A Stit Logic of Responsibility

Extended Abstract

Aldo Iván Ramírez Abarca
Utrecht University
Utrecht, The Netherlands
a.i.ramirezabarca@gmail.com

Jan Broersen
Utrecht University
Utrecht, The Netherlands
J.M.Broersen@uu.nl

ABSTRACT

We present a logic of responsibility. Extending stit theory with epistemic, doxastic, deontic, and intentional modalities, we provide logic-based characterizations of several modes of responsibility.

KEYWORDS

Knowledge Representation; Deontic Logic; Stit Logic

ACM Reference Format:
Aldo Iván Ramírez Abarca and Jan Broersen. 2022. A Stit Logic of Responsibility: Extended Abstract, Main Track. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

Our operational definition for responsibility is as follows: the relation between an agent and circumstances of an environment by which the agent can be either blamed or praised on account of its measure of involvement in the obtaining of those circumstances, all in the context of a particular normative system. Our analysis of this notion is based on Broersen's three *categories* of responsibility: *causal responsibility*, concerning the question "who is the material author of a circumstance in the environment?"; *informational responsibility*, concerning the question "did the author behave consciously while performing the action that brought about the circumstance?"; and *motivational responsibility*, concerning the question "did the author behave intentionally?" In order to reason about these categories, we propose a *decomposition* of responsibility into the following components: **agents**: the bearers of responsibility, the authors of actions; **actions**: the processes by which agents bring about changes or effects in the environment; **knowledge and beliefs**: mental attitudes that constitute explanations for agents' particular choices of action; **intentions**: the agentive states that determine whether an action was done with the purpose of bringing about the effects of that action or not; **ought-to-do's**: the actions that agents should perform, complying with some normative system according to which the agents can be either blamed or praised.

2 A LOGIC OF RESPONSIBILITY

The basic proposal is to extend *act-utilitarian stit theory* [9] to account for the components of responsibility mentioned above. Stit theory, where the acronym 'stit' stands for *seeing to it that*, was created to provide a logic-based account of agency. Its semantics lent itself very naturally to the study both of ought-to-do [9] and of knowledge-influenced action [5, 7, 8]. Here, we add belief, belief-dependent ought-to-do's, and intentions to the mix.

Definition 2.1 (Syntax). 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 $\mathcal{L}_{\mathsf{R}}$ is given by:

$$\varphi := \quad p \mid \neg\varphi \mid \varphi \wedge \psi \mid \Box\varphi \mid [\alpha]\varphi \mid K_\alpha\varphi \mid B_\alpha\varphi \mid I_\alpha\varphi \mid \odot_\alpha\varphi \mid \odot_\alpha^{\mathcal{S}}\varphi \mid \odot_\alpha^{\mathcal{B}}\varphi.$$

$\Box\varphi$ is meant to express the 'historical necessity' of φ, and $[\alpha]\varphi$ stands for 'agent α has seen to it that φ.' These are the basic stit modalities. $K_\alpha\varphi$ stands for 'α knows that φ holds.' $B_\alpha\varphi$ stands for 'α believes that φ holds.' $I_\alpha\varphi$ stands for 'α has an intention to realize φ.' $\odot_\alpha\varphi$ expresses that α objectively ought to have seen to it that φ. $\odot_\alpha^{\mathcal{S}}\varphi$ expresses that α subjectively ought to have seen to it that φ. Finally, $\odot_\alpha^{\mathcal{B}}\varphi$ expresses that α doxastically ought to have seen to it that φ. As for the semantics, the structures on which the formulas of $\mathcal{L}_{\mathsf{R}}$ are evaluated are based on what we call *knowledge-belief-intentions-oughts branching-time frames*.

Definition 2.2 (Frames). A tuple of the form $\langle M, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha \in Ags}, \{\mu_\alpha\}_{\alpha \in Ags}, \tau, \mathbf{Value}\rangle$ is a **finite** knowledge-beliefs-intentions-oughts branching-time frame (*kbiobt*-frame for short) iff $\langle M, \sqsubset, Ags, \mathbf{Choice}, \mathbf{Value}\rangle$ is an act-utilitarian stit frame just as defined in [9], with a finite domain. For $\alpha \in Ags$, $\sim_\alpha$ is the epistemic indistinguishability equivalence relation for α. We define two notions of α's *information set* at $\langle m, h\rangle$: the set $\pi_\alpha^\Box[\langle m, h\rangle] := \{\langle m', h'\rangle; \exists h'' \in H_{m'} s.t. \langle m, h\rangle \sim_\alpha \langle m', h''\rangle\}$ is α's *ex ante information set*; the set $\pi_\alpha[\langle m, h\rangle] := \{\langle m', h'\rangle; \langle m, h\rangle \sim_\alpha \langle m', h'\rangle\}$ is α's *ex interim information set.* For $\alpha \in Ags$, $\mu_\alpha : \mathcal{P}(M \times H) \to [0, 1]$ is a classical discrete probability function such that, for index $\langle m, h\rangle$, $\mu_\alpha(\pi_\alpha[\langle m, h\rangle]) > 0$. τ is a function that assigns to each $\alpha \in Ags$ and index $\langle m, h\rangle$ a topology $\tau_\alpha^{\langle m,h\rangle} \subseteq \mathcal{P}(\pi_\alpha^\Box[\langle m, h\rangle])$. This is the topology of α's *intentions* at $\langle m, h\rangle$. For $\alpha \in Ags$, τ_α must satisfy two conditions: (a) if $\pi_\alpha^\Box[\langle m, h\rangle] = \pi_\alpha^\Box[\langle m', h'\rangle]$, then $\tau_\alpha^{\langle m,h\rangle} = \tau_\alpha^{\langle m',h'\rangle}$; and (b) for every non-empty $U, V \in \tau_\alpha^{\langle m,h\rangle}$, $U \cap V \neq \emptyset$. In other words, every non-empty U is dense. **Value** is a deontic function that assigns to each history h a real number, representing the deontic utility of h.

The equivalence relations $\sim_\alpha$ are the usual indistinguishability relations, borrowed from epistemic logic. The probability functions underlie the semantics for a probability-1 version of belief [3, 4]. The function τ implies a novel proposal for the concept of *intentional action* [see 6]. We adopt the ideas of evidential epistemic logic [2, 10] and associate a topology to each agent at each index. The open sets of any such topology are intentions of that agent for bringing about circumstances. The fact that the non-empty open sets of the topologies are dense implies that an agent's intentions are consistent. Regarding the deontic dimension, the idea is that

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online. © 2022 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

objective, subjective, and doxastic ought-to-do's stem from the optimal actions for an agent: to have seen to it that φ is an obligation of an agent at an index iff φ is an effect of all the optimal actions for that agent and index, where the notion of optimality is based on the deontic value of the histories in those actions—provided by **Value**. At moment m, the set of α's objectively optimal actions is denoted by $\mathbf{Optimal}_{\alpha}^{m}$, and the set of α's subjectively optimal actions is denoted by $\mathbf{SOptimal}_{\alpha}^{m}$. The reader is referred to [1] for the formal definitions of these sets. Here, we identify an agent's *doxastic* sense of ought-to-do with the effects of actions that maximize expected (deontic) utility. For α's action L, α's *expected deontic utility* of L at $\langle m, h\rangle$—denoted by $EU_{\alpha}^{\langle m,h\rangle}(L)$—is defined as $EU_{\alpha}^{\langle m,h\rangle}(L) := \sum_{m' \sim_{\alpha} m, h' \in [L]_{\alpha}^{m'}} \mu_{\alpha}\left(\{h'\} \mid \pi_{\alpha}[\langle m', h'\rangle]\right) \cdot \mathbf{Value}(h')$. Since *kbiobt*-frames are finite, there are actions that maximize α's expected deontic utility at every index, and $\mathbf{EU}_{\alpha}^{\langle m,h\rangle}$ denotes the set of such actions.

Definition 2.3 (Models and Evaluation rules). A *kbiobt*-model $\mathcal{M}$ consists of the tuple that results from adding a valuation function $\mathcal{V}$ to a *kbiobt*-frame, where $\mathcal{V} : P \longrightarrow \mathcal{P}(M \times H)$ assigns to each atomic proposition a set of indices. The novel truth conditions are as follows—for the semantics of the basic stit modalities, the reader is referred to [9]:

$$
\begin{array}{lll}
\mathcal{M}, \langle m, h\rangle \models K_{\alpha}\varphi & \text{iff} & \text{for each } \langle m', h'\rangle \text{ s.t. } \langle m, h\rangle \sim_{\alpha} \langle m', h'\rangle, \\
 & & \mathcal{M}, \langle m', h'\rangle \models \varphi \\
\mathcal{M}, \langle m, h\rangle \models B_{\alpha}\varphi & \text{iff} & \mu_{\alpha}\left(\|\varphi\| \mid \pi_{\alpha}[\langle m, h\rangle]\right) = 1 \\
\mathcal{M}, \langle m, h\rangle \models I_{\alpha}\varphi & \text{iff} & \text{there is } U \in \tau_{\alpha}^{\langle m,h\rangle} \text{ s.t. for all} \\
 & & \langle m', h'\rangle \in U, \mathcal{M}, \langle m', h'\rangle \models \varphi \\
\mathcal{M}, \langle m, h\rangle \models \odot_{\alpha}\varphi & \text{iff} & \text{for each } L \in \mathbf{Optimal}_{\alpha}^{m}, h' \in L \text{ implies that} \\
 & & \mathcal{M}, \langle m, h'\rangle \models \varphi \\
\mathcal{M}, \langle m, h\rangle \models \odot_{\alpha}^{\mathcal{S}}\varphi & \text{iff} & \text{for each } L \in \mathbf{SOptimal}_{\alpha}^{m}, \text{ for each } m' \text{ s.t.} \\
 & & m \sim_{\alpha} m', h' \in [L]_{\alpha}^{m'} \text{ implies that} \\
 & & \mathcal{M}, \langle m, h'\rangle \models \varphi \\
\mathcal{M}, \langle m, h\rangle \models \odot_{\alpha}^{\mathcal{B}}\varphi & \text{iff} & \text{for each } L \in \mathbf{EU}_{\alpha}^{\langle m,h\rangle}, \text{ for each } m' \text{ s.t. } m \sim_{\alpha} m', \\
 & & h' \in [L]_{\alpha}^{m'} \text{ implies that } \mathcal{M}, \langle m, h'\rangle \models \varphi.
\end{array}
$$

where $\|\varphi\|$ is the set $\{\langle m, h\rangle \in M \times H; \mathcal{M}, \langle m, h\rangle \models \varphi\}$ and $[L]_{\alpha}^{m} := \{h \in H_m; \exists h_* \in L \text{ s.t. } \langle m_*, h_*\rangle \sim_{\alpha} \langle m, h\rangle\}$.

3 FORMALIZATION OF MODES OF RESPONSIBILITY

We formalize particular modes of responsibility as formulas of $\mathcal{L}_{\mathsf{R}}$. These modes are sub-categories of Broersen's three categories of responsibility that respectively correspond to the active and passive forms of the notion. The active form concerns contributions, and the passive form concerns omissions. We first introduce the characterizations of these sub-categories (Table 1), and then analyze them with respect to the blame-and-praise assignation that is implied by the deontic context of our logic.

	Active form (contributions)	Passive form (omissions)
Causal responsibility	$[\alpha]\varphi$	$\varphi \land \neg[\alpha]\neg\varphi$
Informational responsibility	$K_{\alpha}[\alpha]\varphi$	$\varphi \land K_{\alpha}\neg[\alpha]\neg\varphi$
Motivational responsibility	$I_{\alpha}[\alpha]\varphi$	$\varphi \land I_{\alpha}\neg[\alpha]\neg\varphi$

Table 1: Basic modes

Therefore, α was causal-active responsible for φ iff α saw to it that φ was the case (causal contribution), and α was causal-passive responsible for φ iff φ was the case and α refrained from preventing this (causal omission). α was informational-active responsible for φ iff α knowingly saw to it that φ (conscious contribution), and α was informational-passive responsible for φ iff φ was the case and α knowingly refrained from preventing this. Finally, α was motivational-active responsible for φ iff α intentionally saw to it that φ (intentional contribution), and α was motivational-passive responsible for φ iff φ was the case and α intentionally refrained from preventing this (intentional omission). Observe that, for all three categories, being active-responsible for φ logically implies being passive-responsible for φ. An important variation of the category of informational responsibility results from using the belief modalities in place of the knowledge ones.

The deontic attributes of our models offer a criterion for deciding when agents could be blamed and when agents could be praised. Specific conjunctions of these deontic modalities, on one hand, and the formulas characterizing the basic modes, on the other, yield nuanced degrees of responsibility. For each φ, there are 8 possible combinations for conjunctions of deontic modalities, according to whether $\Delta\varphi$ or $\neg\Delta\varphi$ holds ($\Delta \in \{\odot_{\alpha}, \odot_{\alpha}^{\mathcal{S}}, \odot_{\alpha}^{\mathcal{B}}\}$). Let us illustrate the interplay between the basic modes and the deontic modalities. For a formula φ, assume that $\odot_{\alpha}\varphi \land \odot_{\alpha}^{\mathcal{S}}\varphi \land \odot_{\alpha}^{\mathcal{B}}\varphi$ holds. Then there are the following degrees of praiseworthiness and blameworthiness: **degrees of praiseworthiness**: *lowest* – $[\alpha]\varphi \land \neg K_{\alpha}[\alpha]\varphi \land \neg I_{\alpha}[\alpha]\varphi$ holds (α is causal-active responsible for φ, but α is not aware of having brought about φ and did not intentionally brought about φ); *low* – $[\alpha]\varphi \land \neg K_{\alpha}[\alpha]\varphi \land I_{\alpha}[\alpha]\varphi$ holds (α is causal-active and motivational-active responsible for φ, but α is not aware of having brought about φ); *middle* – $K_{\alpha}[\alpha]\varphi \land \neg I_{\alpha}[\alpha]\varphi$ holds (α is causal-active and informational-active responsible for φ, but did not intentionally brought about φ); *highest* – $K_{\alpha}[\alpha]\varphi \land I_{\alpha}[\alpha]\varphi$ holds (α is causal-active, informational-active, and motivational-active responsible for φ); **Degrees of blameworthiness** *lowest* – $\neg\varphi \land \neg[\alpha]\varphi \land \neg K_{\alpha}\neg[\alpha]\varphi \land \neg I_{\alpha}\neg[\alpha]\varphi$ holds (α is causal-passive responsible for $\neg\varphi$, but α did not knowingly refrain from having brought about φ and did not intentionally refrain from having brought about φ); *low* – $\neg\varphi \land [\alpha]\varphi \land \neg K_{\alpha}\neg[\alpha]\varphi \land I_{\alpha}\neg[\alpha]\varphi$ holds (α is causal-passive and motivational-passive responsible for φ, but α is not aware of having refrained from bringing about φ); *middle* – $\neg\varphi \land K_{\alpha}\neg[\alpha]\varphi \land \neg I_{\alpha}\neg[\alpha]\varphi$ holds (α is causal-passive and informational-passive responsible for $\neg\varphi$, but α did not intentionally refrain from having brought about φ); *highest* – $\neg\varphi \land K_{\alpha}\neg[\alpha]\varphi \land I_{\alpha}\neg[\alpha]\varphi$ holds (α is causal-passive, informational-passive, and motivational-passive responsible for $\neg\varphi$).

With the exception of the case when $\neg\odot_{\alpha}\varphi \land \neg\odot_{\alpha}^{\mathcal{S}}\varphi \land \neg\odot_{\alpha}^{\mathcal{B}}\varphi$ holds (where to have brought about φ or to refrain from doing so elicits a neutral response), in all the cases given by the possible conjunctions of deontic modalities the same degrees of praiseworthiness and blameworthiness apply. However, the particular combinations of deontic modalities yield situations for which praise or blame could increase or decrease.

REFERENCES

[1] Aldo Iván Ramírez Abarca and Jan Broersen. 2019. A Logic of Objective and Subjective Oughts. In *European Conference on Logics in Artificial Intelligence*. Springer, 629–641.

[2] Alexandru Baltag, Nick Bezhanishvili, Aybüke Özgün, and Sonja Smets. 2016. Justified belief and the topology of evidence. In *International Workshop on Logic,*

Language, Information, and Computation. Springer, 83–103.

[3] Alexandru Baltag and Sonja Smets. 2008. Probabilistic dynamic belief revision. *Synthese* 165, 2 (2008), 179.

[4] Adam Bjorndahl, Joseph Y Halpern, and Rafael Pass. 2017. Reasoning about rationality. *Games and Economic Behavior* 104 (2017), 146–164.

[5] Jan Broersen. 2011. Deontic epistemic stit logic distinguishing modes of mens rea. *Journal of Applied Logic* 9, 2 (2011), 137–152.

[6] Jan M Broersen. 2011. Making a start with the stit logic analysis of intentional action. *Journal of philosophical logic* 40, 4 (2011), 499–530.

[7] Andreas Herzig and Nicolas Troquard. 2006. Knowing how to play: uniform choices in logics of agency. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*. ACM, 209–216.

[8] John Horty and Eric Pacuit. 2017. Action types in STIT semantics. *The Review of Symbolic Logic* 10, 4 (2017), 617–637.

[9] John F. Horty. 2001. *Agency and Deontic Logic*. Oxford University Press.

[10] Johan van Benthem and Eric Pacuit. 2011. Dynamic logics of evidence-based beliefs. *Studia Logica* 99, 1-3 (2011), 61–92.