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.

J. Y. Halpern and A. Perea (Eds.): Theoretical Aspects
of Rationality and Knowledge 2021 (TARK 2021)
EPTCS 335, 2021, pp. 281–294, doi:10.4204/EPTCS.335.27

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 φ if φ 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 responsiblity 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.[1]

Definition 1 (Language) *Given a finite set Ags of agent names, a countable set of propositions P such that $p \in P$ and $\alpha \in Ags$, the grammar for the formal language $\mathscr{L}_{\mathtt{KO}}$ is given by:*

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \psi \mid \Box\varphi \mid [\alpha]\varphi \mid K_\alpha\varphi \mid \odot[\alpha]\varphi \mid \odot_{\mathscr{S}}[\alpha]\varphi \ .$$

$\Box\varphi$ is meant to express the 'historical necessity' of φ ($\Diamond\varphi$ abbreviates $\neg\Box\neg\varphi$). $[\alpha]\varphi$ stands for 'agent α has seen to it that φ.' K_α is the epistemic operator for α, so that $K_\alpha\varphi$ stands for 'agent α knows that φ holds.' $\odot[\alpha]\varphi$ is meant to express that α objectively ought to have seen to it that φ. Finally, $\odot_{\mathscr{S}}[\alpha]\varphi$ is meant to express that α subjectively ought to have seen to it that φ.

As for the semantics, the structures on which we evaluate formulas of the language $\mathscr{L}_{\mathtt{KO}}$ 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 $\langle M, \sqsubset, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha\in Ags}, \mathbf{Value}\rangle$ 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 α and moment m to a partition $\mathbf{Choice}^m_\alpha$ of H_m, where the cells of such a partition represent α's available actions at m. For $m \in M$ and $h \in H_m$, we denote the equivalence class of h in $\mathbf{Choice}^m_\alpha$ by $\mathbf{Choice}^m_\alpha(h)$.* **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 $L \in \mathbf{Choice}^m_\alpha$.* (IA) Independence of agency*: A function s on Ags is called a* selection function *at m if it assigns to each α a member of $\mathbf{Choice}^m_\alpha$. 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 $s \in \mathbf{Select}^m$, $\bigcap_{\alpha\in Ags} s(\alpha) \neq \emptyset$ (see [7] for a discussion of the property).*
- *For $\alpha \in Ags$, $\sim_\alpha$ is the epistemic indistinguishability equivalence relation for agent α, 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 $h'_* \in \mathbf{Choice}^{m_*}_\alpha(h_*)$. 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 $\alpha \in Ags$, we define α'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 φ is taken to be an objective obligation of an agent at a given index iff φ is an

[1] 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 $\beta \in Ags$, we define $\mathbf{State}^m_\beta = \left\{ S \subseteq H_m; S = \bigcap_{\alpha \in Ags-\{\beta\}} s(\alpha), \text{ where } s \in \mathbf{Select}^m \right\}$. For $\alpha \in Ags$ and $m_* \in M$, we first define a general ordering $\leq$ on $\mathscr{P}(H_{m_*})$ 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 $L, L' \in \mathbf{Choice}^{m_*}_\alpha$, $L \preceq L'$ iff for each $S \in \mathbf{State}^{m_*}_\alpha, L \cap S \leq L' \cap S$. The optimal set of actions is the set $\mathbf{Optimal}^{m_*}_\alpha := \{L \in \mathbf{Choice}^{m_*}_\alpha; \text{there is no } L' \in \mathbf{Choice}^{m_*}_\alpha \text{such that } L \prec L'\}$.

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 $\alpha \in Ags$, $m_*, m \in M$, and $L \subseteq H_{m_*}$, L's *epistemic cluster* at m is the set $[L]^m_\alpha := \{h \in H_m; \exists h_* \in L \text{ s.t. } \langle m_*, h_* \rangle \sim_\alpha \langle m, h \rangle\}$. 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 $\mathbf{Choice}^{m_*}_\alpha$ 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 $S \in \mathbf{State}^m_\alpha, [L]^m_\alpha \cap S \leq [L']^m_\alpha \cap S$. Just as in the case of objective ought-to-do's, this ordering allows us to think about a subjectively optimal set of actions $\mathbf{S-optimal}^{m_*}_\alpha := \{L \in \mathbf{Choice}^{m_*}_\alpha; \text{ there is no } L' \in \mathbf{Choice}^{m_*}_\alpha \text{ s. t. } L \prec_s L'\}$, 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 φ 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 $\mathscr{M}$ consists of the tuple that results from adding a valuation function $\mathscr{V}$ to a* eaubt-*frame, where $\mathscr{V} : P \to 2^{M \times H}$ assigns to each atomic proposition a set of moment-history pairs. Relative to a model $\mathscr{M}$, the semantics for the formulas of $\mathscr{L}_{\texttt{KOBO}}$ is defined recursively by the following truth conditions, evaluated at a given index $\langle m, h \rangle$:*

$\mathscr{M}, \langle m,h \rangle \models p$	*iff*	$\mathscr{M}, \langle m,h \rangle \in \mathscr{V}(p)$
$\mathscr{M}, \langle m,h \rangle \models \neg\varphi$	*iff*	$\mathscr{M}, \langle m,h \rangle \not\models \varphi$
$\mathscr{M}, \langle m,h \rangle \models \varphi \wedge \psi$	*iff*	$\mathscr{M}, \langle m,h \rangle \models \varphi$ *and* $\mathscr{M}, \langle m,h \rangle \models \psi$
$\mathscr{M}, \langle m,h \rangle \models \Box\varphi$	*iff*	*for each* $h' \in H_m, \mathscr{M}, \langle m,h' \rangle \models \varphi$
$\mathscr{M}, \langle m,h \rangle \models [\alpha]\varphi$	*iff*	*for each* $h' \in \mathbf{Choice}^m_\alpha(h), \mathscr{M}, \langle m,h' \rangle \models \varphi$
$\mathscr{M}, \langle m,h \rangle \models K_\alpha\varphi$	*iff*	*for each* $\langle m',h' \rangle$ *s.t.* $\langle m,h \rangle \sim_\alpha \langle m',h' \rangle, \mathscr{M}, \langle m',h' \rangle \models \varphi$
$\mathscr{M}, \langle m,h \rangle \models \odot[\alpha]\varphi$	*iff*	*for each* $L \in \mathbf{Optimal}^m_\alpha, h' \in L$ *implies that* $\mathscr{M}, \langle m,h' \rangle \models \varphi$
$\mathscr{M}, \langle m,h \rangle \models \odot_{\mathscr{S}}[\alpha]\varphi$	*iff*	*for each* $L \in \mathbf{S-optimal}^m_\alpha$, *for each* m' *s.t.* $m \sim_\alpha m', h' \in [L]^{m'}_\alpha$ *implies that* $\mathscr{M}, \langle m,h' \rangle \models \varphi$.

Satisfiability, validity on a frame, and general validity are defined as usual. We write $\|\varphi\|$ to refer to the set $\{\langle m,h \rangle \in M \times H; \mathscr{M}, \langle m,h \rangle \models \varphi\}$.

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 agentive 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.[2] 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 $B^{\psi}_{\alpha}\varphi$, which are meant to be read as 'after learning ψ, agent α believes that φ was the case (before the learning).'

Definition 4 (Syntax with conditional-belief) *The grammar for the formal language $\mathscr{L}_{KOB}$ (an extension of $\mathscr{L}_{KO}$) is given by:* $\varphi := p \mid \neg\varphi \mid \varphi \wedge \psi \mid \Box\varphi \mid [\alpha]\varphi \mid K_{\alpha}\varphi \mid B^{\psi}_{\alpha}\varphi \mid \odot[\alpha]\varphi \mid \odot_{\mathscr{S}}[\alpha]\varphi$.

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 $\langle M, \sqsubset, \mathbf{Choice}, \{\sim_{\alpha}\}_{\alpha\in Ags}, \{\mu_{\alpha}\}_{\alpha\in Ags}, \mathbf{Value}\rangle$ such that $M, \sqsubset, \mathbf{Choice}, \{\sim_{\alpha}\}_{\alpha\in Ags}$, and* **Value** *are like in Definition 2, and for every $\alpha \in Ags$, $\mu_{\alpha} : \mathscr{P}(M\times H)\times\mathscr{P}(M\times H) \to [0,1]$ is such that (a) for each $B \subseteq M\times H$, $\mu^{B}_{\alpha} := \mu_{\alpha}(\cdot|B)$ is either a constant function with value 1 or a classical-probability function on $M\times H$, and (b) for every $A,B,C \subseteq M\times H$, $\mu_{\alpha}(A\cap B|C) = \mu_{\alpha}(A|B\cap C)\cdot\mu_{\alpha}(B|C)$.*[3]

A dcpbt-*model $\mathscr{M}$ results from adding a valuation function $\mathscr{V}$ to a* dcpbt-*frame, and the semantics for the formulas of $\mathscr{L}_{KOB}$ over such a model is defined recursively as in Definition 3, with the following additional clause: $\mathscr{M}, \langle m,h\rangle \models B^{\psi}_{\alpha}\varphi$ 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 $q : \mathscr{P}(M\times H)\times\mathscr{P}(M\times H) \to [0,1]$ that meets the requirements (a) and (b) of Definition 5 as a* vf-*function. Therefore, if $p : M\times H \to [0,1]$ is a classical-probability function, then the function $p_c : (M\times H)\times(M\times H) \to [0,1]$, 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.*

[2] 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.

[3] This means that μ_{α} 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 μ^{C}_{α} 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: $K_\alpha\varphi \to B_\alpha^\psi\varphi$ (*Persistence of knowledge*); $B_\alpha^\psi\varphi \to K_\alpha B_\alpha^\psi\varphi$ and $\neg B_\alpha^\psi\varphi \to K_\alpha\neg B_\alpha^\psi\varphi$ (*Full introspection of belief*). Additionally, the following axioms that regard revision are also valid: $B_\alpha^\varphi\varphi$ (*Hypotheses are accepted*); $B_\alpha^\psi\varphi \to (B_\alpha^{\varphi\wedge\psi}\theta \leftrightarrow B_\alpha^\psi\theta)$ and $\neg B_\alpha^\psi\neg\varphi \to (B_\alpha^{\varphi\wedge\psi}\theta \leftrightarrow B_\alpha^\psi(\varphi\to\theta))$ (*Minimality of revision*); $\varphi \to \neg B_\alpha^\varphi\bot$ (*Weak consistency of belief*).[4]

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]).[5]

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

[4]Observe that the semantics for conditional belief implies that an agent's belief of φ given ψ 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.

[5]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,[6] 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—$\mathbf{Value}(h)$—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 $\mathscr{M}$ be a* dcpbt*-model. Let $m \in M$, $h \in H_m$, and $\alpha \in Ags$. Let $L \in \mathbf{Choice}^m_\alpha$. We define α's* expected deontic utility *of L at $\langle m,h\rangle$—denoted by $EU^{\langle m,h\rangle}_\alpha(L)$—as the value given by the following formula:* $EU^{\langle m,h\rangle}_\alpha(L) := \sum_{m' \sim_\alpha m, h' \in [L]^{m'}_\alpha} \mu_\alpha(\{h'\} \mid \pi_\alpha[\langle m',h'\rangle]) \cdot \mathbf{Value}(h')$.

Remark 2 *This means that we calculate α'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 α assigns to those histories, conditional on α's information set at the index where α*

[6] 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 $L \in \mathbf{Choice}^m_\alpha$, we have that $EU^{\langle m,h\rangle}_\alpha(L) = EU^{\langle m,h'\rangle}_\alpha(L)$ 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 $\{EU^{\langle m,h\rangle}_\alpha(L); L \in \mathbf{Choice}^m_\alpha\}$ has a maximum. Therefore, there are actions that maximize α's expected deontic utility at every index, namely the ones whose expected deontic utility is the same as said maximum. We denote by $\mathbf{EU}^{\langle m,h\rangle}_\alpha$ the set of actions that maximize α's expected deontic utility at $\langle m,h\rangle$.[7]

Definition 7 (Full syntax) *The grammar for the formal language $\mathscr{L}_{KOBO}$ (an extension of $\mathscr{L}_{KOB}$) is given by: $\varphi := p \mid \neg\varphi \mid \varphi \wedge \psi \mid \Box\varphi \mid [\alpha]\varphi \mid K_\alpha\varphi \mid B^\psi_\alpha\varphi \mid \odot[\alpha]\varphi \mid \odot_{\mathscr{S}}[\alpha]\varphi \mid \odot_{\mathscr{B}}[\alpha]\varphi$.*

We use the same *dcpbt*-models from Definition 5 to evaluate the formulas of $\mathscr{L}_{\mathsf{KOBO}}$. The truth conditions are the same as before, with the following additional clause: $\mathscr{M}, \langle m,h\rangle \models \odot_{\mathscr{B}}[\alpha]\varphi$ iff for each $L \in \mathbf{EU}^{\langle m,h\rangle}_\alpha$ we have that $[L]^{m'}_\alpha \subseteq \|\varphi\|$ 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 φ iff φ 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 $Ags = \{patient, doctor\}$. Let M and $\sqsubset$ 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*

[7] 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 $EU^{\langle m,h\rangle}_\alpha(L) = E(\mathbf{Value}|\pi_\alpha[\langle m,h'\rangle])$ 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— $\mathbf{Value}(h_1) = 1$*—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 μ_{doctor}, given by the following rules. Let $p_{doctor} : \mathscr{P}(M \times H) \to [0,1]$ be a discrete classical-probability function such that $p_{doctor}(\langle m_i, h_j \rangle) = \frac{9}{4}$ for $i, j \in \{1,2\}$, $p_{doctor}(\langle m_1, h_i \rangle) = \frac{1}{4}$ for $i \in \{3,4\}$, and $p_{doctor}(\langle m_3, h_i \rangle) = \frac{1}{4}$ for $i \in \{3,4\}$. We then define μ_α so that $\mu_{doctor}(A|B) = \frac{p_{doctor}(A \cap B)}{p_{doctor}(B)}$.*

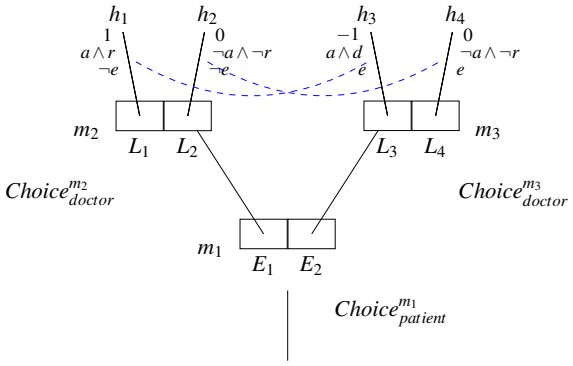

Figure 1: Example from *The Verdict*

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. ($\mathscr{M}, \langle m_2, h_1 \rangle \models e \wedge a \wedge r$).

As for the evaluation of formulas involving the basic stit-theory operators, observe that some instances of it are $\mathscr{M}, \langle m_2, h_1 \rangle \models \Box e$ (*at this index it was the case that it was settled that the patient did not eat*), $\mathscr{M}, \langle m_3, h_3 \rangle \models [doctor]d$ (*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 $j \in \{1,2,3,4\}$, $\mathscr{M}, \langle m_i, h_j \rangle \models \neg K_{doctor} e \wedge \neg K_{doctor} \neg e$ (*at said indices it was the case that the doctor did*

not know whether the patient had eaten or not) and $\mathscr{M},\langle m_i,h_j\rangle \models K_{doctor}[doctor]a \vee K_{doctor}[doctor]\neg a$ (*at said indices it was the case that the doctor knew whether she was supplying the anesthetics or not*). Moreover, we have that $\mathscr{M},\langle m_3,h_3\rangle \models \neg B_{doctor}\neg e$ (*at the actual index it was not the case that the doctor* fully and unconditionally *believed that the patient had not eaten*), that $\mathscr{M},\langle m_3,h_3\rangle \models B^{e}_{doctor}d$ (*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 the patient would die*), and that $\mathscr{M},\langle m_3,h_3\rangle \models B^{e}_{doctor}[doctor]d$ (*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

- $\mathbf{Optimal}^{m_2}_{doctor} = \{L_1\}$, $\mathbf{Optimal}^{m_3}_{doctor} = \{L_4\}$.
- $\mathbf{S-Optimal}^{m_2}_{doctor} = \{L_1,L_2\}$, $\mathbf{S-Optimal}^{m_3}_{doctor} = \{L_3,L_4\}$.
- $\mathbf{EU}^{\langle m_2,h_i\rangle}_{doctor} = \{L_1\}$ ($i \in \{1,2\}$), $\mathbf{EU}^{\langle m_3,h_i\rangle}_{doctor} = \{L_3\}$ ($i \in \{3,4\}$).

Therefore, for instance we have that $\mathscr{M},\langle m_2,h_i\rangle \models \odot[doctor]a \wedge \neg K_{doctor}\odot[doctor]a$ ($i \in \{1,2\}$) (*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 $\mathscr{M},\langle m_3,h_i\rangle \models \odot[doctor]\neg a \wedge \neg K_{doctor}\odot[doctor]\neg a$ ($i \in \{3,4\}$) (*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 $\mathscr{M},\langle m_i,h_j\rangle \models \neg\odot_{\mathscr{S}}[doctor]a \wedge \neg\odot_{\mathscr{S}}[doctor]\neg a$ ($i \in \{1,2\}$ 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 $\mathscr{M},\langle m_i,h_j\rangle \models \odot_{\mathscr{B}}[doctor]a \wedge K_{doctor}\odot_{\mathscr{B}}[doctor]a$ ($i \in \{2,3\}$ 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 $\mathscr{M},\langle m_3,h_3\rangle \models B^{\neg e}_{doctor}(\neg e \wedge \odot[doctor]a)$ (*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 $\odot_{\mathscr{B}}[\alpha]$. 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: $\odot_{\mathscr{B}}[\alpha]\varphi \rightarrow \Diamond K_\alpha\varphi$. We also have that if an agent doxastically ought to have seen to it that φ, then the agent knows that this is settled—the formula $\odot_{\mathscr{B}}[\alpha]\varphi \rightarrow K_\alpha\Box\odot_{\mathscr{B}}[\alpha]\varphi$ 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 $\not\models \odot_{\mathscr{B}}[\alpha]\varphi \rightarrow \odot[\alpha]\varphi$ and that $\not\models \odot[\alpha]\varphi \rightarrow \odot_{\mathscr{B}}[\alpha]\varphi$ (as can be inferred from Example 1). Similarly, we have that $\not\models \odot[\alpha]_{\mathscr{B}}\varphi \rightarrow \odot_{\mathscr{S}}[\alpha]\varphi$ and that $\not\models \odot_{\mathscr{S}}[\alpha]\varphi \rightarrow \odot_{\mathscr{B}}[\alpha]\varphi$. The first invalidity can be inferred from Example 1. The second one can be inferred from a variation of Example 1 as follows: let $\mathbf{Value}(h_1) = 2$, $\mathbf{Value}(h_2) = 1$, $\mathbf{Value}(h_3) = 0$, and $\mathbf{Value}(h_4) = 0$; let p_{doctor} be a discrete classical-probability function such that $p_{doctor}(\langle m_2,h_1\rangle) = \frac{1}{2}$, $p_{doctor}(\langle m_2,h_2\rangle) = \frac{9}{2}$, $p_{doctor}(\langle m_3,h_3\rangle) = \frac{9}{2}$, $p_{doctor}(\langle m_3,h_4\rangle) = \frac{1}{2}$, and p_{doctor} has constant value 0 on

all other indices; let μ_α be defined as in Example 1; then it is the case that $\mathscr{M}, \langle m_i, h_j \rangle \models \odot_{\mathscr{S}}[doctor]a \wedge \odot_{\mathscr{B}}[doctor]\neg a$ ($i \in \{2,3\}$ 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 Λ 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_α. The following axioms and schemata for the interactions of formulas with the given operators:*

$\odot[\alpha](\varphi \to \psi) \to (\odot[\alpha]\varphi \to \odot[\alpha]\psi)$	$(A1)$
$\Box\varphi \to [\alpha]\varphi \wedge \odot[\alpha]\varphi$	$(A2)$
$\Box \odot [\alpha]\varphi \vee \Box\neg \odot [\alpha]\varphi$	$(A3)$
$\odot[\alpha]\varphi \to \odot[\alpha]([\alpha]\varphi)$	$(A4)$
$\odot[\alpha]\varphi \to \Diamond[\alpha]\varphi$	(Oic)
For $n \geq 1$ and pairwise different $\alpha_1, \ldots, \alpha_n$,	
$\bigwedge_{1\leq k\leq n} \Diamond[\alpha_i]\varphi_i \to \Diamond \left(\bigwedge_{1\leq k\leq n}[\alpha_i]\varphi_i\right)$	(IA)
$K_\alpha\varphi \to [\alpha]\varphi$	(OAC)
$\Diamond K_\alpha\varphi \to K_\alpha\Diamond\varphi$	$(Unif-H)$
$\odot_{\mathscr{S}}[\alpha](\varphi \to \psi) \to (\odot_{\mathscr{S}}[\alpha]\varphi \to \odot_{\mathscr{S}}[\alpha]\psi)$	$(A5)$
$\odot_{\mathscr{S}}[\alpha]\varphi \to \odot_{\mathscr{S}}[\alpha](K_\alpha\varphi)$	$(A6)$
$K_\alpha\Box\varphi \to \odot_{\mathscr{S}}[\alpha]\varphi$	$(s.N)$
$\odot_{\mathscr{S}}[\alpha]\varphi \to \Diamond K_\alpha\varphi$	$(s.Oic)$
$\odot_{\mathscr{S}}[\alpha]\varphi \to K_\alpha\Box \odot_{\mathscr{S}}[\alpha]\varphi$	$(s.Cl1)$
$\neg \odot_{\mathscr{S}}[\alpha]\varphi \to K_\alpha\Box\neg \odot_{\mathscr{S}}[\alpha]\varphi$	$(s.Cl2)$
$B_\alpha(\varphi \to \theta) \to (B_\alpha\varphi \to B_\alpha\theta)$	$(A7)$
$(\psi \leftrightarrow \varphi) \to (B_\alpha\theta \leftrightarrow B_\alpha^\varphi\theta)$	$(A8)$
$K_\alpha\varphi \to B_\alpha\varphi$	(PK)
$B_\alpha\varphi \to K_\alpha B_\alpha\varphi$	$(FIB1)$
$\neg B_\alpha\varphi \to K_\alpha\neg B_\alpha\varphi$	$(FIB2)$
$B_\alpha^\varphi\varphi$	(HA)
$B_\alpha\varphi \to (B_\alpha^{\varphi\wedge\psi}\theta \leftrightarrow B_\alpha\theta)$	$(MBR1)$
$\neg B_\alpha\neg\varphi \to (B_\alpha^{\varphi\wedge\psi}\theta \leftrightarrow B_\alpha(\varphi \to \theta))$	$(MBR2)$
$\psi \to \neg B_\alpha\bot$	$(WCon)$
$\odot_{\mathscr{B}}[\alpha](\varphi \to \psi) \to (\odot_{\mathscr{B}}[\alpha]\varphi \to \odot_{\mathscr{B}}[\alpha]\psi)$	$(A9)$
$\odot_{\mathscr{B}}[\alpha]\varphi \to \odot_{\mathscr{B}}[\alpha](K_\alpha\varphi)$	$(A10)$
$K_\alpha\Box\varphi \to \odot_{\mathscr{B}}[\alpha]\varphi$	$(d.N)$
$\odot_{\mathscr{B}}[\alpha]\varphi \to \Diamond K_\alpha\varphi$	$(d.Oic)$
$\odot_{\mathscr{B}}[\alpha]\varphi \to K_\alpha\Box \odot_{\mathscr{B}}[\alpha]\varphi$	$(d.Cl1)$
$\neg \odot_{\mathscr{B}}[\alpha]\varphi \to K_\alpha\Box\neg \odot_{\mathscr{B}}[\alpha]\varphi$	$(d.Cl2)$

- *(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 Λ 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.[8] The proof of completeness in the Appendix also shows that Λ 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.

[8]Nevertheless, one can adapt the correspondence between Kripke models and branching-time models from [40] and [1] to show that Λ 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`.

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 ψ is the case, then the doxastic obligations that such an agent had at the index should in principle be subject to the revision with ψ—just as beliefs are. Formulas of the form $\odot_{\mathscr{B}}[\alpha]^{\psi}\varphi$ 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 ψ 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 & Eunate 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.