alpha^\Psi\varphi \mid \odot[\alpha]\varphi \mid \odot_{\mathcal{S}}[\alpha]\varphi \mid \odot_{\mathcal{B}}[\alpha]\varphi$ .*

We use the same *dcpbt*-models from Definition 5 to evaluate the formulas of  $\mathcal{L}_{\text{KOBO}}$ . The truth conditions are the same as before, with the following additional clause:  $\mathcal{M}, \langle m, h \rangle \models \odot_{\mathcal{B}}[\alpha]\varphi$  iff for each  $L \in \mathbf{EU}_\alpha^{\langle m, h \rangle}$  we have that  $[L]_\alpha^{m'} \subseteq \|\varphi\|$  for all  $m'$  such that  $m \sim_\alpha m'$ . In other words, at a given index it was the case that an agent doxastically ought to have seen to it that  $\varphi$  iff  $\varphi$  is an effect both of all the actions that maximized the agent's expected deontic utility at said index and of the epistemic equivalents of these actions.

## 4.1 Example

In order to illustrate the reach of the semantics introduced above, we present a formal analysis of the example in the introduction, using *dcpbt* models.

**Example 1** *Let  $\text{Ags} = \{\text{patient}, \text{doctor}\}$ . Let  $M$  and  $\Box$  be defined so as to be represented by the diagram in Figure 1. We have three moments ( $m_1 - m_3$ ) and four histories ( $h_1 - h_4$ ). These histories represent the different possibilities in which time may have evolved according to the actions available both to patient and doctor. The actions available to patient at moment  $m_1$  are  $E_1$ , which we interpret as the action of refusing to eat, and  $E_2$ , which we interpret as the action of eating. It is according to such actions that time progressed either into moment  $m_2$  or into moment  $m_3$ . At both these moments, doctor chose from her available actions and executed one of them. At moment  $m_2$ , the actions available to doctor are  $L_1$ , which we interpret as the action of supplying anesthetics, and  $L_2$ , which we interpret as the action of*

<sup>7</sup>Formally, our definition of expected deontic utility is in fact an instance of probability theory's *conditional expectation with respect to an event*—albeit in the setting where conditional probability is primitive. In this case, the so-called *event* is the information set of a given agent at the index of evaluation. Thus, if  $L$  is an available action for an agent at index  $\langle m, h \rangle$ , and if  $E$  denotes the expected value of the random variable **Value** with respect to  $\mu_\alpha(\cdot|H)$  (where we recall that  $H$  is the set of all histories), then  $EU_\alpha^{\langle m, h \rangle}(L) = E(\mathbf{Value}|\pi_\alpha[\langle m, h' \rangle])$  for any  $h' \in L$ .

refusing to supply anesthetics. Similarly, at moment  $m_3$  the actions available to doctor are  $L_3$ , which we interpret as the action of supplying anesthetics, and  $L_4$ , which we interpret as the action of refusing to supply anesthetics.

We model the utilities of the outcomes according to the statement of the example. Therefore, history  $h_1$ , where at  $m_1$  patient refused to eat and at  $m_2$  doctor supplied the anesthetics, gets the highest utility— $\text{Value}(h_1) = 1$ —due to the fact that such a history represents the situation in which the patient got ready for the surgery without any trouble. Histories  $h_2$  and  $h_4$  get a neutral utility of 0: both of them imply that doctor refused to supply anesthetics and thus the patient is not ready for the surgery. History  $h_3$ , on the other hand, gets a negative utility of  $-1$ , since it implies that at  $m_1$  patient ate and at  $m_3$  doctor supplied the anesthetics, leading to patient's death. As implied by the statement of the example, we take  $h_3$  to be the actual history.

The epistemic-doxastic states and ought-to-do's that we focus on are those of doctor, since they illustrate important semantic properties of our logic. We represent doctor's epistemic states with blue dashed lines, so that at both  $m_2$  and  $m_3$ , along every history running through them, doctor did not know whether the patient had eaten or not, but she knew which action she performed. The doxastic states of doctor are represented by the conditional-probability function  $\mu_{\text{doctor}}$ , given by the following rules. Let  $p_{\text{doctor}} : \mathcal{P}(M \times H) \rightarrow [0, 1]$  be a discrete classical-probability function such that  $p_{\text{doctor}}(\langle m_i, h_j \rangle) = \frac{9}{4}$  for  $i, j \in \{1, 2\}$ ,  $p_{\text{doctor}}(\langle m_1, h_i \rangle) = \frac{1}{4}$  for  $i \in \{3, 4\}$ , and  $p_{\text{doctor}}(\langle m_3, h_i \rangle) = \frac{1}{4}$  for  $i \in \{3, 4\}$ . We then define  $\mu_\alpha$  so that  $\mu_{\text{doctor}}(A|B) = \frac{p_{\text{doctor}}(A \cap B)}{p_{\text{doctor}}(B)}$ .

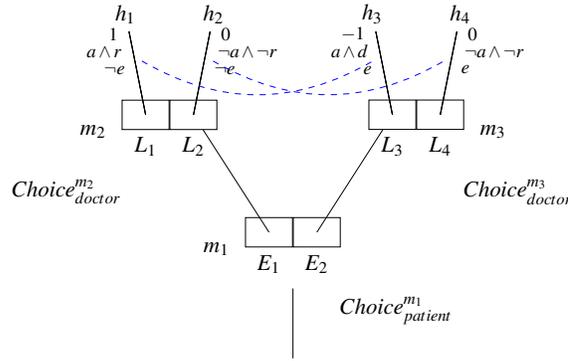


Figure 1: Example from *The Verdict*

Let  $e$  denote the atomic proposition ‘the patient has eaten,’ let  $a$  denote the atomic proposition ‘anesthetics are supplied to the patient,’ let  $r$  denote the atomic proposition ‘the patient is ready for surgery,’ and let  $d$  denote the atomic proposition ‘the patient will die.’ According to Definition 5, these atomic propositions and the formulas that are recursively built with them can be taken as true or false depending on the index of evaluation. For instance, we model the example so that at index  $\langle m_2, h_1 \rangle$  it was the case that *patient* ate, that *doctor* supplied the anesthetics, and that *patient* became ready for surgery. ( $\mathcal{M}, \langle m_2, h_1 \rangle \models e \wedge a \wedge r$ ).

As for the evaluation of formulas involving the basic stit-theory operators, observe that some instances of it are  $\mathcal{M}, \langle m_2, h_1 \rangle \models \Box e$  (at this index it was the case that it was settled that the patient did not eat),  $\mathcal{M}, \langle m_3, h_3 \rangle \models [\text{doctor}]d$  (at this index it was the case that the doctor saw to it that the patient died). As for formulas involving the epistemic-doxastic operators, we have that for  $i \in \{2, 3\}$  and  $j \in \{1, 2, 3, 4\}$ ,  $\mathcal{M}, \langle m_i, h_j \rangle \models \neg K_{\text{doctor}} e \wedge \neg K_{\text{doctor}} \neg e$  (at said indices it was the case that the doctor did

not know whether the patient had eaten or not) and  $\mathcal{M}, \langle m_i, h_j \rangle \models K_{\text{doctor}}[\text{doctor}]a \vee K_{\text{doctor}}[\text{doctor}]\neg a$  (at said indices it was the case that the doctor knew whether she was supplying the anesthetics or not). Moreover, we have that  $\mathcal{M}, \langle m_3, h_3 \rangle \models \neg B_{\text{doctor}}\neg e$  (at the actual index it was not the case that the doctor fully and unconditionally believed that the patient had not eaten), that  $\mathcal{M}, \langle m_3, h_3 \rangle \models B_{\text{doctor}}^e d$  (at the actual index it was the case that if the doctor had learned that the patient had in fact eaten, then she would have fully believed that the patient would die), and that  $\mathcal{M}, \langle m_3, h_3 \rangle \models B_{\text{doctor}}^e[\text{doctor}]d$  (at the actual index it was the case that if the doctor had learned that the patient had in fact eaten, then she would have fully believed that she would kill the patient).

As for formulas involving the deontic operators, we first observe that

- **Optimal** $_{\text{doctor}}^{m_2} = \{L_1\}$ , **Optimal** $_{\text{doctor}}^{m_3} = \{L_4\}$ .
- **S – Optimal** $_{\text{doctor}}^{m_2} = \{L_1, L_2\}$ , **S – Optimal** $_{\text{doctor}}^{m_3} = \{L_3, L_4\}$ .
- **EU** $_{\text{doctor}}^{\langle m_2, h_i \rangle} = \{L_1\}$  ( $i \in \{1, 2\}$ ), **EU** $_{\text{doctor}}^{\langle m_3, h_i \rangle} = \{L_3\}$  ( $i \in \{3, 4\}$ ).

Therefore, for instance we have that  $\mathcal{M}, \langle m_2, h_i \rangle \models \odot[\text{doctor}]a \wedge \neg K_{\text{doctor}} \odot[\text{doctor}]a$  ( $i \in \{1, 2\}$ ) (at moment  $m_2$ , along any history running through it, it was the case that the doctor objectively ought to have supplied the anesthetics, but the doctor did not know that for sure), that  $\mathcal{M}, \langle m_3, h_i \rangle \models \odot[\text{doctor}]\neg a \wedge \neg K_{\text{doctor}} \odot[\text{doctor}]\neg a$  ( $i \in \{3, 4\}$ ) (at moment  $m_3$ , along any history running through it, it was the case that the doctor objectively ought to have refrained from supplying the anesthetics, although the doctor did not know that), that  $\mathcal{M}, \langle m_i, h_j \rangle \models \neg \odot_{\mathcal{S}}[\text{doctor}]a \wedge \neg \odot_{\mathcal{S}}[\text{doctor}]\neg a$  ( $i \in \{1, 2\}$  and  $j \in \{1, 2, 3, 4\}$ ) (at no index it was the case that the doctor either subjectively ought to have supplied the anesthetics or subjectively ought to have refrained from supplying them), that  $\mathcal{M}, \langle m_i, h_j \rangle \models \odot_{\mathcal{B}}[\text{doctor}]a \wedge K_{\text{doctor}} \odot_{\mathcal{B}}[\text{doctor}]a$  ( $i \in \{2, 3\}$  and  $j \in \{1, 2, 3, 4\}$ ) (at every index it was the case that the doctor doxastically ought to have supplied the anesthetics and that she knew that). Observe that we could model the fact that patient's information chart was mistaken by the fact that  $\mathcal{M}, \langle m_3, h_3 \rangle \models B_{\text{doctor}}^e(\neg e \wedge \odot[\text{doctor}]a)$  (at the actual index it was the case that if the doctor had learned that the patient had not eaten (as she did by the mistake in the chart), then she would have fully believed that the patient had not eaten and that she objectively ought to have supplied the anesthetics.) Coupled with the fact that doctor did not know that patient had in fact eaten, the satisfaction of these last two formulas at the actual index should in principle provide a good reason for excusing doctor of actually having caused patient's death.

## 5 Axiomatization and Logic Properties

Since one of the main contributions of the present paper is the introduction of doxastic oughts, we review some of the properties of the semantics for formulas involving the operator  $\odot_{\mathcal{B}}[\alpha]$ . First of all, we must say that this modal operator yields a **KD45** logic. Secondly, the doxastic sense of ought validates a version of Kant's imperative *ought implies can*, as explained by the fact that the following formula is valid with respect to the class of *dcpbt* models:  $\odot_{\mathcal{B}}[\alpha]\varphi \rightarrow \Diamond K_{\alpha}\varphi$ . We also have that if an agent doxastically ought to have seen to it that  $\varphi$ , then the agent knows that this is settled—the formula  $\odot_{\mathcal{B}}[\alpha]\varphi \rightarrow K_{\alpha}\Box \odot_{\mathcal{B}}[\alpha]\varphi$  is valid as well. As for the interaction between this doxastic sense of ought and the objective/subjective ought-to-do's, we have that  $\not\models \odot_{\mathcal{B}}[\alpha]\varphi \rightarrow \odot[\alpha]\varphi$  and that  $\not\models \odot[\alpha]\varphi \rightarrow \odot_{\mathcal{B}}[\alpha]\varphi$  (as can be inferred from Example 1). Similarly, we have that  $\not\models \odot[\alpha]_{\mathcal{B}}\varphi \rightarrow \odot_{\mathcal{S}}[\alpha]\varphi$  and that  $\not\models \odot_{\mathcal{S}}[\alpha]\varphi \rightarrow \odot_{\mathcal{B}}[\alpha]\varphi$ . The first invalidity can be inferred from Example 1. The second one can be inferred from a variation of Example 1 as follows: let **Value**( $h_1$ ) = 2, **Value**( $h_2$ ) = 1, **Value**( $h_3$ ) = 0, and **Value**( $h_4$ ) = 0; let  $p_{\text{doctor}}$  be a discrete classical-probability function such that  $p_{\text{doctor}}(\langle m_2, h_1 \rangle) = \frac{1}{2}$ ,  $p_{\text{doctor}}(\langle m_2, h_2 \rangle) = \frac{9}{2}$ ,  $p_{\text{doctor}}(\langle m_3, h_3 \rangle) = \frac{9}{2}$ ,  $p_{\text{doctor}}(\langle m_3, h_4 \rangle) = \frac{1}{2}$ , and  $p_{\text{doctor}}$  has constant value 0 on

all other indices; let  $\mu_\alpha$  be defined as in Example 1; then it is the case that  $\mathcal{M}, \langle m_i, h_j \rangle \models \odot_{\mathcal{S}}[doctor]a \wedge \odot_{\mathcal{B}}[doctor]\neg a$  ( $i \in \{2, 3\}$  and  $j \in \{1, 2, 3, 4\}$ ).

In order to further the understanding of our logic, we present a sound, complete, and decidable proof system for it.

**Definition 8 (Proof system)** *Let  $\Lambda$  be the proof system defined by the following axioms and rules of inference:*

- (Axioms) *All classical tautologies from propositional logic. The S5 axiom schemata for  $\Box$ ,  $[\alpha]$ ,  $K_\alpha$ . The following axioms and schemata for the interactions of formulas with the given operators:*

$$\begin{aligned}
& \odot[\alpha](\varphi \rightarrow \psi) \rightarrow (\odot[\alpha]\varphi \rightarrow \odot[\alpha]\psi) & (A1) \\
& \Box\varphi \rightarrow [\alpha]\varphi \wedge \odot[\alpha]\varphi & (A2) \\
& \Box\odot[\alpha]\varphi \vee \Box\neg\odot[\alpha]\varphi & (A3) \\
& \odot[\alpha]\varphi \rightarrow \odot[\alpha](\Box\varphi) & (A4) \\
& \odot[\alpha]\varphi \rightarrow \Diamond[\alpha]\varphi & (Oic) \\
& \text{For } n \geq 1 \text{ and pairwise different } \alpha_1, \dots, \alpha_n, \\
& \bigwedge_{1 \leq k \leq n} \Diamond[\alpha_k]\varphi_i \rightarrow \Diamond(\bigwedge_{1 \leq k \leq n} [\alpha_k]\varphi_i) & (IA) \\
& K_\alpha\varphi \rightarrow [\alpha]\varphi & (OAC) \\
& \Diamond K_\alpha\varphi \rightarrow K_\alpha\Diamond\varphi & (Unif - H) \\
& \odot_{\mathcal{S}}[\alpha](\varphi \rightarrow \psi) \rightarrow (\odot_{\mathcal{S}}[\alpha]\varphi \rightarrow \odot_{\mathcal{S}}[\alpha]\psi) & (A5) \\
& \odot_{\mathcal{S}}[\alpha]\varphi \rightarrow \odot_{\mathcal{S}}[\alpha](K_\alpha\varphi) & (A6) \\
& K_\alpha\Box\varphi \rightarrow \odot_{\mathcal{S}}[\alpha]\varphi & (s.N) \\
& \odot_{\mathcal{S}}[\alpha]\varphi \rightarrow \Diamond K_\alpha\varphi & (s.Oic) \\
& \odot_{\mathcal{S}}[\alpha]\varphi \rightarrow K_\alpha\Box\odot_{\mathcal{S}}[\alpha]\varphi & (s.Cl1) \\
& \neg\odot_{\mathcal{S}}[\alpha]\varphi \rightarrow K_\alpha\Box\neg\odot_{\mathcal{S}}[\alpha]\varphi & (s.Cl2) \\
& B_\alpha(\varphi \rightarrow \theta) \rightarrow (B_\alpha\varphi \rightarrow B_\alpha\theta) & (A7) \\
& (\psi \leftrightarrow \varphi) \rightarrow (B_\alpha\theta \leftrightarrow B_\alpha^\varphi\theta) & (A8) \\
& K_\alpha\varphi \rightarrow B_\alpha\varphi & (PK) \\
& B_\alpha\varphi \rightarrow K_\alpha B_\alpha\varphi & (FIB1) \\
& \neg B_\alpha\varphi \rightarrow K_\alpha\neg B_\alpha\varphi & (FIB2) \\
& B_\alpha^\varphi\varphi & (HA) \\
& B_\alpha\varphi \rightarrow (B_\alpha^{\varphi \wedge \psi}\theta \leftrightarrow B_\alpha\theta) & (MBR1) \\
& \neg B_\alpha\neg\varphi \rightarrow (B_\alpha^{\varphi \wedge \psi}\theta \leftrightarrow B_\alpha(\varphi \rightarrow \theta)) & (MBR2) \\
& \psi \rightarrow \neg B_\alpha\perp & (WCon) \\
& \odot_{\mathcal{B}}[\alpha](\varphi \rightarrow \psi) \rightarrow (\odot_{\mathcal{B}}[\alpha]\varphi \rightarrow \odot_{\mathcal{B}}[\alpha]\psi) & (A9) \\
& \odot_{\mathcal{B}}[\alpha]\varphi \rightarrow \odot_{\mathcal{B}}[\alpha](K_\alpha\varphi) & (A10) \\
& K_\alpha\Box\varphi \rightarrow \odot_{\mathcal{B}}[\alpha]\varphi & (d.N) \\
& \odot_{\mathcal{B}}[\alpha]\varphi \rightarrow \Diamond K_\alpha\varphi & (d.Oic) \\
& \odot_{\mathcal{B}}[\alpha]\varphi \rightarrow K_\alpha\Box\odot_{\mathcal{B}}[\alpha]\varphi & (d.Cl1) \\
& \neg\odot_{\mathcal{B}}[\alpha]\varphi \rightarrow K_\alpha\Box\neg\odot_{\mathcal{B}}[\alpha]\varphi & (d.Cl2)
\end{aligned}$$

- (Rules of inference) *Modus Ponens, Substitution, and Necessitation for all modal operators.*

In the Appendix of this work, we discuss all these axioms and schemata, and we show that the proof system  $\Lambda$  is sound and complete with respect to a class of models that are more general than the ones introduced in Definition 5. These models differ from *dcpbt*-models in two main qualities: (1) following [1], they are multi-valued to the extent that instead of only one deontic value function, they have three: one for the objective ought-to-do's, one for the subjective ones, and one for the doxastic ones; and (2) they are Kripke-structures based on domains of possible worlds.<sup>8</sup> The proof of completeness in the Appendix also shows that  $\Lambda$  is *decidable*. This is a consequence of the logic's finite model property, which is shown through obtaining a finite canonical model using arguments typical of modal filtrations.

<sup>8</sup>Nevertheless, one can adapt the correspondence between Kripke models and branching-time models from [40] and [1] to show that  $\Lambda$  is sound and complete with respect to the class of *multi-valued dcpbt* models.

All these results become relevant in a specific line of research where proof systems of deontic logic are intended to help in the testing of ethical behavior of AI through theorem-proving and model-checking (see [3], [28], [11]).

Unfortunately, the Appendix is too long to include it here. Therefore, the reader can only find it at [https://www.researchgate.net/publication/351656805\\_Appendix](https://www.researchgate.net/publication/351656805_Appendix).

## 6 Conclusion

“The performance is sometimes masterful, extremely clever, but the control of the actions, their source, is deranged and depends on various morbid impressions,” says the character Zossimov, in Dostoevsky’s *Crime and Punishment*. From the discussions that appear in this paper, it is clear that what agents know and what they believe at the moment of acting—as well as the obligations that arise according to these knowledge and beliefs—can be interpreted as “sources” of their actions, as some of those “impressions” on which agency depends that Zossimov speaks about.

The main novelty of the present, logic-based, treatment of these “sources” of agency lies in the incorporation of beliefs into deontic stit theory. The relation between (a) a given agent’s doxastic state, (b) the actions that are available to said agent at some point of time, and (c) the deontic utility of such actions, gives us the opportunity to reason about a sense of agentive obligation that is based on the idea of maximizing expected deontic utility. Thus, we end up with a reasonable measure for explaining why agents could have favored certain actions over others, something that is useful in formal analyses of responsibility attribution, for instance.

A prominent feature of our analysis is the use of conditional beliefs. We mentioned that, since stit theory’s account of interactive scenarios would greatly benefit from adopting viewpoints typical of EGT and of epistemic logic, we wanted to introduce a notion of belief that would satisfactorily open up possibilities for belief revision. It must be said, then, that the example discussed in Section 4 does not make heavy use of the theory of belief revision underlying the probabilistic semantics of belief that we introduced. Although this was a choice made more for the sake of simplicity than anything else, it is true that the logic presented here is rather a ‘first step’ toward an appropriate theory of belief-based action and obligation—a theory that would admit revision in *both* the categories of beliefs and obligations. A very interesting problem for future work along these lines, then, regards implementing the ideas of belief revision—in terms of conditional belief—to formalize conditional doxastic oughts. The basic intuition is that, if at some index an agent has learned that  $\psi$  is the case, then the doxastic obligations that such an agent had at the index should in principle be subject to the revision with  $\psi$ —just as beliefs are. Formulas of the form  $\odot_{\mathcal{B}}[\alpha]^{\psi}\phi$  could then capture these revised doxastic oughts, such that possible semantics for these formulas could depend on the restriction of the model’s domain to indices where  $\psi$  holds—just as happens for the version of conditional belief discussed in this paper. In fact, one can find good pointers in this respect in [23, Chapter 4], since a stit-theoretic account of conditional ought-to-do’s is presented there. An adequate axiomatization of such possible conditional doxastic oughts, however, is still a complicated open problem.

In conclusion, this work deals with important questions in the modeling of agency, knowledge, belief, and obligation. We presented logic-based characterizations of these concepts, that allowed us to devise unequivocal representations of interactive scenarios, where agents within an environment choose courses of action through time and where those choices could be traced back both to the epistemic-doxastic states of the agents and to different senses of obligation. The logic developed here lays the groundwork for interesting future research, and there is still plenty of work to do.

## References

- [1] Aldo Iván Ramírez Abarca & Jan Broersen (2019): *A Logic of Objective and Subjective Oughts*. In: *European Conference on Logics in Artificial Intelligence*, Springer, pp. 629–641, doi:10.1007/978-3-030-19570-0\_41.
- [2] Aldo Iván Ramírez Abarca & Jan Broersen (2019): *Stit Semantics for Epistemic Notions Based on Information Disclosure in Interactive Settings*. In: *International Workshop on Dynamic Logic*, Springer, pp. 171–189, doi:10.1007/978-3-030-38808-9\_11.
- [3] Konstantine Arkoudas, Selmer Bringsjord & Paul Bello (2005): *Toward ethical robots via mechanized deontic logic*. In: *AAAI Fall Symposium on Machine Ethics*, pp. 17–23.
- [4] Alexandru Baltag & Sonja Smets (2006): *Conditional doxastic models: A qualitative approach to dynamic belief revision*. *Electronic notes in theoretical computer science* 165, pp. 5–21, doi:10.1016/j.entcs.2006.05.034.
- [5] Alexandru Baltag & Sonja Smets (2006): *The logic of conditional doxastic actions: a theory of dynamic multi-agent belief revision*. In: *Proceedings of ESSLLI Workshop on Rationality and Knowledge*, pp. 13–30.
- [6] Alexandru Baltag & Sonja Smets (2008): *Probabilistic dynamic belief revision*. *Synthese* 165(2), p. 179, doi:10.1007/s11229-008-9369-8.
- [7] N. Belnap, M. Perloff & M. Xu (2001): *Facing the future: agents and choices in our indeterminist world*. Oxford University Press.
- [8] Adam Bjorndahl, Joseph Y Halpern & Rafael Pass (2017): *Reasoning about rationality*. *Games and Economic Behavior* 104, pp. 146–164, doi:10.1016/j.geb.2017.03.006.
- [9] Oliver Board (2004): *Dynamic interactive epistemology*. *Games and Economic Behavior* 49(1), pp. 49–80, doi:10.1016/j.geb.2003.10.006.
- [10] Craig Boutilier et al. (1995): *On the revision of probabilistic belief states*. *Notre Dame Journal of Formal Logic* 36(1), pp. 158–183, doi:10.1305/ndjfl/1040308833.
- [11] Selmer Bringsjord, Konstantine Arkoudas & Paul Bello (2006): *Toward a general logicist methodology for engineering ethically correct robots*. *IEEE Intelligent Systems* 21(4), pp. 38–44, doi:10.1109/MIS.2006.82.
- [12] Jan Broersen (2013): *Probabilistic stit logic and its decomposition*. *International journal of approximate reasoning* 54(4), pp. 467–477, doi:10.1016/j.ijar.2012.08.007.
- [13] Jan Broersen & Aldo Iván Ramírez Abarca (2018): *Formalising Oughts and Practical Knowledge without Resorting to Action Types*. In: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, International Foundation for Autonomous Agents and Multiagent Systems, pp. 1877–1879.
- [14] Bruno De Finetti (1936): *Les probabilités nulles*. Gauthier-Villars.
- [15] Hein Duijf & Jan Broersen (2016): *Representing strategies*. *arXiv preprint arXiv:1607.03355*, doi:10.4204/EPTCS.218.2.
- [16] HWA Duijf (2018): *Let's do it!: Collective responsibility, joint action, and participation*. Ph.D. thesis, Utrecht University.
- [17] Konstantinos Gkikas (2015): *Stable Beliefs and Conditional Probability Spaces*. Ph.D. thesis, Universiteit van Amsterdam.
- [18] Joseph Y Halpern (2010): *Lexicographic probability, conditional probability, and nonstandard probability*. *Games and Economic Behavior* 68(1), pp. 155–179, doi:10.1016/j.geb.2009.03.013.
- [19] Peter J Hammond (1994): *Elementary non-Archimedean representations of probability for decision theory and games*. In: *Patrick Suppes: scientific philosopher*, Springer, pp. 25–61, doi:10.1007/978-94-011-0774-7\_2.
- [20] John C Harsanyi (1967): *Games with incomplete information played by “Bayesian” players, I–III Part I. The basic model*. *Management science* 14(3), pp. 159–182, doi:10.1287/mnsc.14.3.159.

- [21] John Horty (2019): *Epistemic Oughts in Stit Semantics*. *Ergo, an Open Access Journal of Philosophy* 6, doi:10.3998/ergo.12405314.0006.004.
- [22] John Horty & Eric Pacuit (2017): *Action types in stit semantics*. *The Review of Symbolic Logic* 10(4), pp. 617–637, doi:10.1017/S1755020317000016.
- [23] John F. Horty (2001): *Agency and Deontic Logic*. Oxford University Press, doi:10.1093/0195134613.001.0001.
- [24] Richard C Jeffrey (1965): *Ethics and the Logic of Decision*. *The Journal of Philosophy* 62(19), pp. 528–539, doi:10.2307/2023748.
- [25] Edi Karni (2014): *Axiomatic foundations of expected utility and subjective probability*. In: *Handbook of the Economics of Risk and Uncertainty*, 1, Elsevier, pp. 1–39, doi:10.1016/B978-0-444-53685-3.00001-5.
- [26] Daniel Lehmann & Menachem Magidor (1992): *What does a conditional knowledge base entail?* *Artificial intelligence* 55(1), pp. 1–60, doi:10.1016/0004-3702(92)90041-U.
- [27] Emiliano Lorini, Dominique Longin & Eunata Mayor (2014): *A logical analysis of responsibility attribution: emotions, individuals and collectives*. *Journal of Logic and Computation* 24(6), pp. 1313–1339, doi:10.1093/logcom/ext072.
- [28] Yuko Murakami (2004): *Utilitarian deontic logic*. *AiML-2004: Advances in Modal Logic* 287.
- [29] Eric Pacuit & Olivier Roy (2017): *Epistemic Foundations of Game Theory*. In Edward N. Zalta, editor: *The Stanford Encyclopedia of Philosophy*, Summer 2017 edition, Metaphysics Research Lab, Stanford University.
- [30] Martin Peterson (2017): *An introduction to decision theory*. Cambridge University Press, doi:10.1017/9781316585061.
- [31] Karl Raimund Popper (1968): *The Logic of Scientific Discovery*. (Revised Edition.). Hutchinson.
- [32] Alfréd Rényi (1955): *On a new axiomatic theory of probability*. *Acta Mathematica Academiae Scientiarum Hungarica* 6(3-4), pp. 285–335, doi:10.1007/BF02024393.
- [33] Abraham Robinson (1973): *Function theory on some nonarchimedean fields*. *The American Mathematical Monthly* 80(6), pp. 87–109, doi:10.2307/3038223.
- [34] L.J. Savage (1954): *The Foundations of Statistics*. John Wiley and Sons, New York.
- [35] Keiran Sharpe (2018): *On risk and uncertainty, and objective versus subjective probability*. *Economic Record* 94, pp. 49–72, doi:10.1111/1475-4932.12403.
- [36] Katie Steele & H. Orri Stefánsson (2016): *Decision Theory*. In Edward N. Zalta, editor: *The Stanford Encyclopedia of Philosophy*, winter 2016 edition, Metaphysics Research Lab, Stanford University.
- [37] Allard Tamminga (2013): *Deontic logic for strategic games*. *Erkenntnis* 78(1), pp. 183–200, doi:10.1007/s10670-011-9349-0.
- [38] Bas C Van Fraassen (1995): *Fine-grained opinion, probability, and the logic of full belief*. *Journal of Philosophical Logic* 24(4), pp. 349–377, doi:10.1007/BF01048352.
- [39] Heinrich Wansing (2006): *Doxastic decisions, epistemic justification, and the logic of agency*. *Philosophical Studies* 128(1), pp. 201–227, doi:10.1007/s11098-005-4063-x.
- [40] Ming Xu (1994): *Decidability of deliberative stit theories with multiple agents*. In: *International Conference on Temporal Logic*, Springer, pp. 332–348, doi:10.1007/BFb0013997.
- [41] Ming Xu (2015): *Combinations of Stit with Ought and Know*. *Journal of Philosophical Logic* 44(6), pp. 851–877, doi:10.1007/s10992-015-9365-7.

# Epistemic Modality and Coordination under Uncertainty\*

Giorgio Sbardolini

ILLC, University of Amsterdam  
Amsterdam, The Netherlands

`g.sbardolini@uva.nl`

Communication facilitates coordination, but coordination might fail if there's too much uncertainty. I discuss a scenario in which vagueness-driven uncertainty undermines the possibility of publicly sharing a belief. I then show that asserting an epistemic modal sentence, 'Might  $\phi$ ', can reveal the speaker's uncertainty, and that this may improve the chances of coordination despite the lack of a common epistemic ground. This provides a game-theoretic rationale for epistemic modality. The account draws on a standard relational semantics for epistemic modality, Stalnaker's theory of assertion as informative update, and a Bayesian framework for reasoning under uncertainty.

Shiv and Logan want to spend time together over the coming weekend. They prefer to go to the beach if it will be sunny and to a café if it will be raining, but they will only go to either place if the other goes. Their predicament is the familiar one of a coordination problem [19]. In a variant known as the signalling game, Shiv and Logan coordinate by sending a signal, i.e. an utterance that reveals information that is initially available to only one of the players [23, 8, 27, 24].

It's relatively well-understood how Shiv and Logan coordinate if the relevant information (it will be sunny, it will be raining) is public. Roughly,  $q$  is public information within a group just in case all members of the group believe that  $q$ , all believe that all believe that  $q$ , all believe that all believe that all believe that  $q$ , and so on. Sometimes, however, a belief may fail to be public. For example: Shiv thinks that it will be raining, but she's not very confident, and indeed she expects that, reasonably, Logan thinks that it will be sunny. In this case Shiv does not believe that she and Logan share the belief that it will be raining. The belief fails to be public. Is there still a way for Shiv and Logan to coordinate for the weekend, despite the uncertainty?

In this paper, I provide a rational reconstruction of the use of epistemic modals in a signalling game. I will present a game-theoretic rationale for epistemic possibility talk: revealing one's uncertainty to the interlocutors can improve one's expected utility despite lack of public information. I will employ a general Bayesian framework for reasoning under uncertainty [12, 11, 14, 18]. The model will show that, in conditions of uncertainty specified below, rational agents can improve their chances of coordination by means of sentences that make an epistemic hedge.

My focus is on sentences of the form 'Might  $\phi$ ', where an epistemic possibility operator takes wide scope. Epistemic possibility modal auxiliaries and adverbs, and expressions of close kin (*might, perhaps, maybe, possibly, probably*) are natural resources to employ in case coordination is challenged by the lack of an epistemic common ground. If Shiv uttered (1) in the story above, for example, she would be naturally understood as suggesting to go to a café for the weekend—almost as if she simply asserted that it will be raining, while at the same time hedging that assertion.

## 1. It might be raining.

---

\*This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No 758540) within the project *EXPRESS: From the Expression of Disagreement to New Foundations for Expressivist Semantics*. Special thanks to Luca Incurvati, Leïla Bussière, and to the participants of the 2020 Dutch Research School in Philosophy conference for discussion.

- (a) Let's stay in.
- (b) Let's go out.

In her context, it is very natural for Shiv to continue (1) as in (1a), much less natural (and not even so coherent) to continue as in (1b). Therefore, even if the speaker is not in a position to outright assert that it will rain, the interlocutors might still coordinate on going to a café by means of something less committal than that assertion.

## 1 Failures of Coordination

A game is a set of players  $I$ , and sets of actions  $x_i$  and utility functions  $u_i$  for each player  $i$  in  $I$ . In a coordination game between two players, each having two actions,  $a$  and  $b$ , the players' utility functions are summarized in Table 1. Shiv (the Column player  $S$ ) prefers to stay in,  $a$ , if Logan (the Row player  $L$ ) stays in, and prefers to go out,  $b$ , if Logan goes out. Same for Logan.

		S	
		a	b
L	a	1, 1	0, 0
	b	0, 0	1, 1

Table 1: Coordination Game

This particular assignment of preferences makes the agents *indifferent* as to whether they stay in or go out so long as each does what the other does. How do they coordinate? Either could commit to an action, and then inform the other about their commitment. But if they are indeed indifferent, they might as well toss a (fair) coin, and resolve to stay in if and only if it lands heads. Suppose furthermore that only one of them,  $S$ , can see the coin: she will then report to  $L$  the outcome of the coin toss by sending a signal ('It's tails!'). Coordination is then easily achieved [23, 27].

Shiv and Logan can coordinate by "pivoting" on the weekend being rainy, rather than the coin landing tails, as well as on any other proposition. Let  $q$  and  $\bar{q}$  be two mutually exclusive propositions. The agents mutually know that they prefer  $a$  if  $q$  and  $b$  if  $\bar{q}$ . If Shiv believes that  $q$ , she may signal so, and thereby share her belief with Logan. Neither Shiv nor Logan has reason to deceive the other, since either receives a positive payoff just in case the other does (as shown in Table 1). Therefore, once a belief is shared by signalling, it typically becomes public: both believe it, both believe that both believe it, and so on. In a coordination game, if  $q$  is public between Shiv and Logan, the rational (i.e. utility maximizing) choice for both is  $a$ .<sup>1</sup>

The crucial idea is that signalling turns a belief into public belief. Sometimes, however, beliefs fail to be public. This may be for a number of reasons. In more accidental cases, people are distracted, uncurious, or unintelligent. The impasse here is "solved" by a sleight of idealization. Let's assume that Shiv and Logan are Bayesian agents who do not suffer from such accidental shortcomings of rationality. Their credences are mathematically coherent, and they update by Bayes' rule. Still, there are complex cases of failure of public belief known to the literature.

---

<sup>1</sup>This is version of the traditional signalling game described by David Lewis [19], with a couple of qualifications. (i) Lewis talked about common knowledge, rather than common belief, but the stronger condition doesn't add much at this stage: people can coordinate on something false, so long as enough people believe it. (ii) Lewis worked with the notion of a Nash equilibrium, but the idea of solving the game by reference to an event with an independent prior probability (the coin lands heads, the weekend will be rainy) leads in fact to a generalization known as correlated equilibrium, with which I work in the current paper [1, 29]. Games of this kind have played an important role in our understanding of language [24].

In the following scenario, Shiv and Logan are planning for dinner. However, they are try to coordinate on something vague. The model of vagueness below is discussed in [4], and inspired by the well-known case of two generals' failing to coordinate an attack [7, 21].

**Vagueness.** Shiv and Logan just moved to a new town. They have been told about a great restaurant. If the restaurant is close, they would like to go there for dinner, but since they are quite tired after a day of moving and unpacking, they prefer to eat in if the restaurant is far. They can either go out or stay in, but if either goes while the other doesn't, both will eat alone and be miserable.

In Vagueness, Shiv and Logan are facing a Lewisian coordination problem. Sometimes, however, a restaurant is neither definitely close, nor definitely far. There is no sharp boundary between close and far, and, in the "borderline area", each may think that the restaurant is close while the other thinks that it's far. They will coordinate only if a particular belief is public, but in borderline cases, neither believes that they believe the same thing. Vagueness undermines the possibility of sharing a belief in public.

It would be natural for Shiv and Logan to be uncertain about what to do, in their situation. Vagueness has often been linked to uncertainty [6, 20]. A common way to describe uncertainty is in terms of degrees of confidence. Let's say that an agent  $i$  thinks that  $q$  just in case  $i$  has some positive degree of confidence that  $q$  is the case. Then  $i$  thinks that  $q$  just in case  $p(q_i) > 1 - p(q_i)$ , i.e.  $i$  expects that  $q$  is more likely than not. In the borderline area, Shiv may think that the restaurant is close, although her confidence remains low: indeed, below a relevant threshold. Above the threshold, Shiv thinks that the restaurant is definitely close, or, as I shall say, she *believes* that it is close.

What is a confidence threshold? In ordinary life, many factors contribute to an agent's confidence level. In the context of the game, a qualitative characterization helps: an agent believes that  $q$  just in case she thinks that  $q$ , and thinks that others think that  $q$  as well. That is, one believes that  $q$  just in case one is confident enough that  $q$  is the case to think that others think that  $q$  as well. And so someone who thinks that the restaurant is close is uncertain so long as she has a reasonable expectation that others are not of the same opinion.<sup>2</sup>

By assuming that confidence thresholds and shared attitudes line up, the failures of coordination in Vagueness are failures to share a belief about what others think. In other words, coordination fails because a belief isn't public. To make this point precise, let there be at least three discrete states  $w_1, w_2, w_3$  in the agents' environment. In  $w_1$  the restaurant is close ( $q$ ), in  $w_3$  it is far ( $\bar{q}$ ), and in  $w_2$  it is neither close nor far. As far as the agents know, prior to the interaction, any of these worlds might be theirs.

Crucially, the agents' doxastic states are not aligned in the borderline area. For concreteness, let's assume that Shiv does not distinguish  $w_1$  and  $w_2$ , in which the restaurant looks close to her, and Logan does not distinguish  $w_2$  and  $w_3$ , in which the restaurant looks far to him. (The converse possibility is analogous, and omitted.) Thus, the agents partition the logical space differently. A partition is a set of jointly exhaustive and mutually exclusive subsets of the logical space  $W$ . Each agent  $i$  has their own partition  $\Pi_i$ , which represents how they distinguish possibilities. Let  $\pi_i(w) \in \Pi_i$  be the cell of  $\Pi_i$  to

<sup>2</sup>Of course, this is not an analysis of *think* and *believe*, but a stipulation for describing a probability distribution over propositions. The stipulation plausibly fits with at least some informal uses of *think* and *believe*. The terminology, however, may sound misleading. Does belief imply certainty? Not if *certainty* means 'lack of Cartesian doubt', but it might if it means 'having high-ish confidence about what others think'. I mean the latter. Likewise, there is a sense in which one could be confident that something is the case while believing that others disagree: but this just shows that there are other characterizations of what confidence thresholds are, besides what I offered. That's fine. It's worth keeping in mind that Shiv and Logan are supposedly rational epistemic peers: if they think that  $q$  expecting that someone else does not think that  $q$ , who is equally rational and in the same epistemic position, then they should be less confident about their judgement. I assume that they are. That's the relevant sense of *confidence*.

which  $w$  belongs. An agent  $i$  ‘fails to distinguish’  $w$  from any  $w'$  that belongs to  $\pi_i(w)$ . I will say that  $i$  thinks that  $q$  at  $w$  just in case  $\pi_i(w) \subseteq q$ . Figure 1 represents the agents’ doxastic states with respect to the three possibilities that are salient to them at the beginning of the interaction. Furthermore, I assume that Figure 1 is how the agents understand their own beliefs as well as those of the others.

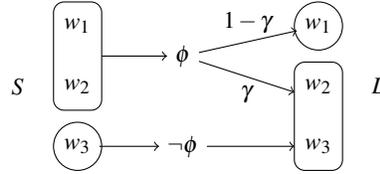


Figure 1: Signalling game under uncertainty

Partitions in Figure 1 are generated as follows. Let’s consider a Sorites series  $S$  of states  $t_1, \dots, t_n$ . The restaurant is definitely close in  $t_1$  (it is downstairs), it is definitely far in  $t_n$  (it is two time zones away), and the remaining states form a linearly ordered progression from near to far. Shiv and Logan go through a ‘forced march’ [16]. For all states in  $S$ , they judge whether the restaurant is close,  $q$ , or far,  $\bar{q}$ . The judgment has to be made even if the agents hesitate. Both Shiv and Logan think that  $q$  in  $t_1$  since the restaurant is definitely close. For some  $x$  between 1 and  $n$ , Shiv will presumably flip and think that  $\bar{q}$  in  $t_x$ . For some  $y$  between 1 and  $n$ , Logan will flip too. There comes a point during the forced march at which they both judge ‘It’s far’, not necessarily the same point. Thus, both Shiv and Logan think that  $q$  in all  $t < \min(t_x, t_y)$ , and both think that  $\bar{q}$  in all  $t > \max(t_x, t_y)$ . Therefore, we may pool  $S$  into three ‘uber’ states  $w_1, w_2, w_3$  without loss of generality, and continue working with uber states (or worlds).

$$\begin{aligned} w_1 &= \{t \in S : t < \min(t_x, t_y)\} \\ w_2 &= \{t \in S : \min(t_x, t_y) \leq t \leq \max(t_x, t_y)\} \\ w_3 &= \{t \in S : \max(t_x, t_y) < t\} \end{aligned}$$

The picture could be complicated by adding more players, each drawing the line between  $q$  and  $\bar{q}$  at a different point. It would still be possible to pool all states in  $S$  into those in which  $q$  is true for every player, those in which  $\bar{q}$  is true for every player, and the rest. Thus, there is a supervaluational description of the Vagueness game environment, in which players take the place of ‘precisifications’ [4].

Finally, let’s suppose that Shiv and Logan have two signals,  $\phi$  and  $\neg\phi$ : ‘The restaurant is close’ and ‘The restaurant is far’. For the semantics, let  $\phi$  be true at  $w_1$  and false at  $w_3$ , so  $\neg\phi$  is false at  $w_1$  and true at  $w_3$ . I prefer to remain neutral on further details of the semantics of vague terms, such as ‘close’ and ‘far’. In order to be more specific, one could say that  $\phi$  and  $\neg\phi$  are truth-valueless at  $w_2$ . Alternatively, one could keep classical logic, following [30]. My discussion does not depend on the logic of vague terms.

Sending signals reveals to the signal receiver what the signal sender thinks, and this is how coordination is ordinarily reached, if it is. For if they both think the same then, through signalling, they both come to believe that they both think the same. But uncertainty can undermine coordination. Suppose that the agents know that their interaction is as depicted in Figure 1. Suppose first that Shiv thinks that the restaurant is far. Then she sends  $\neg\phi$ . Upon receiving  $\neg\phi$ , Logan thinks that Shiv thinks that the restaurant is far, for Shiv has no reason to deceive Logan. Therefore, a signal  $\neg\phi$  is how Logan can distinguish a possibility in which Logan thinks that the restaurant is far and Shiv does too, from one in which he thinks that the restaurant is far but Shiv doesn’t. But Logan thinks that the restaurant is far in any circumstance in which Shiv does, hence if  $\neg\phi$  is sent, Shiv and Logan believe that it’s far. Moreover,

since they can reason to this point, they believe that they believe that it's far, and so on. Therefore, if  $\neg\phi$  is sent, the belief that the restaurant is far becomes public, and they coordinate on staying in.

If instead Shiv thinks that the restaurant is close, she sends  $\phi$ , but public belief might fail. For Shiv fails to distinguish a possibility in which both think that the restaurant is close from another possibility,  $w_2$ , in which she thinks that the restaurant is close but Logan doesn't. These cases can be interpreted as those in which Shiv wrongly guesses what Logan thinks. For Shiv's signal  $\phi$  ('The restaurant is close!') would come across to Logan as an error of judgement: Logan would think that the restaurant is far, and realize by Shiv's signal that she thinks that it's close. Since Shiv is aware of this, if she thinks that the restaurant is close, she may not think that Logan thinks it is.

Coordination equilibria still exist, although within the limits given by the agents' uncertainty. Let  $\gamma$  be the agents' shared prior concerning the chance that they give different judgments as to whether the restaurant is close or far, and let  $\delta$  be a real number between 0 and 1. The relevant possibilities can then be evaluated as follows.

$$p(w_1) = \delta(1 - \gamma) \quad p(w_2) = \gamma \quad p(w_3) = (1 - \delta)(1 - \gamma)$$

That is, with chance  $\gamma$  the agents guess that they don't think what the other thinks, i.e., they are in  $w_2$ , and they assign complementary probabilities to the rest of the cases by splitting them over  $\delta$  and  $1 - \delta$ .<sup>3</sup> Under the plausible assumption that an agent  $i$  chooses  $a$  only if  $i$  thinks  $q$ , we can calculate the expected utility of  $a$  for  $i$ . Let  $u_i$  be  $i$ 's utility function, and  $j$  be  $i$ 's opponent.

$$eu_i(a) = p(q_i \& q_j) \cdot u_i(a, a) + p(q_i \& \bar{q}_j) \cdot u_i(a, b)$$

By Table 1,  $u_i(a, b) = 0$  for both agents, hence by Bayes' rule,

$$eu_i(a) = p(q_i) \cdot p(q_j|q_i) \cdot u_i(a, a)$$

Consider  $S$  first. The probability  $p(q_S)$  that  $S$  thinks that  $q$  is  $\delta(1 - \gamma) + \gamma$ , and the conditional probability  $p(q_L|q_S)$  that  $L$  thinks that  $q$  while  $S$  thinks that  $q$  is just the proportion of cases in which  $L$  thinks  $q$  out of those in which  $L$  does:  $\frac{\delta(1-\gamma)}{\delta(1-\gamma)+\gamma}$ . Hence  $eu_S(a) = \delta(1 - \gamma)$ .

Consider  $L$ . The probability that  $L$  thinks that  $q$  is  $p(w_1) = \delta(1 - \gamma)$ , and the probability that  $S$  thinks that  $q$  given that  $L$  thinks that  $q$  is just 1, for all cases in which  $S$  thinks that  $q$  are cases in which  $L$  does too. It follows that  $eu_L(a) = \delta(1 - \gamma) = eu_S(a)$ . Parallel reasoning shows that  $eu_S(b) = eu_L(b) = (1 - \delta)(1 - \gamma)$ . Consequently, the coordination equilibrium  $(a, a)$  obtains only if  $eu_S(a) > eu_S(b)$  and  $eu_L(a) > eu_L(b)$ , hence only if  $\delta > 1 - \delta$ . Similarly, the  $(b, b)$  equilibrium obtains only if  $1 - \delta > \delta$ .

I have assumed in the preceding paragraph that  $\gamma \neq 1$ . This seems reasonable, since  $\gamma$  represents the chance that an agent doesn't think as the other does. In other words, so long as doxastic misalignment isn't inevitable, coordination equilibria exist under the conditions just derived. While reasonable, this conclusion is not very strong. For even if the agents are rational, and know by the proof above that coordination equilibria exist, it doesn't follow that they will coordinate. The uncertainty may still be too impressive for them to take action.

An upper bound on  $\gamma$  would help. Earlier I assumed that a necessary condition for  $i$  to choose  $a$  is that  $i$  thinks that  $q$ . It seems plausible to say that a sufficient condition for  $i$  to choose  $a$  is that both think that

<sup>3</sup>Probabilities are only assigned to sets of possible worlds, in order to uniformly represent an agent's credal state. Therefore,  $p(w)$  is strictly speaking a function from the singleton  $\{w\}$  to a real number, not a function of a world.

$q$ , i.e., that both consider  $q$  more likely than not. That is,  $a$  is a best response for both if  $p(q_i \& q_j) > 1/2$ . Then  $a$  is played if  $\delta(1 - \gamma) > 1/2$ , i.e., if

$$\gamma < 1 - \frac{1}{2\delta} \quad (\text{Confidence Threshold for } a)$$

By similar reasoning,  $b$  is played if

$$\gamma < 1 - \frac{1}{2(1-\delta)} \quad (\text{Confidence Threshold for } b)$$

These inequalities are the confidence thresholds for coordination on  $(a, a)$  and  $(b, b)$  respectively. They are derived assuming that  $\delta$  is neither 0 nor 1, but the generality of the conclusion is not lessened. Such values would trivialize the interaction, for  $\delta = 0$  would mean that the agents do not consider world  $w_1$  a genuine possibility, and on the other hand  $\delta = 1$  would mean that  $w_3$  is not a genuine possibility, given how  $p(w_1)$  and  $p(w_3)$  were defined above.

The value of  $1/2$  as tipping point for action has been chosen somewhat arbitrarily, but the conclusion is representative of a general point. Two conditions characterize the existence of a coordination equilibrium in conditions of uncertainty. For the  $(a, a)$  outcome, it must be that  $\delta > 1 - \delta$  and  $\gamma < 1 - \frac{1}{2\delta}$ ; for  $(b, b)$ , that  $1 - \delta > \delta$  and  $\gamma < 1 - \frac{1}{2(1-\delta)}$ . Uncertainty undermines the agents' confidence that something is the case, so that a belief fails to be public. Nevertheless, if the chance  $\gamma$  that their thinking differently is not too high, coordination may still obtain. The inequalities  $CTa$  and  $CTb$  specify what "not too high" means.

## 2 Coordination in Times of Uncertainty

Confronted with a failure to coordinate beliefs, rational agents could change their mind, of course. However, revising judgements doesn't eliminate vagueness: Shiv and Logan would simply go one step further in the forced march. This may be as good as it gets, if the agents' common language is indeed limited to  $\phi$  and  $\neg\phi$ . Shiv and Logan will then have to learn to live with the occasional failures of coordination. On the other hand, if the agents' language includes epistemic vocabulary, then they could make their uncertainty manifest, and this potentially matters for their attempt to coordinate. To characterize this idea, I will begin with a standard relational semantic for *might*.

The idea that more is communicated in conversation than the semantic content of what the interlocutors say goes back to H. P. Grice [15]. Gricean reasoning has a strategic nature, and an appreciation of this point has led to a more systematic game-theoretic understanding of it [3, 12, 11, 5, 2, 22]. Furthermore, recent work has emphasized the connection between Gricean reasoning and more general Bayesian models of inference under uncertainty that have wide applications in the study of human cognition [13, 14, 18]. The result is a framework for probabilistic inference and back-and-forth reasoning whose outline I will follow in the next sections.

Suppose that, besides the two signals  $\phi$  and  $\neg\phi$ , the agents' language includes epistemic vocabulary. They can utter sentences such as 'The restaurant might be close' and 'The restaurant might be far', namely  $\diamond\phi$  and  $\diamond\neg\phi$ , respectively. The general idea is that a sentence  $\diamond\phi$  is true just in case  $\phi$  is compatible the information some agent has. Roughly, such information is an agent's evidence, or doxastic mental state. For simplicity we may take the relevant agents to be the participants in the game, though of course this would be implausible for the purposes of natural language semantics. If so, then  $\diamond\phi$  is true at a world  $w$  in the game model of Figure 1 just in case there is some agent  $i$  who thinks that  $\phi$  in  $w$ . By

this light, in  $w_2$  it is true to say, ‘It might be that  $\phi$  and it might be that  $\neg\phi$ ’. This seems the right thing to say when one is uncertain, as Shiv and Logan are in  $w_2$ .

More formally, we define a semantic model over the game of Figure 1. The model  $(I, W, \Pi, \llbracket \cdot \rrbracket^{c,s})$  includes a set  $I$  of players, a set  $W$  of worlds, an interpretation function  $\llbracket \cdot \rrbracket^{c,s}$  relative to the context and a variable assignment (superscripts henceforth omitted), and a set of partitions  $\Pi = \{\Pi_i : i \in I\}$  of  $W$ , one partition for each player. A rough but standard Kratzerian semantics for  $\diamond$  can be given in terms of  $\Pi$  [17, 9].

$$\llbracket \diamond\phi \rrbracket = \lambda w. \exists i \in I. \exists \pi_i \in \Pi_i. \exists w' \in \pi_i(w) : w' \in \llbracket \phi \rrbracket$$

Intuitively,  $\diamond\phi$  is true at  $w$  iff there is a  $w'$  accessible from  $w$  such that  $\phi$  is true at  $w'$ . A world is accessible from another just in case they belong to the same cell of some agent’s partition. Thus, epistemically accessible worlds are those that some agent finds indistinguishable on the basis of their doxastic perspective prior to communication. It is straightforward to check that accessibility, thus defined in terms of doxastic partitions, is reflexive, symmetric, and not transitive.

In order to calculate the pragmatic effects of manifesting uncertainty by asserting that  $\diamond\phi$ , let’s refer to the set  $\{w_1, w_2, w_3\}$  as Shiv and Logan’s *common ground* at time 0,  $cg(0)$ : the worlds the interlocutors jointly consider to be possible, at the beginning of their interaction. Following Stalnaker [28, 25, 26], conversation is a cooperative enterprise whereby interlocutors narrow down the common ground. The task for the listener is to figure out which world is actual, given what the speaker said. A simple hypothesis is that worlds in the common ground, at any time, have equal chances of being actual. On the basis of this hypothesis, base-rate probabilities may be easily calculated for any time  $t$ .

$$\text{For all times } t \text{ and for all } w \text{ in } cg(t) : p(w) = \frac{1}{|cg(t)|}$$

Therefore, once the agents narrow down the common ground to  $\{w\}$ , the probability that  $w$  is the actual world is 1. In  $cg(0)$ , we have  $p(w_1) = p(w_2) = p(w_3) = 1/3$ .

Suppose for illustration that Shiv thinks the restaurant is close, but can’t tell if Logan thinks so as well. For all she knows (we could say, semantically ascending), the actual world is  $w_1$  or  $w_2$ : after all, she cannot distinguish these two possibilities. In both  $w_1$  and  $w_2$ ,  $\diamond\phi$  is true, as for both worlds there is an agent, namely Shiv, who thinks that the restaurant is close at those worlds. Therefore (semantically descending), Shiv believes that  $\diamond\phi$  is true. Thus, she asserts so. An assertion is a proposal to update the common ground, by eliminating possibilities that are incompatible with the semantic content of the assertion [28]. By the standard semantics I assumed above, and the Stalnakerian dynamics of assertion, an assertion of  $\diamond\phi$  rules out  $w_3$ , which is incompatible with the truth of the assertion. After Shiv’s assertion that  $\diamond\phi$ , the possibility that the restaurant is definitely far is no longer relevant for either Shiv or Logan.

Logan may now reason that if Shiv had meant to suggest that  $w_1$  is the actual world, she would have sent  $\phi$  (‘The restaurant is close!’) right away, for  $w_1$  is the world in which both think that the restaurant is close. But she didn’t send  $\phi$ : she wasn’t confident enough for that. So, she doesn’t think that the restaurant is definitely close. Since  $w_2$  is the only other possibility left,  $w_2$  must be the actual world according to the speaker. Thus the agents become aware of a distinction between confidence levels by using epistemic language.

This reasoning can be formalized in a Bayesian framework. At time 1, after the update, the listener  $L$  has equal priors for the worlds in  $cg(1)$ , i.e.  $p(w_1) = p(w_2) = 1/2$ . Moreover,  $L$  expects  $S$  to be truthful. Since a truthful speaker could send only  $\phi$  and  $\diamond\phi$  in  $cg(1)$ ,  $L$  holds even priors for the events that these

signals are sent, i.e.  $p(\phi) = p(\diamond\phi) = 1/2$ . Finally,  $L$  expects  $S$  to send  $\phi$  in  $w_1$ , not in  $w_2$ . For an assertion that  $\phi$  reveals the speaker's belief that the restaurant is close, but the speaker believes that the restaurant is close only in  $w_1$ . Therefore,  $L$ 's conditional probability for the event that  $\phi$  is sent given that  $w_1$  is the actual world is at least nearly 1. Conversely for  $\diamond\phi$ .

$$\begin{aligned} p(\phi|w_1) &\approx 1 & p(\diamond\phi|w_1) &\approx 0 \\ p(\phi|w_2) &\approx 0 & p(\diamond\phi|w_2) &\approx 1 \end{aligned}$$

The last step is for  $L$  to update by Bayes rule. The posterior probability that a world is actual is calculated by the listener by conditionalizing on the evidence, namely the observation that  $\diamond\phi$  was sent.

$$p'(w_2) = \frac{p(\diamond\phi|w_2) \cdot p(w_2)}{p(\diamond\phi)} \approx \frac{1 \cdot 1/2}{1/2} \approx 1$$

From the observation that ‘It might be raining’ was uttered, with the semantics it has, and given what else could have been uttered, the listener draws a conclusion about the speaker's confidence level: in the actual world the restaurant is neither definitely close nor definitely far, and in particular the speaker thinks that it's close but she is not confident. The listener's inference is a defeasible one, and not a semantic entailment. Like ordinary pragmatic reasoning, its conclusion is not packaged in the semantic content of the sentence that was uttered by the speaker. Thus, the sentence ‘It might be raining’ is not about the speaker's credal state (or anybody else's, for that matter). Yet it supports a Bayesian inference to a conclusion about the speaker's credences.

### 3 Strategic Hedging

How do rational agents react to someone's assertion that the restaurant might be close? Earlier I assumed that an agent goes out if they think that the restaurant is close, and stays in if they think that it's far. In  $w_2$  the restaurant is neither close nor far, and Shiv thinks that it's close while Logan thinks that it's far. Consequently, they don't coordinate. The assumption is rather crude, however, for we might want to say that uncertainty comes with indecision [20].

If Shiv says ‘The restaurant might be close’, she signals her uncertainty to Logan, who infers it as a good Bayesian. Moreover, Shiv might expect this inference to be of some consequence. Logan would have to take Shiv's uncertainty into account. At the very least, Shiv might expect that Logan hesitates before taking action, once ‘Might  $\phi$ ’ is asserted. I will assume that she does. I will now show that, as a consequence, it's reasonable for Logan to go out, when he is told that the restaurant might be close, even if he thinks that it's far. That is, the chances of coordination improve despite failure of public belief.

Shiv's expectation about Logan's reaction to her utterance kick-starts an expectation-building process. For she will have higher-order expectations about what her reaction will be to what she expects Logan's reaction to her utterance is, and so on. The result is essentially an instance of iterated reasoning between speaker and listener. Taking notice of each other, the interlocutors adjust their propensity to act.<sup>4</sup>

Shiv and Logan's reasoning about each others' actions takes place under the assumption that their mental states are incompatible. In other words, we are in  $w_2$ , and the agents correctly inferred this

<sup>4</sup>There are two slightly different frameworks one could use to reconstruct this process: *iterated best response* models of pragmatic reasoning [12], and *rational speech act* models [10]. The discussion in this section is inspired mainly by the latter, but could be carried out in the former setting with some adjustments.

by Bayesian reasoning as above. Since the agents don't change their mind concerning the restaurant's location as they go through the expectation-building process below, probabilities are normalized at each step. Thus, their mental states remain incompatible throughout, insofar as beliefs can be surmised by dispositions to act. Nevertheless, we will see that Shiv and Logan's expected utilities increase. I indicate with  $p_i(x)$  the probability that agent  $i$  performs action  $x$ , and break down the reasoning in several steps.<sup>5</sup>

**Step 0:** prior to the use of epistemic modals.  $S$  thinks that  $q$ , so she chooses  $a$ . This fixes the speaker's prior, which is  $p_S$  at step 0. At the same time,  $L$  thinks that  $\bar{q}$ , so he chooses  $b$ . Coordination at this stage inevitably fails.

$$\begin{aligned} p_S^0(a) &= 1 & p_S^0(b) &= 0 \\ p_L^0(a) &= 0 & p_L^0(b) &= 1 \end{aligned}$$

**Step 1:** using epistemic modals.  $S$  signals  $\diamond\phi$  and expects that  $L$  hesitates.  $L$ 's expected hesitation is a matter of randomly choosing  $a$  or  $b$ . This fixes the listener's prior, which is  $p_L$  at step 1.

$$p_L^1(a) = p_L^1(b) = 0.5$$

**Step 2:** expectation-building.  $S$  reflects on her action in response to the listener's prior. Each next step from now on is obtained by normalizing an agent's prior with the other player's.

$$p_S^1(a) = \frac{p_S^0(a)}{\sum_{i \in I} p_i(a)} = \frac{p_S^0(a)}{p_S^0(a) + p_L^1(a)} = \frac{1}{1 + 0.5} \approx 0.666$$

**Step 3:** as in the previous step.

$$p_L^2(a) = \frac{p_L^1(a)}{p_L^1(a) + p_S^2(a)} = \frac{0.5}{0.5 + 0.666} \approx 0.428$$

By proceeding in this way, the probability that  $S$  chooses  $a$  in  $w_2$  tends approximately to 0.6, and the probability that  $L$  does so tends approximately to 0.4. Conversely for  $b$ . The step-wise process can be summarized by a system of equations. This is an inductive definition of a function  $f_a(n)$  that maps a number  $n$  that counts the steps, to the probability that an agent takes action  $a$  at step  $n$ . The probability of doing  $a$  for the speaker is given by  $f_a(2n)$ , i.e., for steps indexed by even numbers, whereas the probability of doing  $a$  for the listener is given by  $f_a(2n + 1)$ . A similar series can be defined for  $b$ .

$$\begin{aligned} f_a(0) &= 1 \\ f_a(1) &= 1/2 \\ f_a(n) &= \frac{f_a(n-2)}{f_a(n-1) + f_a(n-2)} \end{aligned}$$

$f_a$  defines a divergent sequence of probabilities, oscillating between approximately 0.4 and approximately 0.6. This can be observed by simple calculation. Analytic proof is quite involved, and left out of the paper.

While they go through the (first few steps of the) stepwise process, the agents' dispositions to act are incompatible throughout, as an effect of normalizing probabilities. However, the margin by which such

<sup>5</sup>More precisely,  $p_i(x)$  is short for  $p_i(x|q_S \ \& \ \bar{q}_L)$ : the conditional probability that  $i$  does  $x$  given that  $S$  thinks that  $q$  and  $L$  thinks that  $\bar{q}$ . We are holding fixed that we are in  $w_2$ , in which the condition  $q_S \ \& \ \bar{q}_L$  holds.

incompatibility causes failures of coordination is reduced with each step. So, they still share no public belief, but their expected utility is higher. Recall that the value of an agent's expected utility for action  $a$ , as calculated above, was:

$$eu_i(a) = p(q_i \& q_j) \cdot u_i(a, a) = \delta(1 - \gamma)$$

However, this equation assumes that one gets a payoff for  $a$  just in case both think that  $q$ . But one's payoff for  $a$  increases, via  $f_a$ , also in proportion to the probability that  $S$  thinks that  $q$ ,  $L$  thinks that  $\bar{q}$ , but both do  $a$ . So we revise the notion of expected utility, indexing it to the number of steps.

$$\begin{aligned} eu_i^n(a) &= eu_i(a) + p(q_S \& \bar{q}_L) \cdot p_S^n(a) \cdot p_L^n(a) \cdot u_i(a, a) \\ &= \delta(1 - \gamma) + \gamma \cdot p_S^n(a) \cdot p_L^n(a) \cdot u_i(a, a) \end{aligned}$$

At Step 0, the listener doesn't think that  $q$ , thus  $p_L^0(a) = 0$ . Therefore, the overall expected utility of  $a$  at 0 is simply  $eu_i(a)$ , as above. Assuming instead that we are at Step 3:

$$eu_i^3(a) = \delta(1 - \gamma) + \gamma \cdot 0.666 \cdot 0.428 = \delta(1 - \gamma) + \gamma \cdot 0.285$$

More generally, for all actions  $x$ , and for all  $n \geq 0$ , the agents' expected utility monotonically increases with the sequence of steps.

$$eu_i^0(x) \leq eu_i^n(x)$$

This argument is fairly abstract, but it's a mathematical reconstruction of a plausible conclusion. A reasoning process can be defined on the basis of the agents' expectations, in reaction to the uncertainty manifested by an assertion of 'Might  $\phi$ '. The base step of the induction is the intuitively plausible idea that the speaker, having signalled her uncertainty, expects the listener to hesitate before acting. Based on this, the speaker reflects on how to react to the listener's hesitation, on how the listener would react to her reaction, and so on. The agents need not have perfect powers of reasoning. They need not follow the induction to infinity. It suffices that one or two steps are taken, and already the use of  $\diamond\phi$  leads to higher expected utility.

If the restaurant is neither close nor far, going out is reasonable not only for Shiv (who thinks with little confidence that the restaurant is close), but also for Logan (who thinks with little confidence that it's far), in response to the speaker's assertion that it might be close. This choice is *reasonable* in the very concrete sense of expected utility maximization. Thus, by hedging one's assertion in conditions of uncertainty via the use of epistemic possibility modals, the chances of coordination improve although public belief fails.

## 4 Conclusion

In this paper, I presented a "proof of concept" for the use of epistemic modal expressions in signalling games in which uncertainty (about what another player thinks) undermines coordination. Vagueness may trigger uncertainty of this kind, since it undermines the belief that others think in the same way as we do. However, by using 'Might  $\phi$ ', we hedge our assertions and make uncertainty manifest. This can be seen by a straightforward application of Bayes' rule, on the basis of a standard semantics for *might* and the Stalnakerian pragmatics of assertion as informative update. In turn, a manifestation of uncertainty may lead interlocutors to accommodate their actions with what they expect the others' actions will be, even though their doxastic mental states remain incompatible throughout. Coordination under uncertainty is facilitated by the strategic assertion of 'Might  $\phi$ '.

By necessity, the view I presented applies only to particular contexts, formalized as particular kinds of games. By no means I suggest that the interaction I described is the only effect epistemic possibility modals have in an interactive setting. The semantics for epistemic modality I adopted is somewhat rough but standard, and could be fine-tuned for the purposes of natural language semantics. The rational speech act model I adopted is an abstract formalization of the computational import of epistemic signalling, but could be understood as an element of a cognitively plausible picture of bounded rationality in interaction.

## References

- [1] Robert J. Aumann (1987): *Correlated Equilibrium as an Expression of Bayesian Rationality*. *Econometrica* 55(1), pp. 1–18, doi:10.2307/1911154.
- [2] Anton Benz & Jon Stevens (2018): *Game-Theoretic Approaches to Pragmatics*. *Annual Review of Linguistics* 4(1), pp. 173–191, doi:10.1146/annurev-linguistics-011817-045641.
- [3] Herbert H. Clark (1996): *Using Language*. Cambridge: Cambridge University Press, doi:10.1017/cbo9780511620539.
- [4] Kris De Jaegher (2003): *A Game-Theoretic Rationale for Vagueness*. *Linguistics and Philosophy* 26(5), pp. 637–659, doi:10.1023/A:1025853728992.
- [5] Kris De Jaegher & Robert van Rooij (2014): *Game-Theoretic Pragmatics under Conflicting and Common Interests*. *Erkenntnis* 79, pp. 769–820, doi:10.1007/s10670-013-9465-0.
- [6] Dorothy Edgington (1992): *Validity, Uncertainty and Vagueness*. *Analysis* 52, pp. 193–204, doi:10.1093/analys/52.4.193.
- [7] Ronald Fagin, Joseph Y. Halpern, Yoram Moses & Moshe Vardi (1995): *Reasoning About Knowledge*. MIT Press, doi:10.7551/mitpress/5803.001.0001.
- [8] Joseph Farrell & Matthew Rabin (1996): *Cheap Talk*. *Journal of Economic Perspectives* 10(3), pp. 103–118, doi:10.1257/jep.10.3.103.
- [9] Kai von Fintel & Anthony S. Gillies (2011): *‘Might’ Made Right*. In Andy Egan & Brian Weatherson, editors: *Epistemic Modality*, Oxford University Press, pp. 108–130, doi:10.1093/acprof:oso/9780199591596.003.0004.
- [10] Michael Frank (2017): *Rational Speech Act Models of Pragmatic Reasoning in Reference Games*. <https://osf.io/x9mre/>.
- [11] M. Franke, T. De Jager & R. Van Rooij (2009): *Relevance in Cooperation and Conflict*. *Journal of Logic and Computation* 22(1), pp. 23–54, doi:10.1093/logcom/exp070.
- [12] Michael Franke (2011): *Quantity implicatures, exhaustive interpretation, and rational conversation*. *Semantics and Pragmatics* 4, doi:10.3765/sp.4.1.
- [13] Noah Goodman & Andreas Stuhlmüller (2013): *Knowledge and Implicature: Modeling Language Understanding as Social Cognition*. *Topics in Cognitive Science* 5, pp. 173–184, doi:10.1111/tops.12007.
- [14] Noah D. Goodman & Michael C. Frank (2016): *Pragmatic Language Interpretation as Probabilistic Inference*. *Trends in Cognitive Sciences* 20(11), pp. 818–829, doi:10.1016/j.tics.2016.08.005.
- [15] H. Paul Grice (1975): *Logic and Conversation*. In M. Ezcurdia & R. J. Stainton, editors: *The Semantics-Pragmatics Boundary in Philosophy*, Broadview Press, pp. 47–59, doi:10.1057/9780230005853\_5.
- [16] Terence Horgan (1994): *Robust Vagueness and the Forced-March Sorites Paradox*. *Philosophical Perspectives* 8, pp. 159–188, doi:10.2307/2214169.
- [17] Angelica Kratzer (2012): *What ‘Must’ and ‘Can’ Must and Can Mean*. Oxford: OUP, doi:10.1093/acprof:oso/9780199234684.003.0001.
- [18] Daniel Lassiter & Noah D. Goodman (2015): *Adjectival vagueness in a Bayesian model of interpretation*. *Synthese* 194(10), pp. 3801–3836, doi:10.1007/s11229-015-0786-1.

- [19] David K. Lewis (1969): *Convention: A Philosophical Study*. Wiley-Blackwell, doi:10.1002/9780470693711.
- [20] John MacFarlane (2016): *Vagueness as Indecision*. *Aristotelian Society Supplementary Volume* 90, pp. 255–283, doi:10.1093/arisup/akw013.
- [21] Yoram Moses, Danny Dolev & Joseph Y. Halpern (1986): *Cheating husbands and other stories: A case study of knowledge, action, and communication*. *Distributed Computing* 1, pp. 167–176, doi:10.1007/bf01661170.
- [22] Prashant Parikh (2019): *Communication and Content*. Berlin, Germany: Language Science Press.
- [23] Matthew Rabin (1990): *Communication between rational agents*. *Journal of Economic Theory* 51(1), pp. 144–170, doi:10.1016/0022-0531(90)90055-o.
- [24] Brian Skyrms (2010): *Signals*. Oxford University Press, doi:10.1093/acprof:oso/9780199580828.001.0001.
- [25] Robert Stalnaker (1999): *Context and Content*. Oxford: Oxford University Press, doi:10.1093/0198237073.001.0001.
- [26] Robert Stalnaker (2002): *Common Ground*. *Linguistics and Philosophy* 25, pp. 701–721, doi:10.1023/A:1020867916902.
- [27] Robert Stalnaker (2006): *Saying and Meaning, Cheap Talk and Credibility*. In: *Game Theory and Pragmatics*, Palgrave Macmillan UK, pp. 83–100, doi:10.1057/9780230285897\_2.
- [28] Robert Stalnaker (2008): *Assertion*. In: *Formal Semantics*, Blackwell, pp. 147–161, doi:10.1002/9780470758335.ch5.
- [29] Peter Vanderschraaf (1995): *Convention as correlated equilibrium*. *Erkenntnis* 42(1), pp. 65–87, doi:10.1007/bf01666812.
- [30] Timothy Williamson (1994): *Vagueness*. Routledge, doi:10.4324/9780203014264.

# Communication Pattern Models: An Extension of Action Models for Dynamic-Network Distributed Systems

Diego A. Velázquez\*  
velazquez-diego@ciencias.unam.mx

Armando Castañeda†  
armando.castaneda@im.unam.mx

David A. Rosenblueth  
drosenbl@unam.mx

Universidad Nacional Autónoma de México

Halpern and Moses were the first to recognize, in 1984, the importance of a formal treatment of knowledge in distributed computing. Many works in distributed computing, however, still employ informal notions of knowledge. Hence, it is critical to further study such formalizations. Action models, a significant approach to modeling dynamic epistemic logic, have only recently been applied to distributed computing, for instance, by Goubault, Ledent, and Rajsbaum. Using action models for analyzing distributed-computing environments, as proposed by these authors, has drawbacks, however. In particular, a direct use of action models may cause such models to grow exponentially as the computation of the distributed system evolves. Hence, our motivation is finding compact action models for distributed systems. We introduce *communication pattern models* as an extension of both ordinary action models and their update operator. We give a systematic construction of communication pattern models for a large variety of distributed-computing models called *dynamic-network models*. For a proper subclass of dynamic-network models called *oblivious*, the communication pattern model remains the same throughout the computation.

## 1 Introduction

A formal treatment of the concept of *knowledge* is important yet little studied in the distributed-computing literature. Authors in distributed computing often refer to the knowledge of the different agents or processes, but typically do so only *informally*. Hence, a formal basis of knowledge in distributed computing would increase the power to prove formal results. The first step in this direction was taken by Halpern and Moses [11]. A topological approach [12] to distributed computing resulted in a further connection [10] with epistemic logic. Such a connection uses epistemic “action models” [1, 8] to capture communication between agents. We observe, however, that the action models proposed in [10] for the *Iterated Immediate Snapshot* (IIS) computing model [4, 5] not only vary at each communication round, but each action model itself is structurally isomorphic to the resulting epistemic model, which paradoxically, requires knowing the desired result beforehand. Our objective is to develop a different connection between *dynamic-network models* [7, 14, 16] (which include the IIS model), and *Dynamic Epistemic Logic* (DEL) [1, 8], more appropriate for computing knowledge change in these systems.

**Context.** In a distributed system, communication is typically performed either by sending and receiving messages, or by writing to, and reading from, a shared memory. The communication patterns (i.e., who communicated with whom) that can occur may change from model to model. When designing and analyzing distributed systems, it is often the case that authors informally refer to what an agent “knows” after an agent performs some action. There is indeed a formal connection between distributed systems

---

\*Diego A. Velázquez is a doctoral student at the Programa de Doctorado en Ciencia e Ingeniería de la Computación, Universidad Nacional Autónoma de México, and is the recipient of a fellowship from CONACyT.

†Armando Castañeda was supported by PAPIIT project IN108720.

and epistemic logic: this connection was initiated by Halpern and Moses in 1984 [11], showing that distributed systems can be rigorously studied from an epistemic-logic viewpoint. Roughly, a distributed protocol is studied through an *epistemic model* with each of its states representing a possible *configuration* of the protocol. Since its discovery, the epistemic-based approach to distributed systems has been fruitful, as shown in the book by Fagin, Halpern, Moses, and Vardi [9].

An important connection between distributed computing and topology was discovered in three independent papers by Borowsky and Gafni [3], Herlihy and Shavit [13], and Saks and Zaharoglou [18] in 1993, and since then this approach has provided useful techniques to show a number of important results in this field. The book by Herlihy, Kozlov, and Rajsbaum [12] provides a comprehensive description of this connection.

Recently, Goubault, Ledent, and Rajsbaum have shown [10] that the epistemic-based approach can be directly connected to the topology-based approach to distributed systems. The topological approach studies a distributed protocol through its topological representation: a geometric object, called *simpli-cial complex*, where each of its faces is associated with a *configuration* of the protocol. In essence, Goubault, Ledent, and Rajsbaum established [10] a correspondence between the topological description of distributed protocols and epistemic models.

A second interesting result of these authors is that the communication patterns allowed in the IIS distributed model can also be described using epistemic-logic tools from DEL: the communication in a distributed model can be modeled with an action model capturing the communication events that can occur, and the *restricted modal product* operator shows how knowledge evolves after agents exchange information in a *communication round*. (A more thorough summary of [10] appears in Section 6.)

We observe that the action models of [10] describing communication in the IIS model have drawbacks: First, such action models are different for each communication round (an ideal representation of communication would not depend on the communication rounds that have been executed so far). In addition, the size of such action models grows exponentially as the computation develops. Moreover, such action models are structurally isomorphic to the epistemic models we wish to compute. The action models of [10], therefore, not only are not useful for computing the epistemic model resulting from a communication event, but are not a succinct representation of the communication that can happen in the IIS model. This phenomenon is opposite to the description of communication in the topological approach, where we have a geometric and compact description of the communication in the IIS model: the communication is clearly described as a subdivision [12, Chapter 11].<sup>1</sup>

**Contributions.** We are interested in the following question: in the spirit of the action-model approach to DEL, is it possible to describe the communication in a distributed model in a compact manner? As a first step, we try to salvage the approach of [10], by attempting to find an action model applicable to every communication round for *two* agents with binary inputs in the IIS model. We exhibit a family of action models with a constant number of events, although each event is labeled with a precondition formula whose size does increase at each communication round. For obtaining these action models, it was crucial to know *in advance* the epistemic model after a communication round. We have not been able to find a similar family for *three or more* agents yet. The case of *m*-ary inputs for  $m \geq 3$  would be even harder to analyze.

The drawbacks of the action models proposed in [10], together with our unsuccessful efforts to find action models for IIS of small size, are motivations for investigating a different approach. We hence consider an extension of action models that allows us to easily derive models of small size. Moreover, we study not only the IIS model but also a larger class of message-passing models called *dynamic-network*

---

<sup>1</sup>Informally, a subdivision results from dividing the faces of a geometrical object into more faces preserving its shape.

*models* [7, 14, 16]. Roughly speaking, in a dynamic-network model, the agents execute infinite sequences of communication rounds. In each round, the agents communicate according to a *communication pattern* that specifies who communicates with whom in that round. A proper subclass of dynamic-network models are those known as *oblivious* that are specified with a set of communication patterns that can occur in any round, regardless of the communication patterns that have occurred so far in the execution. The IIS model can be defined as an oblivious dynamic-network model.

Our main contribution is a simple but powerful extension to the existing action models and its restricted modal product. For every dynamic-network model, we systematically define an infinite sequence of *communication pattern models* that represent how knowledge changes when agents communicate in the *full-information protocol*, hence making our approach amenable to be extended to automated formal verification of distributed systems. For the case of oblivious models, the communication pattern model remains the same all through the execution. Hence, we are able to model communication of oblivious dynamic-network models in constant space.

**Structure of this paper.** The rest of this paper is structured as follows. Section 2 gives an overview of drawbacks arising from a straightforward use of action models in some contexts in multi-agent systems and outlines our solution to overcome such shortcomings. After establishing notation and definitions in Section 3, we explain, in Section 4, an attempt to improve on [10] within the IIS model. Section 5 presents communication pattern models, our modification of action models. Comparison with existing work appears in Section 6, and Section 7 concludes this paper.

## 2 An Overview of Our Proposal

We first motivate our proposal by pointing out a limitation of action models and the restricted modal product that arises in some contexts when modeling the communication that can happen in a multi-agent system. Roughly speaking, sometimes it is impossible to have action models of “small size”. Our discussion here is informal as we are interested in high-level ideas at the moment, hence delaying formal definitions for the next sections.

**The issue.** Let us consider the well-known *coordinated attack* problem where two agents  $a$  and  $b$  wish to schedule an attack. Agent  $a$  has two possible preferences for scheduling the attack,  $n$  for noon or  $d$  for dawn, while agent  $b$  has no initial preference and wishes to learn  $a$ 's. Communication is unreliable: whenever an agent sends a message, such a message can get lost. The epistemic model  $M$  modeling the initial situation before any communication occurs has two worlds, one in which  $a$  prefers to attack at dawn and another one in which  $a$  prefers to attack at noon;  $b$  cannot distinguish between these two worlds. See model  $M$  in Fig. 1.

An action model is a generalization of an epistemic model, where vertices, called *events*, are labeled with arbitrary formulas (as opposed to sets of propositional variables) called *preconditions*. The restricted modal product of an epistemic model  $M$  and an action model  $A$ , denoted  $M \otimes A$ , is an epistemic model where each world is a pair  $(w, e)$ , such that  $w$  is a world in  $M$ ,  $e$  is an event in  $A$ , and the precondition of  $e$  holds in  $w$ . Worlds  $(w, e)$  and  $(w', e')$  are connected with each other for agent  $a$  if both  $w$  and  $w'$  are connected in  $M$  for  $a$ , and  $e$  and  $e'$  are connected in  $A$  for  $a$ . World  $(w, e)$  is labeled with the same label as that of  $w$ . (Formal definitions of action model and restricted modal product appear in Subsect. 4.2.)

A simple action model  $A$  modeling that  $a$  sends its preference to  $b$  has three events: one for each preference  $p \in \{d, n\}$  modeling that  $b$  successfully receives the preference,  $p$ , of  $a$ , with a precondition

specifying that the event can happen only if  $p$  is the preference of  $a$ , and a third event, modeling that  $a$ 's message gets lost, with precondition  $\top$ . Agent  $b$  can distinguish between all events since it either receives  $a$ 's message or not, but  $a$  cannot distinguish between events because messages can get lost. See action model  $A$  in Fig. 1. The restricted modal product  $M \otimes A$  contains four worlds, one for each combination of  $a$ 's initial preference and successful/unsucessful communication.

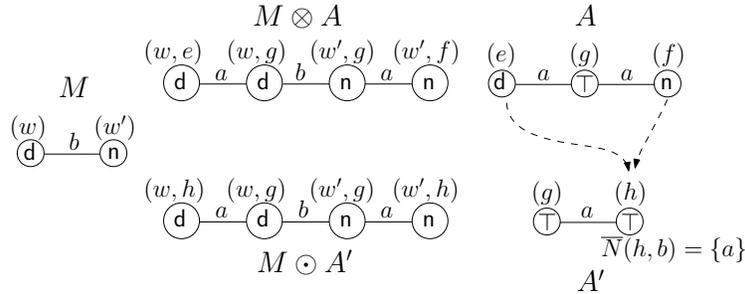


Figure 1: A smaller action model for the coordinated attack problem using our approach.

We observe that the action model  $A$  has the following inconvenience. If  $a$  has  $x > 2$  preferences to schedule the attack instead of only two, a natural generalization of  $A$  has  $x + 1$  events:  $A$  is akin to a star with a “central” event modeling that  $a$ 's message gets lost, and one event for each of the possible preferences of  $a$ . Thus, the size of the action model is proportional to the size of  $a$ 's *input space*.

Can we design a smaller action model for this situation? Can we design an action model with only two events, one corresponding to the case that  $a$ 's message gets lost and another corresponding to the case that  $a$ 's message (with distinct contents, either  $d$  or  $n$ ) reaches  $b$ ? The answer is no. It is easy to see that if we have an action model  $A'$  with only one event  $h$  corresponding to  $a$ 's successful communication, and unavoidably with precondition  $d \vee n = \top$  (see  $A'$  in Fig. 1, discarding for the moment the set  $\bar{N}(h, b) = \{a\}$ ), then  $M \otimes A'$  has again four worlds but now  $b$  cannot distinguish between the worlds  $(w, h)$  and  $(w', h)$  corresponding to the cases where the communication was successful (which is incorrect). A similar situation happens if  $a$  has more than two preferences: it is impossible to have an action model with an event that models the case that  $a$ 's communication is successful. We cannot get any smaller action model in this situation (it might be possible, however, to do so in further rounds) because the action models are designed to deal with “interpreted” events, namely, an event includes the information of the message encoded in its precondition, hence it has a limited ability to represent that *some* information is sent from one agent to another.

This property of action models is a problem in some contexts. Specifically, when studying computability in a given distributed model, it is often the case that the analysis is performed on *protocols* in which agents proceed in a sequence of *rounds* of communication, and in each round every agent sends *all* the information it has collected so far to all other agents; these protocols are called *full-information* in the distributed-computing literature. To be able to reason in the style of DEL, we would like to have an action model modeling the communication events that can happen in a round, and update the epistemic model with the help of the restricted modal product in each round. Drawbacks of the approach in [10] are that the size of the action models proposed there grows exponentially in the number of round and that such action models are structurally isomorphic to the resulting epistemic models. As we will see later, for the case of two agents, we have been able to find a family of action models with a constant number of actions (although the preconditions of the action model do change from round to round) but it is unclear how to find action models with this property for other cases.

**A glimpse of our solution.** Coming back to our initial example, how do we “fix” the problem in  $M \otimes A'$ , i.e., that  $b$  cannot distinguish between worlds  $(w, h)$  and  $(w', h)$ ? Our solution is based on the following observation:  $b$  must be able to distinguish between the two worlds because (1) it receives a message from  $a$  in the event  $h$  of  $A'$ , and (2)  $a$  can distinguish between the  $w$  and  $w'$  in  $M$ . Therefore,  $a$  must send information that makes  $b$  able to distinguish between the two worlds in  $M \otimes A'$ .

We define an extension to the action model formalism, which equips an action model with an additional function  $\bar{N}$  that maps every pair  $(e, a)$  to a set of agents. Intuitively,  $\bar{N}(e, a)$  contains the agents that  $a$  receives messages from when the event  $e$  happens. The restricted modal product is modified by adding two conditions when updating the accessibility relation of an epistemic model. Such conditions say that an agent  $a$  cannot distinguish between two worlds  $(w, e)$  and  $(w', e')$  if and only if  $a$  receives messages from the same set of agents in  $e$  and  $e'$  (i.e.,  $\bar{N}(e, a) = \bar{N}(e', a)$ ) and each of these agents cannot distinguish between  $w$  and  $w'$  (namely,  $\forall a' \in \bar{N}(e, a), w \sim_{a'} w'$ ). The idea is that if those agents sending information to  $a$  cannot distinguish between  $w$  and  $w'$ , then there is no information they send to  $a$  making  $(w, e)$  and  $(w', e')$  distinguishable to  $a$ . The new product is denoted  $\odot$ . Using this formalism, for the coordinated attack problem, we are able to define a *communication pattern model*  $A'$  with a *single* event (called communication pattern in our context)  $h$  corresponding to the case in which  $a$ 's message reaches  $b$ . In  $A'$ ,  $\bar{N}(h, b)$  is set to  $\{a\}$ , and  $\bar{N}$  is set to  $\emptyset$  in any other case. Figure 1 shows the model  $A'$ . Furthermore, the action model is correct regardless of the size of  $a$ 's input space, meaning that the very same action model produces the desired epistemic model if  $M$  represents the situation that  $a$  has  $x > 2$  initial preferences.

### 3 Analyzing distributed computing models

In this section, we give some introductory definitions and fix the notation. We assume some familiarity with basic epistemic logic. We refer to the language of multiagent epistemic logic as  $\mathcal{L}_K$ . Additionally, we consider a non-empty finite set of agents  $Ag = \{a_1, \dots, a_n\}$  and a non-empty finite set of propositions  $Props$ , unless specified otherwise.

**Our models of interest.** We are interested in *dynamic-network models* [7, 14, 16], in which a set of  $n \geq 2$  failure-free agents proceed in an infinite sequence of *synchronous* rounds of communication. Each agent is a *state machine*. In each round, the communication is specified with a *communication graph*, namely a directed graph whose vertex set is  $Ag$ , with each edge  $(a_i, a_j)$  indicating that a message from  $a_i$  to  $a_j$  is successfully delivered in that round. The *in-neighborhood* of an agent  $a_i$  in a communication graph  $G$ , namely the set of agents  $a_j$  such that  $(a_j, a_i)$  is an arrow in  $G$ , is denoted  $N_G^-(a_i)$ . Let  $CP_{Ag}$  denote the set with all communication graphs with vertex set  $Ag$ . Thus, a dynamic-network model  $Adv$  is specified with a set of infinite sequences of graphs of  $CP_{Ag}$ , that we call *adversary*. Intuitively, we say that an adversary  $Adv$  is *oblivious* if in every round, any communication graph in a given set can happen, regardless of the communication graphs that have happened in previous rounds. This is formalized as follows. We say that a finite sequence  $S$  of communication graphs is a *prefix* of  $Adv$  if  $S$  is a prefix of a sequence in  $Adv$ . An adversary  $Adv$  is oblivious if there exists a non-empty subset  $X \subseteq CP_{Ag}$  such that the graphs in  $X$  are the prefixes of  $Adv$  of length one, and for every finite sequence  $S$  that is a prefix of  $Adv$ , it holds that  $S \cdot G$  is a prefix of  $Adv$ , for every graph  $G \in X$ . Thus, an oblivious adversary is simply specified through the set  $X$  of communication graphs; we will say that  $Adv = X$ .

**Protocols.** Each agent locally executes a *protocol* that specifies the messages that the agent sends in a round, depending on the local state of the agent at the beginning of the round. Each agent starts the

computation with a private input, which is the state of the agent at the beginning of the first round. Since we are interested in modeling how knowledge can evolve through the computation, we assume that in every round every agent attempts to communicate to everybody all it knows so far. Formally, every agent locally executes the *full-information* protocol, namely, in every round an agent sends to all other agents all the information such an agent has collected so far. Therefore, the full-information protocol captures all that an agent can know in an execution. The full-information protocol is an important tool in distributed-computing computability research.

**Executions and configurations.** An *execution*  $E$  of an adversary  $Adv$  is a pair  $(I, S^\infty)$ , where  $I = (v_1, v_2, \dots, v_n)$  is an *input vector* denoting that agent  $a_i$  starts with input  $v_i$ , with  $v_i$  belonging to an *input space*, denoted  $In$ , and  $S^\infty$  is a sequence of  $Adv$ . An  $r$ -*execution* of  $Adv$  is a pair  $(I, S)$ , where  $I$  is an input vector and  $S$  is a prefix of  $Adv$  with  $|S| = r$ . A *configuration*  $C$  is an  $n$ -tuple whose  $i$ -th position is a local state of agent  $a_i$  (thus input vectors are configurations). We say that  $a_i$  *does not distinguish* between configurations  $C$  and  $C'$  if and only if  $C(i) = C'(i)$ . An  $r$ -execution  $(I, S)$  *ends* at a configuration  $C$  if each agent  $a_i$  has the local state  $C(i)$  after the execution of the sequence of communication rounds described by  $S$  with the inputs stated by  $I$ ; alternatively, we say that  $C$  is the configuration at the *end* of  $(I, S)$ . Note that for the empty sequence, denoted  $[], I$  is the configuration at the end of the 0-execution  $(I, [])$ , for every input vector  $I$ .

**Our representation.** We use *epistemic models*, that we name  $M^r$ , for representing the  $r$ -executions of a given adversary  $Adv$ . An *epistemic model* for  $Ag$  and a set of propositions  $Props$  is a triple  $M = (W, \sim, L)$ , where  $W$  is a finite set of worlds,  $\sim : Ag \rightarrow \wp(W \times W)$  assigns an equivalence relation to each agent, and  $L : W \rightarrow \wp(Props)$  assigns a set of true-valued propositions to each world. Each world in  $M^r$  represents an  $r$ -execution and the accessibility relations represent the indistinguishability relations over the configurations at the end of the  $r$ -executions of  $Adv$ .

**The initial epistemic model  $(M^0)$ .** We build the initial epistemic model  $M^0 = (W^0, \sim^0, L^0)$  for  $Ag$  and  $In$  with  $Props = \{in_{a,v} \mid a \in Ag \wedge v \in In\}$  so that  $W^0 = \{I \mid I \text{ is an input vector for } Ag \text{ and } In\}$ ,  $I \sim_{a_i}^0 I'$  if and only if  $I(i) = I'(i)$ , and  $L(I) = \{in_{a_i,v} \in Props \mid I(i) = v\}$ . The epistemic model  $M^0$  for the agents  $Ag = \{a, b\}$  and binary inputs  $In = \{0, 1\}$  is depicted in Fig. 2.

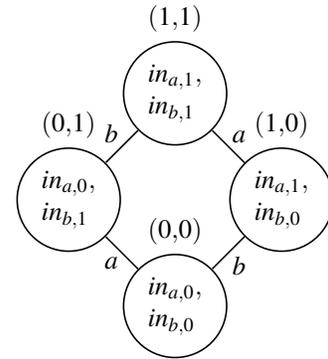


Figure 2: Model  $M^0$  for agents  $a$  and  $b$  with binary inputs.

## 4 Action models and the IIS model

In this section, we first present the IIS model. Next, we give the definition of action models. Finally, we exhibit our best action-model solution of modeling IIS for agents  $a$  and  $b$  with binary inputs.

### 4.1 Iterated Immediate Snapshot distributed-computing model

The IIS model [5] is a fundamental model that fully captures what can be solved in asynchronous wait-free shared-memory systems with process-crash failures. We can define IIS as a (failure-free) synchronous oblivious dynamic-network adversary. The set describing the adversary is as follows. For every

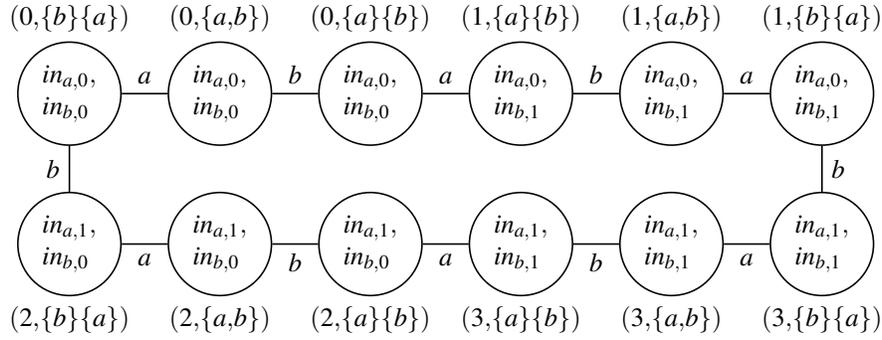


Figure 3: Epistemic model  $M_{\text{IIS}}^1$  that represents the configurations at the end of the first round of the full-information protocol for two agents  $a$  and  $b$  with binary inputs in the IIS model.

sequence of non-empty subsets of  $\text{Ag}$ ,  $S = [C_1, C_2, \dots, C_k]$ , satisfying that  $\text{Ag} = \bigcup C_i$  and  $C_i \cap C_j = \emptyset$  whenever  $i \neq j$ , the adversary has the communication graph with a directed edge  $(a, b)$  for every pair of agents  $a \in C_i, b \in C_j$  with  $1 \leq i \leq j \leq k$ . We say that  $C_i$  is a *concurrency class*. In Fig. 3, we show the epistemic model  $M_{\text{IIS}}^1$  that represents the configurations at the end of the first round of the full-information protocol for processes  $a$  and  $b$  with binary input in the IIS model.

## 4.2 Action models

Action models were introduced in [1] as a general way to model dynamics of knowledge via events.

**Definition 1** (Action model). *An action model  $A$  is a triple  $(E, R, \text{Pre})$ , where  $E$  is a non-empty finite set of events,  $R : P \rightarrow \wp(E \times E)$  is a function that associates each agent with a relation over the set of events, and  $\text{Pre} : E \rightarrow \mathcal{L}_K$  is a function that associates each event with a precondition.*

**Definition 2** (Syntax). *Let  $A = (E, R, \text{Pre})$  be an action model over  $\text{Ag}$  and Props. The language  $\mathcal{L}_{\otimes}$  is given by the following BNF  $\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi \mid [(A, e)]\varphi$ , where  $a \in \text{Ag}$ , and  $p \in \text{Props}$ ,  $e \in E$  and  $(A, e)$  is an update.*

**Definition 3** (Restricted modal product). *Let  $M = (W, \sim, L)$  be an epistemic model over  $\text{Ag}$  and Props. Let  $A = (E, R, \text{Pre})$  be an action model.  $M' = (W', \sim', L') = M \otimes A$  is defined as follows:*

- $W' = \{(w, e) \in W \times E \mid M, w \models \text{Pre}(e)\}$
- $\sim'_a = \{((w, e), (w', e')) \in W' \times W' \mid w \sim_a w' \wedge e R_a e'\}$
- $L'((w, e)) = L(w)$

**Definition 4** (Semantics). *Let  $M = (W, R, L)$  be an epistemic model over  $\text{Ag}$  and Props. Let  $A = (E, R, \text{Pre})$  be an action model. Let  $p \in \text{Props}$  be a proposition. Let  $w, w' \in W$  be worlds. Let  $a \in \text{Ag}$  be an agent. Let  $e \in E$  be an event. Let  $\varphi, \psi \in \mathcal{L}_{\otimes}$  be formulas.*

$$\begin{aligned}
M, w \models p & \quad \text{iff } p \in L(w) \\
M, w \models \neg\varphi & \quad \text{iff } M, w \not\models \varphi \\
M, w \models \varphi \wedge \psi & \quad \text{iff } M, w \models \varphi \text{ and } M, w \models \psi \\
M, w \models K_p\varphi & \quad \text{iff } M, w' \models \varphi \text{ for all } w' \text{ such that } w R(p) w' \\
M, w \models [(A, e)]\varphi & \quad \text{iff } M, w \models \text{Pre}(e) \text{ implies } M \otimes A, (w, e) \models \varphi
\end{aligned}$$

### 4.3 Our best action-model solution for IIS

We now present our best action-model approach of modeling IIS for agents  $a$  and  $b$  with binary inputs. We exploit the fact that for two-agent IIS, *the epistemic models will always be bipartite graphs*. We can hence partition the set of worlds in  $M^i$  into two sets  $W_1^i$  and  $W_2^i$  so that any pair of distinct worlds in the same set can be distinguished by both agents. For each set  $W_j^i$ , we use three events to represent the different sequences of concurrency classes that can happen in a round:  $\{a\}\{b\}$ ,  $\{a,b\}$ , and  $\{b\}\{a\}$  (see Fig. 4). Thus we have six events: three for operating with the worlds in  $W_1^i$ , and three for operating with the worlds in  $W_2^i$ . The sketch of the action model is shown in Fig. 5. In such a sketch, the preconditions,  $\phi_1$  and  $\phi_2$ , change from round to round.  $\phi_j$  is a disjunction of formulas identifying the worlds in  $W_j^i$ . A formula identifying a world is a conjunction of the formulas describing the local state of each agent. In Appendix A, we define functions that compute an epistemic logic formula that describes the local state of an agent. For the first round, if we consider  $W_1^0 = \{(0,0), (1,1)\}$  and  $W_2^0 = \{(0,1), (1,0)\}$ , the preconditions are:  $\phi_1 = (in_{a,0} \wedge in_{b,0}) \vee (in_{a,1} \wedge in_{b,1})$ , and  $\phi_2 = (in_{a,0} \wedge in_{b,1}) \vee (in_{a,1} \wedge in_{b,0})$ .

This approach appears to be a succinct representation of the full-information execution dynamics. There are, however, still issues. We would like to represent communication defined by an oblivious model *just once* because the allowed communication patterns are the same regardless of the round. All correct action models we have been able to find have preconditions that change from round to round. Moreover, the size of the formulas we get from the  $\phi$  functions grows exponentially in the number of rounds. This suggests that in certain cases, a straightforward application of action models might not be ideal.

We have not been able to find a similar family of action models for three agents. We would need to analyze if the corresponding epistemic models are always  $n$ -partite, and how we could join all the needed events. Finding action models for the case of  $m$ -ary inputs for  $m \geq 3$  would be even harder. Making things worse, the analysis might be different in distinct models: we would need to study each model to take advantage of its own characteristics. All these facts motivated us to look for a different and more appropriate approach.

## 5 Communication pattern models

Intuitively, a communication pattern model can be viewed as a non-directed graph whose vertices have two labels: a formula and a communication graph.

### 5.1 Definition of communication pattern models

First, we define our *communication pattern models*. Then, we define the syntax of our language. After that, we define our restricted modal product. Finally, we define our language semantics.

**Definition 5** (Communication pattern model).  $\mathcal{P}$  is a tuple  $(CP, R, Pre, \bar{N})$ , where  $CP$  is a non-empty finite set whose elements are called communication patterns,  $R : Ag \rightarrow \wp(CP \times CP)$  is a function that as-

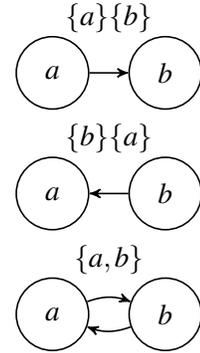


Figure 4: Communication graphs for two-agent IIS.

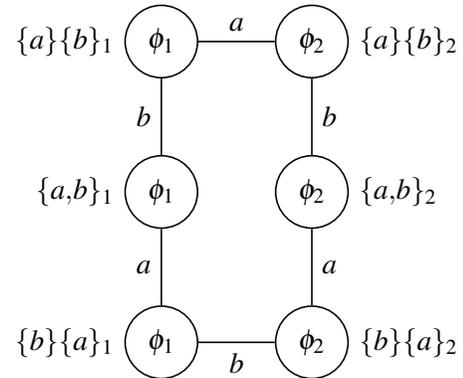


Figure 5: Sketch of the action model for two-agent IIS with binary inputs.

sociates each agent with an equivalence relation over the set of communication patterns,  $Pre : CP \rightarrow \mathcal{L}_K$  is a function that associates each communication pattern with a precondition, and  $\bar{N} : CP \times Ag \rightarrow \wp(Ag)$  is a function that associates a (communication pattern, agent)-pair with a subset of  $Ag$ .

We can think of communication patterns  $cp \in CP$  as communication events. The  $\bar{N}$  function describes the communication graph associated with a communication pattern:  $\bar{N}(cp, a)$  is the in-neighborhood of  $a$  in such a communication graph.

**Definition 6** (Syntax). *Let  $\mathcal{P}$  be a communication pattern model over  $Ag$  and Props. The language  $\mathcal{L}_{\odot}$  is given by the following BNF:*

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_a\varphi \mid [(\mathcal{P}, cp)]\varphi$$

where  $p \in Props$ ,  $a \in Ag$ ,  $cp \in CP$  and  $(\mathcal{P}, cp)$  is an update.

**Definition 7** (Restricted modal product). *Let  $M = (W, \sim, L)$  be an epistemic model over  $Ag$  and Props. Let  $\mathcal{P} = (CP, R, Pre, \bar{N})$  a communication pattern model. Let  $a \in Ag$  an agent.  $(W', \sim', L') = M' = M \odot \mathcal{P}$  is defined as follows:*

- $W' = \{(w, cp) \in W \times CP \mid w \in W \wedge cp \in CP \wedge M, w \models Pre(cp)\}$
- $\sim'_a = \{((w, cp), (w', cp')) \in W' \times W' \mid w \sim_a w' \wedge cp R_a cp' \wedge \frac{\bar{N}(cp, a) = \bar{N}(cp', a)}{w \sim_{a'} w' \forall a' \in \bar{N}(cp, a)}\}$
- $L'((w, cp)) = L(w)$

Intuitively, the first underlined condition requires agent  $a$  to receive information from the same set of processes in both  $cp$  and  $cp'$ , and the second one requires all processes in such a set to send the same information since such processes are required not to distinguish between  $w$  and  $w'$ .

**Definition 8** (Semantics). *Let  $M = (W, \sim, L)$  be an epistemic model over  $Ag$  and Props. Let  $w, w' \in W$  be worlds. Let  $a \in Ag$  be an agent. Let  $\mathcal{P} = (CP, R, Pre, \bar{N})$  be a communication pattern model. Let  $cp \in CP$  be a communication pattern. Let  $\varphi, \psi \in \mathcal{L}_{\odot}$  be formulas.*

$$\begin{aligned} M, w \models p & \quad \text{iff } p \in L(w) \\ M, w \models \neg\varphi & \quad \text{iff } M, w \not\models \varphi \\ M, w \models \varphi \wedge \psi & \quad \text{iff } M, w \models \varphi \text{ and } M, w \models \psi \\ M, w \models K_a\varphi & \quad \text{iff } M, w' \models \varphi \text{ for all } w' \text{ such that } w \sim_a w' \\ M, w \models [(\mathcal{P}, cp)]\varphi & \quad \text{iff } M, w \models Pre(cp) \text{ implies } M \odot \mathcal{P}, (w, cp) \models \varphi \end{aligned}$$

**Action models and communication pattern models.** Communication pattern models are at least as general as action models. Notice that we can build a *degenerate* communication pattern model given an action model. Let  $A = (E, R, Pre)$  be an action model. We build a communication pattern model  $\mathcal{P} = (E, R, Pre, \bar{N})$  so that  $\bar{N}(e, a) = \emptyset \forall (e, a) \in E \times Ag$ . It is easy to see that  $M \otimes A = M \odot \mathcal{P}$  holds.

## 5.2 Communication pattern models for arbitrary adversaries

Consider any adversary  $Adv$  and the initial model  $M^0$  defined in Section 3. Here we define an infinite sequence  $\mathcal{P}^1, \mathcal{P}^2, \dots$  of communication pattern models that succinctly model the evolution of knowledge in the executions of  $Adv$ . More precisely, Theorem 1 in the next section will show that the epistemic model  $M^r = (W^r, \sim^r, L^r) = M^0 \odot \mathcal{P}^1 \odot \mathcal{P}^2 \odot \dots \odot \mathcal{P}^r$  captures how knowledge changes after  $r$  rounds of communication.

For every  $i \geq 1$ , the communication pattern model  $\mathcal{P}^i = (CP^i, R^i, Pre^i, \bar{N}^i)$  is defined as follows:

- $CP^i = \{cp \in CP_{Ag} \mid \exists \text{ an } i\text{-execution } (I, S \cdot cp) \text{ of } Adv\}$ .
- For every  $a \in Ag$ ,  $R_a^i = \{(cp, cp') \in CP^i \times CP^i \mid N_{cp}^-(a) = N_{cp'}^-(a)\}$ .
- For every  $(cp, a) \in CP^i \times Ag$ ,  $\bar{N}^i(cp, a) = N_{cp}^-(a)$ .
- For every  $cp \in CP^i$ , let  $\mathscr{W}_{cp}^{i-1} = \{(I, S) \mid \exists \text{ an } i\text{-execution } (I, S \cdot cp) \text{ of } Adv\}$ . Thus,  $Pre^i(cp) = \bigvee_{(I, S) \in \mathscr{W}_{cp}^{i-1}} \varphi(I, S)$ , where  $\varphi(I, S) = \bigvee_{0 \leq i \leq n} \varphi_i(a_i, C(i))$ . See Appendix A for the definition of  $\varphi_i$ .

**The case of oblivious dynamic-network models.** Following the definition of  $CP^i$ , we can see that, for any oblivious adversary  $Adv$ ,  $CP^i = Adv$ , for each  $i \geq 1$ . Thus, all  $\mathscr{P}^i$  have the same set of communication patterns. Moreover, for each  $cp \in CP^i$ ,  $\mathscr{W}_{cp}^{i-1}$  contains *all*  $(i-1)$ -executions of  $Adv$ , and hence  $Pre^i(cp)$  can be set to  $\top$ . Therefore,  $\mathscr{P}^1 = \mathscr{P}^2 = \dots$ . The communication pattern model representing dynamics for IIS with agents  $a$  and  $b$  is depicted in Fig. 6. For clarity, the function  $\bar{N}$  is not depicted; however, it can be obtained from the in-neighborhoods of the communication graphs in Fig. 4. It is worth observing that the communication pattern in Fig. 6, omitting  $\bar{N}$ , and the usual modal product  $\otimes$  do not model IIS for two agents, not even for the first round. Namely,  $M^1 = M^0 \otimes \mathscr{P}_{\text{two-IIS}}$  has “undesirable” pairs in agent relations which make  $M^1$  structurally different from a 12-cycle, which is the structure of the epistemic model for two processes with binary inputs after one round of communication in IIS (see Fig. 3).

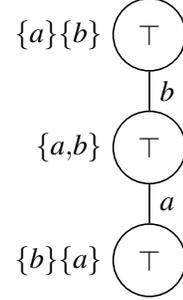


Figure 6: Communication pattern model  $\mathscr{P}_{\text{two-IIS}}$  for two-agent IIS.

### 5.3 The $\odot$ product reflects the change in local states through rounds

The dynamic epistemic logic that we present is focused on reasoning about computations. In particular, we are interested in modeling how *configurations* change in the full-information protocol. A key point is that when updating an epistemic model with our modal product, the resulting epistemic model models how the local states of agents change. Theorem 1 below states that our communication pattern models do model knowledge dynamics. The theorem formalizes this claim using the following notion.

Let  $Adv$  be an adversary. For every  $i \geq 0$ , we define the set  $\mathscr{C}_{Adv}^i = \{C \mid \text{there is an } i\text{-execution } (I, S) \text{ of } Adv \text{ that ends in the configuration } C\}$ . Let  $\mathscr{P}^1, \mathscr{P}^2, \dots$  be an infinite sequence of communication pattern models. We say that the sequence  $\mathscr{P}^1, \mathscr{P}^2, \dots$  *reflects* the adversary  $Adv$  if for each  $r \geq 1$ , there is a bijection  $f^r : W^r \rightarrow \mathscr{C}_{Adv}^r$  such that  $w \sim_{a_i} w'$  if and only if  $a_i$  does not distinguish between  $f^r(w)$  and  $f^r(w')$ , where  $M^0$  is the initial epistemic model and  $M^r = (W^r, \sim^r, L^r) = M^0 \odot \mathscr{P}^1 \odot \mathscr{P}^2 \odot \dots \odot \mathscr{P}^r$ . If  $\mathscr{P}^1 = \mathscr{P}^2 = \dots$ , we simply say that  $\mathscr{P}^1$  reflects  $Adv$ .

**Theorem 1** (Main result). *Let  $Adv$  be an adversary and  $\mathscr{P}^1, \mathscr{P}^2, \dots$  be the communication pattern models built from  $Adv$ , as described in Subsection 5.2. Then,  $\mathscr{P}^1, \mathscr{P}^2, \dots$  reflects  $Adv$ .*

Let  $\mathscr{E}_{Adv}^r$  be the set of all  $r$ -executions of  $Adv$ . Let  $I, I'$  be two input vectors for  $Ag$  and  $In$ . The proof of Theorem 1 will be as follows. First, we will present two lemmas whose proof we omit because of space restrictions. Then, we will prove by induction that  $w_r \sim_{a_i} w'_r$  if and only if  $a_i$  does not distinguish between  $f^r(w_r)$  and  $f^r(w'_r)$ .

Consider  $E_{r+1} = (I, [cp_1, cp_2, \dots, cp_r, cp_{r+1}]) \in \mathscr{E}_{Adv}^{r+1}$ , and  $E_r = (I, [cp_1, cp_2, \dots, cp_r]) \in \mathscr{E}_{Adv}^r$ . We define  $g^r : \mathscr{E}_{Adv}^r \rightarrow \mathscr{E}_{Adv}^r$  as follows:

$$g^0((I, [])) = I.$$

$$g^{r+1}(E_{r+1}) = C_{r+1} = (C_{r+1}(1), C_{r+1}(2), \dots, C_{r+1}(n))$$

where

$$C_{r+1}(i)(j) = \begin{cases} g^r(E_r)(j) & \text{if } a_j \in N_{cp_{r+1}}^-(a_i) \cup \{a_i\} \\ \perp & \text{otherwise} \end{cases}.$$

**Lemma 1.**  $g^r$  is a bijection.

Consider  $w_r = (\dots((I, cp_1), cp_2) \dots, cp_r) \in W^r$ . We define  $h^r : W^r \rightarrow \mathcal{E}_{Adv}^r$  as follows:

$$h^r(w_r) = (I, [cp_1, cp_2, \dots, cp_r]).$$

**Lemma 2.**  $h^r$  is a bijection.

Now, we start with the proof of Theorem 1.

*Proof.* We define

$$f^r : W^r \rightarrow \mathcal{C}^r = g^r \circ h^r.$$

Since  $g^r$  and  $h^r$  are bijective,  $f^r$  is bijective.

Now we prove, by induction on the round number  $r$ , that the epistemic model  $M^r$  reflects indistinguishability between configurations.

**Base case.**

Consider  $I, I' \in W^0$ ,  $C_I = f^0(I) = (I(1), I(2), \dots, I(n))$ , and  $C_{I'} = f^0(I') = (I'(1), I'(2), \dots, I'(n))$ . By construction of  $\sim^0$ ,  $I \sim_{a_i}^0 I'$  if and only if  $I(i) = I'(i)$  holds. Since  $a_i$  does not distinguish between  $C_I$  and  $C_{I'}$  if and only if  $I(i) = I'(i)$  holds,  $I \sim_{a_i}^0 I'$  if and only if  $a_i$  does not distinguish between  $C_I$  and  $C_{I'}$  holds.

**Inductive hypothesis.**

Consider  $M^r = (W^r, \sim^r, L^r) = M^0 \odot \mathcal{P}^1 \odot \mathcal{P}^2 \odot \dots \odot \mathcal{P}^r$ , and  $w_r, w'_r \in W^r$ . We assume that  $f_r : W^r \rightarrow \mathcal{C}_{Adv}^r$  satisfies that  $w_r \sim_{a_i}^r w'_r$  if and only if  $a_i$  does not distinguish between  $f^r(w_r)$  and  $f^r(w'_r)$ .

**Inductive step.**

Consider  $w_{r+1} = (w_r, cp_{r+1})$ ,  $w'_{r+1} = (w'_r, cp'_{r+1}) \in W^{r+1}$ . We need to prove that  $w_{r+1} \sim_{a_i}^{r+1} w'_{r+1}$  if and only if  $a_i$  does not distinguish between  $f^{r+1}(w_{r+1})$  and  $f^{r+1}(w'_{r+1})$ .

Consider  $w_{r+1}, w'_{r+1} \in W^{r+1}$ . By definition of  $f^{r+1}$ , we know that

$$f^{r+1}(w_{r+1}) = C_{r+1} = (C_{r+1}(1), C_{r+1}(2), \dots, C_{r+1}(n))$$

where

$$C_{r+1}(i)(j) = \begin{cases} f^r(w_r)(j) & \text{if } a_j \in \bar{N}^{r+1}(cp_{r+1}, a_i) \cup \{a_i\} \\ \perp & \text{otherwise} \end{cases}$$

and

$$f^{r+1}(w'_{r+1}) = C_{r+1}' = (C'_{r+1}(1), C'_{r+1}(2), \dots, C'_{r+1}(n))$$

where

$$C'_{r+1}(i)(j) = \begin{cases} f^r(w'_r)(j) & \text{if } a_j \in \bar{N}^{r+1}(cp'_{r+1}, a_i) \cup \{a_i\} \\ \perp & \text{otherwise} \end{cases}.$$

By the definition of  $\odot$ ,  $w_{r+1} \sim_{a_i}^{r+1} w'_{r+1}$  if and only if  $w_r \sim_{a_i}^r w'_r$ ,  $cp_{r+1} R_{a_i}^{r+1} cp'_{r+1}$ ,  $\bar{N}^{r+1}(cp_{r+1}, a_i) = \bar{N}^{r+1}(cp'_{r+1}, a_i)$ , and  $w_r \sim_{a_j}^r w'_r \forall a_j \in \bar{N}^{r+1}(cp_{r+1}, a_i)$ .

$C_{r+1}(i)(j) = \perp$  if and only if  $C'_{r+1} = \perp$  holds because by construction of  $\mathcal{P}^{r+1}$ ,  $cp_{r+1} R_{a_i}^{r+1} cp'_{r+1}$  if and only if  $\bar{N}^{r+1}(cp_{r+1}, a_i) = \bar{N}^{r+1}(cp'_{r+1}, a_i)$  holds. Then,  $C_{r+1}(i)(j) = C'_{r+1}(i)(j)$  holds if and only if  $f^r(w_r)(j) = f^r(w'_r)(j)$  holds for all agents in  $\bar{N}^{r+1}(cp_{r+1}, a_i)$ . By the inductive hypothesis, we have

that  $f^r(w_r)(j) = f^r(w'_r)(j) \forall a_j \in \overline{N}^{r+1}(cp_{r+1}, a_i)$  holds. Then,  $w_{r+1} \sim_{a_i}^{r+1} w'_{r+1}$  holds if and only if  $C_{r+1}(i)(j) = C'_{r+1}(i)(j)$ .  $C_{r+1}(i)(j) = C'_{r+1}(i)(j)$  holds if and only if  $C_{r+1}(i) = C'_{r+1}(i)$  holds. Hence,  $C_{r+1}(i) = C'_{r+1}(i)$  holds if and only if  $a_i$  does not distinguish between  $C_{r+1}$  and  $C'_{r+1}$ .  $\square$

**Corollary 1** (Constant space). *Modeling an oblivious adversary Adv with communication pattern models require constant space.*

*Proof.* Let  $\mathcal{P}$  be the communication pattern model for Adv built as described in Subsection 5.2. By Theorem 1,  $\mathcal{P}$  reflects Adv. Moreover,  $\mathcal{P}$  remains the same in all rounds.  $\square$

## 6 Related work

The formal treatment of knowledge in distributed computing was pioneered by Halpern and Moses in [11]. Perhaps their most important result is having proved that common knowledge amounts to simultaneity. The book by Fagin, Halpern, Moses, and Vardi [9] was pivotal, as it summarized numerous results and compared different approaches to studying many aspects of knowledge in a system of agents.

Action models first appeared in [1]. Such a formalism, however, was only considered for modeling evolution of knowledge in distributed systems, as far as we know, in [10], by Goubault, Ledent, and Rajsbaum and in [17], by Pflieger and Schmid.

Closer to our work is [10], where the authors exhibit a tight connection between the topological approach [12] to distributed processing and Kripke models. A second contribution of [10] is employing the *restricted modal product* operator of action models to model knowledge change between agents after a round. A third important result is employing action models to represent “tasks”. A task is the equivalent of a function in distributed computability. The task defines the possible inputs to the agents, and for each set of inputs, it specifies the set of outputs that the agents may produce. By representing the task itself, the possibility of solving a task amounts to the existence of a certain simplicial map.

The objective of [17], which uses action models as well, is that of obtaining lower limits on the number of bits necessary for implementing a protocol that is specified with an initial epistemic model and an infinite sequence of action models that describe how the epistemic model is updated through an infinite sequence of communication rounds. Like us, [17] uses dynamic-network models. Unlike us, [17] assumes that the action model are given. As a result, [17] does not build an action model and is not concerned with the size of the action models.

The work in [2] exhibits drawbacks similar to the ones we found when using the action model framework in other contexts. The authors propose an extension of epistemic models adding a function and an update mechanism. Adding such a function decreases the number of events needed to represent certain problems. Our proposal, however, can be directly applied to the context of distributed systems by the communication between agents.

## 7 Concluding Remarks

The formalization of knowledge in the distributed-computing literature has still to have a more significant impact. The evidence is that many papers in distributed computing refer to knowledge informally.

At the same time, in the epistemic-logic literature, the formalism of action models has emerged as an important mechanism for modeling the evolution of knowledge. Hence, the works by Goubault, Ledent,

and Rajsbaum [10], establishing a connection between action models and a topological approach to distributed systems, and by Pflieger and Schmid [17], modeling a dynamic-network protocol by an initial epistemic model and an infinite sequence of action models, are relevant.

The approach of [10] operates an action model with an epistemic model capturing knowledge at a certain point in time, to obtain a new epistemic model for knowledge after one round of communication. We observed however, that the action models proposed in [10] for IIS have certain inconveniences. Such action models are structurally isomorphic to the desired epistemic model, hence the number of events grows exponentially in the round number.

We proposed a family of action models with six events, for the case of two agents with binary inputs, whose preconditions change from round to round. For obtaining such a family however, we needed to know in advance the structure of the epistemic models in further rounds. Furthermore, the analysis for more agents or even more inputs seems to be more difficult. Hence, a generalization of such a family is unclear for IIS. Moreover, the analysis would depend on how the epistemic models change in different distributed-computing models.

To overcome these disadvantages, we proposed an extension of action models for dealing with communication patterns, called *communication pattern models*. Our models work for a large variety of distributed-computing models, called dynamic-network models. Using our extension, we were able define communication pattern models systematically for every round of execution in the full-information protocol. In the case of oblivious models, which includes IIS, the communication pattern model remains the same all through the computation. In either case, our approach can be applied in automated distributed-systems verification. We emphasize the fact that communication pattern models as presented in this work are designed to deal with the full-information protocol. We plan to analyze modifying definitions to deal with arbitrary protocols.

Communication pattern models were presented as an extension of action models. It is possible, however, to present the same idea with a set of communication graphs. When analyzing arbitrary dynamic-network models, there should be a precondition for each communication pattern. When analyzing oblivious models there is no need of such precondition because it is always true. An advantage of presenting communication pattern models as an extension of action models is that of studying how an action model can be seen in an agent-communication perspective.

An alternative approach to modeling distributed systems epistemically is by the use of *interpreted* systems, as in [11], or in the more recent papers by Castañeda, Gonczarowski, and Moses [6], as well as Moses [15]. In these works, protocols are modeled explicitly, and indistinguishability is generated directly from the local states; consequently there is no need for a communication pattern model (or an action model) that models the dynamics of the system. Since we use epistemic models and communication pattern models, we need to show that the indistinguishability relation that they generate coincides with the one based on local states in the corresponding model, which is shown in Theorem 1. A benefit of our approach, however, is that the communication pattern models that we compute are arguably a succinct representation of the communication that can occur in a model.

## 8 Acknowledgment

We should like to thank Hans van Ditmarsch for his insightful comments.

## References

- [1] Alexandru Baltag, Lawrence S. Moss & Slawomir Solecki (1998): *The logic of public announcements, common knowledge, and private suspicions*. In: *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 1998)*, pp. 43–56. Available at <http://dl.acm.org/citation.cfm?id=645876.671885>.
- [2] Adam Bjorndahl & Will Nalls (2021): *Endogenizing Epistemic Actions*. *Studia Logica*, doi:10.1007/s11225-020-09937-8.
- [3] Elizabeth Borowsky & Eli Gafni (1993): *Generalized FLP impossibility result for t-resilient asynchronous computations*. In: *Proceedings of the Twenty-Fifth ACM Symposium on Theory of Computing (STOC 1993)*, pp. 91–100, doi:10.1145/167088.167119.
- [4] Elizabeth Borowsky & Eli Gafni (1993): *Immediate atomic snapshots and fast renaming (extended abstract)*. In: *Proceedings of the Twelfth Annual ACM Symposium on Principles of Distributed Computing (PODC 1993)*, pp. 41–51, doi:10.1145/164051.164056.
- [5] Elizabeth Borowsky & Eli Gafni (1997): *A simple algorithmically reasoned characterization of wait-free computation (extended abstract)*. In: *Proceedings of the Sixteenth Annual ACM Symposium on Principles of Distributed Computing (PODC 1997)*, ACM, pp. 189–198, doi:10.1145/259380.259439.
- [6] Armando Castañeda, Yannai A. Gonczarowski & Yoram Moses (2014): *Unbeatable Consensus*. In Fabian Kuhn, editor: *Distributed Computing*, Springer Berlin Heidelberg, p. 91–106, doi:10.1007/978-3-662-45174-8\_7.
- [7] Bernadette Charron-Bost & André Schiper (2009): *The Heard-Of Model: Computing in Distributed Systems with Benign Faults*. *Distributed Comput.* 22(1), pp. 49–71, doi:10.1007/s00446-009-0084-6.
- [8] Hans van Ditmarsch, Wiebe van der Hoek & Barteld Kooi (2008): *Dynamic Epistemic Logic*. Springer, doi:10.1007/978-1-4020-5839-4.
- [9] Ronald Fagin, Joseph Y. Halpern, Yoram Moses & Moshe Y. Vardi (1995): *Reasoning about Knowledge*. MIT Press, doi:10.7551/mitpress/5803.001.0001.
- [10] Eric Goubault, Jérémy Ledent & Sergio Rajsbaum (2018): *A simplicial complex model for dynamic epistemic logic to study distributed task computability*. In: *Proceedings of the Ninth International Symposium on Games, Automata, Logics, and Formal Verification (GandALF 2018)*, pp. 73–87, doi:10.4204/EPTCS.277.6.
- [11] Joseph Y. Halpern & Yoram Moses (1984): *Knowledge and common knowledge in a distributed environment*. In: *Proceedings of the Third ACM Symposium on Principles of Distributed Computing (PODC 1984)*, pp. 50–61, doi:10.1145/79147.79161.
- [12] Maurice Herlihy, Dmitry Kozlov & Sergio Rajsbaum (2014): *Distributed Computing Through Combinatorial Topology*. Morgan-Kaufmann, doi:10.1016/C2011-0-07032-1.
- [13] Maurice Herlihy & Nir Shavit (1993): *The asynchronous computability theorem for t-resilient tasks*. In: *Proceedings of the Twenty-Fifth ACM Symposium on Theory of Computing (STOC 1993)*, pp. 111–120, doi:10.1145/167088.167125.
- [14] Fabian Kuhn & Rotem Oshman (2011): *Dynamic networks: models and algorithms*. *SIGACT News* 42(1), pp. 82–96, doi:10.1145/1959045.1959064.
- [15] Yoram Moses (2016): *Relating Knowledge and Coordinated Action: The Knowledge of Preconditions Principle*. *Electronic Proceedings in Theoretical Computer Science* 215, p. 231–245, doi:10.4204/eptcs.215.17.
- [16] Thomas Nowak, Ulrich Schmid & Kyrill Winkler (2019): *Topological Characterization of Consensus under General Message Adversaries*. In Peter Robinson & Faith Ellen, editors: *Proceedings of the 2019 ACM Symposium on Principles of Distributed Computing (PODC 2019)*, ACM, pp. 218–227, doi:10.1145/3293611.3331624.
- [17] Daniel Pflieger & Ulrich Schmid (2018): *On knowledge and communication complexity in distributed systems*. In: *International Colloquium on Structural Information and Communication Complexity (SIROCCO 2018)*, Springer, pp. 312–330, doi:10.1007/978-3-030-01325-7\_27.

- [18] Michael Zacks & Fotios Zaharoglou (1993): *Wait-free k-set agreement is impossible: the topology of public knowledge*. In: *Proceedings of the Twenty-Fifth ACM Symposium on Theory of Computing (STOC 1993)*, pp. 101–110, doi:10.1145/167088.167122.

## A Views and epistemic formulas

Here, we first show a way of thinking about local states in distributed computing called *views*. We then give a formal way of representing such views with an epistemic-logic formula.

In the distributing-computing literature, it is common to regard the local states of the agents as their views. We can think of a view of an agent as a single variable whose value changes from round to round. Such a view takes different values depending on the round.

**Definition 9 (View).** Consider  $S_k = [cp_1, cp_2, \dots, cp_k]$ , and  $S_{k+1} = S_k \cdot cp_{k+1}$  so that  $(I, S_{k+1})$  is a  $k+1$ -execution. The view of an agent  $a_i$  in a execution  $(I, S)$ ,  $view(a_i, (I, S))$  for short, in the full-information protocol is defined inductively as follows:

$$view(a_i, (I, [])) = I(i).$$

$$view(a_i, (I, S_{k+1})) = [view[1], view[2], \dots, view[n]]$$

where

$$view[j] = \begin{cases} view(a_j, (I, S_k)) & \text{if } a_j \in N_{cp_{k+1}}^- \cup \{a_i\} \\ \perp & \text{otherwise} \end{cases}$$

In the full-information protocol, each agent tries to communicate its whole local state to the other agents. If  $a_i$  receives a message from  $a_j$ ,  $a_i$  will know all that  $a_j$  knew in the previous round, otherwise  $a_i$  will not be able to know what  $a_j$  could know.

Now, we formalize the notion of views building an epistemic logic formula for the view of  $a_i$ .

**Definition 10.** Let  $Views^k$  be the set of all possible views of the agents in  $Ag$  at the end of the  $k$ -th round. Let  $Views_i^k$  be the set of all possible views of the agent  $a_i$  at the end of the  $k$ -th round. Consider  $view = [view[1], view[2], \dots, view[n]] \in Views^{k+1}$ . We define the functions  $\varphi_k : P \times Views^k \rightarrow \mathcal{L}_K$ , for all  $k \in \mathbb{N} \cup \{0\}$  as follows:

$$\varphi_0(a_i, v) = in_{a_i, v}.$$

where  $v \in In$ .

$$\varphi_{k+1}(a_i, view) = \bigwedge_{j=1}^n \begin{cases} K_{a_i}(\varphi_k(a_j, view[j])) & \text{if } view[j] \neq \perp \\ \bigwedge_{view' \in Views_j^k} \neg K_{a_i}(\varphi_k(a_j, view')) & \text{otherwise} \end{cases}$$