A Logic of Evaluation

Emiliano Lorini

A Logic of Evaluation

Emiliano Lorini
IRIT-CNRS, Toulouse University
Toulouse, France
Emiliano.Lorini@irit.fr

ABSTRACT

We present a logic of evaluation which clarifies the relationship between knowledge, values and preferences of multiple agents in an interactive setting. Evaluation is a fundamental concept for understanding how an ethical agent's decision is affected by her values. We provide a complete axiomatics for the logic and present a dynamic extension by the concept of value expansion. We show that value expansion indirectly affects the agents' preferences by inducing a preference upgrade operation.

ACM Reference Format:
Emiliano Lorini. 2021. A Logic of Evaluation. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 9 pages.

1 INTRODUCTION

An autonomous agent is, by definition, endowed with endogenous motivations, commonly called goals, which determine her preferences, thereby indirectly influencing her decision-making process. Evaluation is a core aspect of agent autonomy. It is the bridge between the agent's goals and preferences through the lens of the agent's knowledge (or beliefs). This aspect is emphasized in [29] in which an agent's evaluation of a situation (outcome, state, etc.) is defined to be a belief of an evaluating agent about the goodness (or usefulness) of the situation with regard to the agent's goals.

The concept of evaluation is paramount for psychological theories of action [20] and emotion [30] as well as for cognitive theories of knowledge and beliefs [1]. It is also highly relevant for ethical multiagent systems and, more generally, for machine ethics, one of the central areas of AI nowadays [3, 18, 42]. Indeed, as pointed out by [17], for an autonomous agent to be ethical and to behave responsibly, some of her goals must reflect values and norms with which she is expected to comply and which take other agents and their welfare into consideration. This includes both abstract values such as justice, fairness, reciprocity and equity and more concrete ones such as "greenhouse gas emissions are reduced" and "social distancing measures are adopted for fighting against covid-19".[1] A typical example of ethical autonomous agent is a robot whose set of values includes the respect for human integrity [43].

In order to supply her expected functionality, an ethical agent should be capable of computing her preference ordering over the alternatives directly from her values and then use it, together with her knowledge and belief, as input of her decision-making process. This is what Figure 1 highlights. Specifically, by evaluating how

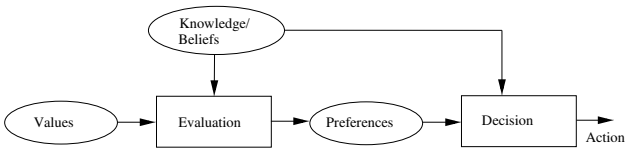

Figure 1: From values to preferences via evaluation

good a situation is, the ethical agent makes an epistemic judgment about whether and how much the situation promotes the achievement of her values. Some evaluations are comparative in the sense that the agent's assessment of the goodness of a certain situation is made in relation to the goodness of another situation, depending on which and how many values are satisfied in each of them.[2]

This view of evaluation is in line with the philosophical doctrine of pluralistic consequentialism [35, 36] and with recent theories of reason-based choice [15, 16] according to which an ethical agent has to weigh different, and sometimes conflicting, values and properties to assess the relative goodness of a given option or alternative in her choice set. It is also in accord with existing computational models of ethical deliberation and planning in robotics [4, 41] and in AI [2, 12, 33, 37] in which the influence of values on decision and the evaluative component are made explicit.

The aim of this paper is to introduce a logic of evaluation which helps to clarify the relationship between knowledge, values and preferences of multiple ethical agents in a multi-agent setting. The interest of having a multi-agent account of values and evaluation lies in the possibility of modeling interactive situations in which: (i) an agent's value may concern other agents' well-being, safety and integrity, and (ii) agents' decisions are interdependent so that the possibility for an agent to achieve her values may depend on what other agents decide to do. The development of a language with a suitable syntax and semantics for specifying value preferences of ethical agents in the context of sociotechnical systems (STSs) was recently put forward by Murukannaiah et al. [31] as a research challenge. In this paper, we make a first step to meet this challenge. We are not the first to propose a logical analysis of ethical reasoning. There are approaches based on preference logic [21], event calculus (ASP) [9], temporal-epistemic logic [28], BDI (belief, desire, intention) agent language [13] and classical higher-order logic (HOL) [8]. Nonetheless, none of them has taken the concept of evaluation into account. The latter, we believe, is crucial for elucidating the connection between values and preferences.

[1] According to contemporary theories of human motivation in philosophy [34] and in economics [22], goals of a rational agent may originate either (i) from somatically-marked motivations such as desires or physiological needs, or (ii) from ethical considerations, moral values and norms. In other words, there are desire-based goals and value-based goals. In this paper, we will focus exclusively on the latter category.

[2] This aspect is emphasized by Dewey in his theory of valuing, where he defines evaluation as "...an activity of rating, an act that involves comparison..." [14, p. 5].

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

The paper is organized as follows. In Section 2, we discuss the conceptual background with special emphasis on the notion of value. In Section 3, we present the language and the semantics of our logic of evaluation, while in Section 4 we concentrate on the notion of choice, in connection with knowledge and value, and show how it can be expressed in the language. Section 5 is devoted to present complete axiomatics for our logic. In Section 6, we move from a static to a dynamic perspective by extending our logic with value expansion operators. We show that value expansion indirectly affects an agent's preferences by inducing a preference upgrade operation in the sense of [40]. A selection of abbreviated proofs is given in the technical annex at the end of the paper.

2 CONCEPTUAL BACKGROUND

We conceive evaluation as the computation of a preference ordering over a set of epistemic alternatives from a set of actual values. The evaluating agent determines whether a situation that she considers possible is at least as good as another situation that she considers possible in the light of which — and eventually how many — of her values are satisfied in each situation. The evaluation criterion we study in the paper is purely qualitative. We assume that an agent will consider her epistemic alternative v at least as good as her epistemic alternative w if and only if the set of her values satisfied at w is included in the set of her values satisfied at v.

Evaluation is also leveraged by the agent to identify ideal situations in her set of epistemic alternatives, namely, those alternatives at which all her values are satisfied. Clearly, if the agent has conflicting values, i.e., values that cannot be concomitantly satisfied, then the set of ideal situations is empty since there is no epistemic alternative at which two conflicting values are satisfied. In this case, the agent can only identify subideal situations, i.e., situations at which only some of her values but not all of them are satisfied.[3]

Our logic of evaluation is grounded on the three primitive concepts of knowledge, value and preference, in accordance with the conceptual framework depicted in Figure 1. They will be reflected in the language of the logic by corresponding modal operators of the form K_i (knowledge), V_i (value) and $[\preceq_i]$ (preference), where i is an agent identifier. The operator K_i captures the usual S5 notion of knowledge for logically omniscient and fully introspective agents [19]. The operator $[\preceq_i]$ is a standard S4 betterness modality [40] reflecting a partial order over agent i's set of epistemic alternatives. The operator V_i captures a logically weak notion of value whose only positive property is closure under logical equivalence. We see the latter as the minimal property for values. This assumption is justified by the fact that the notion of value we study is normative (in opposition to descriptive), in the sense that it pertains to a rational resource-unbounded agent who is capable of immediately detaching a new value that ψ from her value that φ and her knowledge that φ and ψ have the same extension.

The reason why the value operator is parameterized by an agent identifier i is that values are assumed to be agent-relative. This does not conflict with the objectivist view of morality, namely the view that all agents start from the same set of ultimate ethical principles which determine the goodness or rightness of any state of affairs or action.[4] The variability of values between two agents may stem from the difference between their epistemic states, despite the fact that the two agents share the same set of ultimate ethical principles. For example, suppose Ann and Bob share the ultimate ethical principle that "unfair acts are deplorable". Nonetheless, they have different interpretations of what 'fairness' means, given their different epistemic states. For instance, Ann is a fervent utilitarian, while Bob is Rawlsian: according to Ann an act is fair if and only if it is aimed at maximizing collective utility while, according to Bob, an act is fair if and only if it is aimed at maximizing utility of the most disadvantaged person. Consequently, Ann and Bob have divergent values about implementation of fairness.

The first negative property of values we want to discuss is that they are not necessarily closed under conjunction. Suppose Mary is organizing a party. Her two dear friends Ann and Bob have recently split up and are actually in a bad relationship. According to Mary, inviting Ann and inviting Bob are both valuable options, since it would be unfair to exclude one of her dearest friends from the party. Nonetheless, if Ann and Bob meet at the party, they will find themselves in an embarrassing situation which may negatively affect the other guests. For this reason, according to Mary, inviting both Ann and Bob to the party is not a valuable option. This example shows that an agent can have conflicting values, without having the value of complying with both of them. In formal terms, it is reasonable to assume that the following formula should be satisfiable:

$$\mathsf{V}_i\varphi \wedge \mathsf{V}_i\psi \wedge \neg\mathsf{V}_i(\varphi \wedge \psi).$$

A further negative property of values is that they are not necessarily closed under disjunction. Suppose Ann is a doctor working at the hospital intensive care unit (ICU). A first patient in a critical situation arrives at the unit. Saving her/his life is a value for Ann, since her commitment is to save people's lives. Few minutes later a second patient is taken to the unit. Saving the second patient's life is also a value for Ann. In this situation, Ann does not have the disjunctive value to save the first's patient life or to save the second's patient life. Indeed, she is motivated to save the lives of both patients. This illustrates that an agent can have the value that φ and the value that ψ, — eventually coupled with the conjunctive value that $\varphi \wedge \psi$ —, without having the disjunctive value that $\varphi \vee \psi$. In other words, the following formula should be satisfiable:

$$\mathsf{V}_i\varphi \wedge \mathsf{V}_i\psi \wedge \neg\mathsf{V}_i(\varphi \vee \psi).$$

Note that the satisfiability of the previous formula should be independent from the fact that agent i knows that φ and ψ cannot occur together. For example, Ann can still want to save the lives of both patients, notwithstanding her knowledge that she is unable to do so (e.g., because the ICU is not equipped with two automatic ventilators for covid-19 treatment).

Values do not satisfy weakening either. Specifically, the fact that an agent has the conjunctive value that φ and ψ does not necessarily imply that she has the value that φ. In other words, it is reasonable to assume that the following formula should be satisfiable:

$$\mathsf{V}_i(\varphi \wedge \psi) \wedge \neg\mathsf{V}_i\varphi.$$

[3] This situation is typical of moral struggles, as defined by Levi, which "...are provoked by inconsistencies between value commitments and information concerning the kinds of decision problems which arise..." [25, p. 8].

[4] See [32, 35] for further discussion about the compability between moral objectivism and agent-relativity of ethical values.

For example, suppose Mary and Bob are discussing about the actions for schools during the covid-19 outbreak. According to Mary, children should return to school (φ) with application of social distancing and the use of masks (ψ). This does not imply that, according to Mary, children should return to school regardless of the fact that social distancing is applied and masks are used.

More generally, since values do not satisfy closure under disjunction or weakening, they are not closed under logical consequence. This means that the fact that ψ is a logical consequence of φ and an agent considers φ a valuable state of affairs does not necessarily imply that the agent considers ψ a valuable state of affairs.

We conclude this section with a remark on the concept of realism for values. The question is whether a rational agent can have values which are incompatible with her knowledge and beliefs. For example, may a rational agent be a fervent pacifist, notwithstanding her knowledge that aggressivity is a natural disposition of human nature and human conflict is unavoidable? In this paper, we study two different logics: the logic of (possibly non-realistic) values and the logic of realistic values. The latter logic assumes that rational agents instantaneously discard non-realistic values. The former is less demanding, as it is prone to accept that an agent may have utopian values in her mind and simply filter them out at a subsequent stage. This is in line with [11], according to whom non-realistic goals and values are not actively pursued and are not taken into consideration during the evaluation process, thereby having no influence on preference formation and decision-making.

3 LOGICAL FRAMEWORK

In this section, we first present a multi-agent language for describing the relationship between values, knowledge and preferences of multiple agents. Then, we introduce the notion of evaluative model and show how to interpret our language relative to it.

Let Atm be a countable infinite set of atomic propositions and let Agt be a finite set of agents. We define the language $\mathcal{L}$ by the following grammar:

$$\varphi \quad ::= \quad p \mid \mathsf{id}_i \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \mathsf{K}_i\varphi \mid \mathsf{V}_i\varphi \mid [\preceq_i]\varphi,$$

where p ranges over Atm and i ranges over Agt.

The formula $\mathsf{K}_i\varphi$ has to be read "agent i knows that φ", while $\mathsf{V}_i\varphi$ has to be read "agent i considers φ a valuable state of affairs" or, simply, "agent i has the value that φ". The formula $[\preceq_i]\varphi$ has to be read "φ is true at all states that according to agent i are at least as good as the current one". Finally, the special atomic formula id_i is meant to stand for "the actual world is an ideal world for agent i". The dual of the epistemic operator K_i and of the preference operator $[\preceq_i]$ are defined as usual: $\widehat{\mathsf{K}}_i\varphi =_{def} \neg\mathsf{K}_i\neg\varphi$ and $\langle\preceq_i\rangle\varphi =_{def} \neg[\preceq_i]\neg\varphi$. The extended set of atomic formulas Atm^+ is defined as follows:

$$Atm^+ = Atm \cup \bigcup_{i\in Agt} \{\mathsf{id}_i\}.$$

For notational convenience, elements of Atm^+ are noted $x, y, \ldots$

The following definition introduces evaluative models.

Definition 3.1 (Evaluative model). An evaluative model (EM) is a tuple $M = (W, (\equiv_i)_{i\in Agt}, N, (\preceq_i)_{i\in Agt}, V)$ where:

- W is a non-empty set of worlds or states,
- $\equiv_i$ is an equivalence relation on W,
- $N : Agt \times W \longrightarrow 2^{2^W}$ is a neighbourhood function,
- $\preceq_i$ is a partial preorder on W,
- $V : W \longrightarrow 2^{Atm^+}$ is a valuation function,

and that satisfies the following constraints, for every $w, v \in W$, $X \subseteq W$ and $i \in Agt$:

(C1) $\preceq_i \subseteq \equiv_i$,
(C2) if $X \in N(i, w)$ then $X \subseteq \equiv_i(w)$,
(C3) if $w \equiv_i v$ then $N(i, w) = N(i, v)$,
(C4) $w \preceq_i v$ if and only if $Sat_M(i, w) \subseteq Sat_M(i, v)$,
(C5) $\mathsf{id}_i \in V(w)$ if and only if $Sat_M(i, w) = N(i, w)$,

with $\equiv_i(w) = \{v \in W : w \equiv_i v\}$ and $Sat_M(i, w) = \{X \in N(i, w) : w \in X\}$. The class of EMs is denoted by **M**.

The equivalence relation $\equiv_i$ is an epistemic indistinguishability relation. In particular, for each $w \in W$, $\equiv_i(w)$ is agent i's information set at world w. The neighbourhood function specifies the agents' values at each world. Specifically, if $X \in N(i, w)$ then X is a value of agent i at world w, i.e., at w agent i considers X a valuable (or ideal) state of affairs. The relation $\preceq_i$ specifies a preference ordering for agent i over the possible worlds. $Sat_M(i, w)$ is the set of agent i's values that are satisfied at w. According to Constraint C1, an agent's preference ordering is relative to worlds in her information set. Similarly, according to Constraint C2, an agent's value is relative to worlds in her information set. Constraint C3 captures introspection for values, i.e., an agent's set of values should be the same in all worlds in the agent's information set. Constraints C4 and C5 are the central properties of evaluation. They are used to compute an agent's preference ordering and set of ideal worlds from her set of values. According to C4, world v is for agent i at least as good as world w if and only if the set of i's values satisfied at w is included in the set of i's values satisfied at v.[5] According to C5, a world is ideal for an agent if and only if it satisfies all her values. The set of worlds that agent i considers ideal at world w of model M (*aka* i's set of subjectively ideal worlds at w of M) is:

$$I_M(i, w) = \big\{v \in \equiv_i(w) : \mathsf{id}_i \in V(v)\big\}.$$

Since $\equiv_i$ is an equivalence relation, if $w \equiv_i v$ then $I_M(i, w) = I_M(i, v)$. Moreover, thanks to Constraints C3 and C5, if $v \in I(i, w)$ then $v \in \bigcap_{X\in N(i,w)} X$. This means that an agent's subjectively ideal world is a world that satisfies all actual values of the agent.

We introduce the usual notation for strict preference, indifference and incomparability:

$$\begin{aligned}
\prec_i &= \{(w, v) \in W \times W : w \preceq_i v \text{ and } v \not\preceq_i w\},\\
\succeq_i &= \{(w, v) \in W \times W : v \preceq_i w\},\\
\succ_i &= \{(w, v) \in W \times W : w \succeq_i v \text{ and } v \not\succeq_i w\},\\
\sim_i &= \{(w, v) \in W \times W : w \preceq_i v \text{ and } v \preceq_i w\},\\
\|_i &= \{(w, v) \in W \times W : w \not\preceq_i v \text{ and } v \not\preceq_i w\}.
\end{aligned}$$

The following definition precisely defines the realism condition for values we briefly discussed in Section 2.

Definition 3.2 (Value realism). Let $M = (W, (\equiv_i)_{i\in Agt}, N, (\preceq_i)_{i\in Agt}, V)$ be an EM. We say that M satisfies value realism (VR)

[5] A similar idea of computing a preorder from a neighbourhood structure or a priority graph can be found in evidence logic [38, 39] and in the logic of preference [26].

if and only if, for every $w \in W$, $X \subseteq W$ and $i \in Agt$,

$$\text{if } X \in N(i,w) \text{ then } \equiv_i(w) \cap X \neq \emptyset.$$

As we pointed out in Section 2, an agent can have conflicting values. The following definition restricts to the limit case in which an agent's values are globally consistent, i.e., there exists at least one subjectively ideal world for the agent.

Definition 3.3 (Value consistency). Let $M = (W, (\equiv_i)_{i\in Agt}, N, (\preceq_i)_{i\in Agt}, V)$ be an EM. We say that M satisfies value consistency (VC) if and only if, for every $w \in W$ and $i \in Agt$, we have $I_M(i,w) \neq \emptyset$.

Formulas of the language $\mathcal{L}$ are interpreted with respect a pointed evaluative model, i.e., an evaluative model and a world in it.

Definition 3.4 (Satisfaction relation). Let $M = (W, (\equiv_i)_{i\in Agt}, N, (\preceq_i)_{i\in Agt}, V)$ be an EM and let $w \in W$. Then:

$$\begin{aligned}
M, w \models x &\iff x \in V(w),\\
M, w \models \neg\varphi &\iff M, w \not\models \varphi,\\
M, w \models \varphi \wedge \psi &\iff M, w \models \varphi \text{ and } M, w \models \psi,\\
M, w \models \mathsf{K}_i\varphi &\iff \forall v \in W : \text{if } w \equiv_i v \text{ then } M, v \models \varphi,\\
M, w \models \mathsf{V}_i\varphi &\iff ||\varphi||^M_{i,w} \in N(i,w),\\
M, w \models [\preceq_i]\varphi &\iff \forall v \in W : \text{if } w \preceq_i v \text{ then } M, v \models \varphi,
\end{aligned}$$

with $||\varphi||^M_{i,w} = \{v \in W : M, v \models \varphi\} \cap \equiv_i(w)$.

Note that the interpretation of agent i's value operator V_i is relative to i's subjective truth set for φ (i.e., $||\varphi||^M_{i,w}$), namely, the worlds in i's information set in which φ is true.[6]

For every $X \subseteq \{VR, VC\}$, we denote by $\mathbf{M}_X$ the class of EMs satisfying every property in X, where *VR* and *VC* are, respectively, the value realism and value consistency condition of Definitions 3.2 and 3.3. Since value consistency implies value realism, we have $\mathbf{M}_{VC} \subseteq \mathbf{M}_{VR}$. Clearly, $\mathbf{M}_\emptyset = \mathbf{M}$. For each model class $\mathbf{M}_X$, notions of validity and satisfiability for formulas in $\mathcal{L}$ relative to $\mathbf{M}_X$ are defined in the usual way. We write $\models_{\mathbf{M}_X} \varphi$ to denote the fact that φ is valid relative to the class $\mathbf{M}_X$.

4 CHOICE OPERATOR

In this section we define a rational choice operator and study its relationship with the notions of value and knowledge. As a preliminary step towards this definition, we introduce the following notion of best alternative as a world in the agent's information set which is either at least as good or incomparable with any other world in the agent's information set.

Definition 4.1 (Best alternatives). Let $M = (W, (\equiv_i)_{i\in Agt}, N, (\preceq_i)_{i\in Agt}, V)$ be an EM, let $i \in Agt$ and let $w \in W$. Agent i's set of best alternatives at w is defined as follows:

$$Best(i,w) = \big\{v \in \equiv_i(w) : \forall u \in \equiv_i(w), u \preceq_i v \text{ or } u \|_i v\big\}.$$

A candidate choice for agent i is a $\sim_i$-equivalence class relative to agent i's set of best alternatives.

Definition 4.2 (Candidate choices). Let $M = (W, (\equiv_i)_{i\in Agt}, N, (\preceq_i)_{i\in Agt}, V)$ be an EM, let $i \in Agt$ and let $w \in W$. Agent i's set of candidate choices at w, denoted by $Choice(i,w)$, is the partition of the set $Best(i,w)$ induced by the indifference relation $\sim_i$.

[6] A similar interpretation for the notion of explicit belief is used in [5].

Intuitively, a candidate choice is a state of affairs that it would be rational for an agent to try to attain.

The previous notion of candidate choice is syntactically expressed through the following abbrevation:

$$\mathsf{C}_i\varphi =_{def} \widehat{\mathsf{K}}_i(\varphi \wedge [\preceq_i]\varphi),$$

where $\mathsf{C}_i\varphi$ has to be read "agent i can rationally choose that φ".

As the following proposition highlights, the operator C_i correctly represents the idea that an agent can rationally choose that φ if, from her point of view, φ is a consequence of one her candidate choices.

Proposition 4.3. *Let $M = (W, (\equiv_i)_{i\in Agt}, N, (\preceq_i)_{i\in Agt}, V)$ be an EM, let $i \in Agt$ and let $w \in W$. Then,*

$M, w \models \mathsf{C}_i\varphi$ *if and only if* $\exists X \in Choice(i,w)$ *such that* $X \subseteq ||\varphi||^M_{i,w}$.

It is worth noting that, under the assumption of value consistency, the choice operator C_i becomes normal. Indeed, we have the following validity for the class $\mathbf{M}_{VC}$:

$$\models_{\mathbf{M}_{VC}} \big(\mathsf{C}_i\varphi \wedge \mathsf{C}_i(\varphi \to \psi)\big) \to \mathsf{C}_i\psi \tag{1}$$

Moreover, the rule of necessitation is admissible for the class $\mathbf{M}$:

$$\text{If } \models_{\mathbf{M}} \varphi \text{ then } \models_{\mathbf{M}} \mathsf{C}_i\varphi \tag{2}$$

The reason for the normality of the operator C_i in $\mathbf{M}_{VC}$ is that the set of candidate choices $Choice(i,w)$ is a singleton in every model of this class. Specifically, $Choice(i,w)$ only includes the set of ideal worlds which satisfy all agent i's values at w. Therefore, under the value consistency assumption, an agent has exactly one candidate choice, hence the formula $\mathsf{C}_i\varphi$ can simply be read "agent i rationally chooses that φ". The following example illustrates the interrelation between the concepts of knowledge, value and choice.

Example 4.4. There are two mobile robots 1 and 2 who have to collect objects in a room and an obstacle obstructing the access to the battery charger station. Each robot can decide either to remove the obstacle or to exploit the other robot by letting it remove the obstacle. In order to formalize the example, we assume that the set Atm includes the following atomic propositions with their associated meanings: r_1 ("robot 1 removes the obstacle"), r_2 ("robot 2 removes the obstacle"), c_1 ("robot 1 charges its battery"), c_2 ("robot 2 charges its battery"), f ("the access to the battery charger station is free of obstacles"), e_1 ("robot e_1 is exploited") and e_2 ("robot 2 is exploited"). Each robot is motivated by two values, namely, the value of preserving its own well-functioning by keeping its battery charged and the value of refraining from exploiting the other robot. This is captured by the following abbreviation:

$$\varphi_1 =_{def} \mathsf{V}_1c_1 \wedge \mathsf{V}_2c_2 \wedge \mathsf{V}_1\neg e_2 \wedge \mathsf{V}_1\neg e_1.$$

Since each robot has no additional values, it knows that a situation in which it keeps its battery charged without exploiting the other robot is an ideal situation:

$$\varphi_2 =_{def} \mathsf{K}_1\big((c_1 \wedge \neg e_2) \to \mathsf{id}_1\big) \wedge \mathsf{K}_2\big((c_2 \wedge \neg e_1) \to \mathsf{id}_2\big).$$

Furthermore, each robot knows that the only way to charge a battery is by freeing the access to the battery charger station, and

that the only way to free the access to the battery charger station is by one of them removing the obstacle:

$$\varphi_3 =_{def} \bigwedge_{i\in\{1,2\}} \Big(\mathsf{K}_i\big((c_1 \vee c_2) \to f\big) \wedge \mathsf{K}_i\big(f \to (r_1 \vee r_2)\big)\Big).$$

Finally, each robot knows that if one removes the obstactle while the other does not do it, then this counts as an act of exploitation:

$$\varphi_4 =_{def} \bigwedge_{i\in\{1,2\}} \mathsf{K}_i\Big(\big((r_1 \wedge \neg r_2) \to e_1\big) \wedge \big((r_2 \wedge \neg r_1) \to e_2\big)\Big).$$

It is routine exercise to check that, under the previous four hypotheses and the assumption that the robots' values are consistent, each robot can rationally choose to remove the obstacle, but it cannot rationally choose to refrain from doing it:

$$\models_{\mathbf{M}_{VC}} (\varphi_1 \wedge \varphi_2 \wedge \varphi_3 \wedge \varphi_4) \to \bigwedge_{i\in\{1,2\}} (\mathsf{C}_i r_i \wedge \neg \mathsf{C}_i \neg r_i).$$

If, moreover, both robots have knowledge about the four hypotheses, then they also have knowledge about each other's choices:

$$\models_{\mathbf{M}_{VC}} \bigwedge_{i\in\{1,2\}} \big(\mathsf{K}_i(\varphi_1 \wedge \varphi_2 \wedge \varphi_3 \wedge \varphi_4) \to \mathsf{K}_i \bigwedge_{i\in\{1,2\}} (\mathsf{C}_i r_i \wedge \neg \mathsf{C}_i \neg r_i)\big).$$

5 AXIOMATIZATION

In this section, we present an axiomatics for the set of validities of the language $\mathcal{L}$. In order to prove its completeness relative to the semantics defined in Section 3, it is useful to define a weaker semantics based on quasi-models.

Definition 5.1 (Quasi-model). A quasi-evaluative model (quasi-EM) is like an EM, as defined in Definition 3.1, except that Constraints C4 and C5 are replaced by the following four constraints, for every $w, v \in W$, $X \subseteq W$ and $i \in Agt$:

- **(C6)** if $w \preceq_i v$ then $Sat_M(i, w) \subseteq Sat_M(i, v)$,
- **(C7)** if $\mathsf{id}_i \in V(w)$ then $Sat_M(i, w) = N(i, w)$,
- **(C8)** if $w \equiv_i v$ and $\mathsf{id}_i \in V(v)$ then $w \preceq_i v$,
- **(C9)** if $w \preceq_i v$ and $\mathsf{id}_i \in V(w)$ then $\mathsf{id}_i \in V(v)$.

For every $X \subseteq \{VR, VC\}$, the class of quasi-EMs satisfying each semantic property in X is denoted by $\mathbf{QM}_X$.

It is routine exercise to verify that every EM is also a quasi-EM. Truth conditions of formulas in $\mathcal{L}$ relative to quasi-EMs are just like truth conditions relative to EMs (Definition 3.4). For every $X \subseteq \{VR, VC\}$, notions of validity and satisfiability relative to the classes $\mathbf{QM}_X$ are also defined in the usual way.

As the following theorem highlights, the language $\mathcal{L}$ is not expressive enough to distinguish the semantics based on quasi-models from the semantics based on models.

Theorem 5.2. *Let $\varphi \in \mathcal{L}$ and $X \subseteq \{VR, VC\}$. Then, φ is satisfiable for the class $\mathbf{M}_X$ if and only if it is satisfiable for the class $\mathbf{QM}_X$.*

In the following definition, we precisely define a family of logics of evaluation. We will show that they are sound and complete relative to both the quasi-model and the model semantics.

Definition 5.3 (Logic). We define LEV (Logic of Evaluation) to be the extension of classical propositional logic given by the following axioms and rule of inference:

$$\big(\mathsf{K}_i\varphi \wedge \mathsf{K}_i(\varphi \to \psi)\big) \to \mathsf{K}_i\psi \qquad (\mathbf{K}_{\mathsf{K}_i})$$
$$\mathsf{K}_i\varphi \to \varphi \qquad (\mathbf{T}_{\mathsf{K}_i})$$
$$\mathsf{K}_i\varphi \to \mathsf{K}_i\mathsf{K}_i\varphi \qquad (\mathbf{4}_{\mathsf{K}_i})$$
$$\varphi \to \mathsf{K}_i\widehat{\mathsf{K}}_i\varphi \qquad (\mathbf{B}_{\mathsf{K}_i})$$
$$\big([\preceq_i]\varphi \wedge [\preceq_i](\varphi \to \psi)\big) \to [\preceq_i]\psi \qquad (\mathbf{K}_{[\preceq_i]})$$
$$[\preceq_i]\varphi \to \varphi \qquad (\mathbf{T}_{[\preceq_i]})$$
$$[\preceq_i]\varphi \to [\preceq_i][\preceq_i]\varphi \qquad (\mathbf{4}_{[\preceq_i]})$$
$$\mathsf{K}_i\varphi \to [\preceq_i]\varphi \qquad (\mathbf{Mix}_{\mathsf{K}_i,[\preceq_i]})$$
$$\mathsf{K}_i(\varphi \leftrightarrow \psi) \to (\mathsf{V}_i\varphi \to \mathsf{V}_i\psi) \qquad (\mathbf{Mix1}_{\mathsf{K}_i,\mathsf{V}_i})$$
$$\mathsf{V}_i\varphi \to \mathsf{K}_i\mathsf{V}_i\varphi \qquad (\mathbf{Mix2}_{\mathsf{K}_i,\mathsf{V}_i})$$
$$(\mathsf{V}_i\varphi \wedge \varphi) \to [\preceq_i]\varphi \qquad (\mathbf{Mix}_{\mathsf{V}_i,[\preceq_i]})$$
$$(\mathsf{V}_i\varphi \wedge \mathsf{id}_i) \to \varphi \qquad (\mathbf{Mix}_{\mathsf{V}_i,\mathsf{id}_i})$$
$$[\preceq_i]\varphi \to \mathsf{K}_i(\mathsf{id}_i \to \varphi) \qquad (\mathbf{Mix1}_{[\preceq_i],\mathsf{id}_i})$$
$$\mathsf{id}_i \to [\preceq_i]\mathsf{id}_i \qquad (\mathbf{Mix2}_{[\preceq_i],\mathsf{id}_i})$$
$$\frac{\varphi}{\mathsf{K}_i\varphi} \qquad (\mathbf{Nec}_{\mathsf{K}_i})$$

For every $X \subseteq \{\mathbf{Real}_{\mathsf{V}_i}, \mathbf{Cons}_{\mathsf{V}_i}\}$, we define LEV_X to be the extension of the logic LEV by each axiom in X, where $\mathbf{Real}_{\mathsf{V}_i}$ and $\mathbf{Cons}_{\mathsf{V}_i}$ are the following axioms:

$$\mathsf{K}_i\varphi \to \neg\mathsf{V}_i\neg\varphi \qquad (\mathbf{Real}_{\mathsf{V}_i})$$
$$\widehat{\mathsf{K}}_i\mathsf{id}_i \qquad (\mathbf{Cons}_{\mathsf{V}_i})$$

Note that the base logic LEV is the same as $\mathrm{LEV}_\emptyset$. For each logic LEV_X, notions of theorem and consistency are defined in the usual way. We denote by $\vdash_{\mathrm{LEV}_X} \varphi$ the fact that φ is a theorem of LEV_X.

We have all S5-principles for the epistemic operator K_i (Axioms $\mathbf{K}_{\mathsf{K}_i}$, $\mathbf{T}_{\mathsf{K}_i}$, $\mathbf{4}_{\mathsf{K}_i}$ and $\mathbf{B}_{\mathsf{K}_i}$, and Rule $\mathbf{Nec}_{\mathsf{K}_i}$) and the S4-principles for the betterness operator $[\preceq_i]$ (Axioms $\mathbf{K}_{[\preceq_i]}$, $\mathbf{T}_{[\preceq_i]}$ and $\mathbf{4}_{[\preceq_i]}$). According to Axiom $\mathbf{Mix}_{\mathsf{K}_i,[\preceq_i]}$, if an agent knows that φ, then φ has to be true at all worlds which are for the agent at least as good as the current one. Indeed, an agent's preference is relative to her epistemic state. Axiom $\mathbf{Mix1}_{\mathsf{K}_i,\mathsf{V}_i}$ is a sort of 'co-extensionality' principle for values: if φ and ψ are equivalent (co-extensional) according to agent i, then i has the value that φ if and only if she has value that ψ. According to Axiom $\mathbf{Mix2}_{\mathsf{K}_i,\mathsf{V}_i}$, an agent has introspection over her values. Axioms $\mathbf{Mix}_{\mathsf{V}_i,[\preceq_i]}$ and $\mathbf{Mix}_{\mathsf{V}_i,\mathsf{id}_i}$ are the principles relating values with preferences. According to $\mathbf{Mix}_{\mathsf{V}_i,[\preceq_i]}$, if the actual world satisfies a certain value, then all better worlds should also satisfy it. According to $\mathbf{Mix}_{\mathsf{V}_i,\mathsf{id}_i}$, if the actual world is an ideal world, then it should satisfy all values. According to Axiom $\mathbf{Mix1}_{[\preceq_i],\mathsf{id}_i}$, if φ is true at all better worlds for the agent then, according to the agent, is true at all ideal worlds. Finally, according to Axiom $\mathbf{Mix2}_{[\preceq_i],\mathsf{id}_i}$, all better worlds of an ideal world are also ideal worlds. Axioms $\mathbf{Real}_{\mathsf{V}_i}$ and $\mathbf{Cons}_{\mathsf{V}_i}$ are, respectively, the value realism and the value consistency axiom.

The following rule of necessitation for the preference operator and rule of equivalence for the value operator are derivable in LEV

by means of Axiom $\mathbf{Mix}_{\mathsf{K}_i,[\preceq_i]}$, Axiom $\mathbf{Mix1}_{\mathsf{K}_i,\mathsf{V}_i}$ and Rule $\mathbf{Nec}_{\mathsf{K}_i}$:

$$\frac{\varphi}{[\preceq_i]\varphi} \tag{3}$$

$$\frac{\varphi \leftrightarrow \psi}{\mathsf{V}_i\varphi \leftrightarrow \mathsf{V}_i\psi} \tag{4}$$

Let f_c be the correspondence function associating each axiom in $\{\mathbf{Real}_{\mathsf{V}_i}, \mathbf{Cons}_{\mathsf{V}_i}\}$ to its corresponding semantic property in $\{VR, VC\}$:

$$f_c(\mathbf{Real}_{\mathsf{V}_i}) = VR,$$
$$f_c(\mathbf{Cons}_{\mathsf{V}_i}) = VC.$$

Our first result is about soundness and completeness relative to quasi-models.

Theorem 5.4. *Let $X \subseteq \{\boldsymbol{Real}_{\mathsf{V}_i}, \boldsymbol{Cons}_{\mathsf{V}_i}\}$. Then, the logic* LEV_X *is sound and complete for the class* $\mathbf{QM}_{\{f_c(x):x\in X\}}$.

Soundness and completeness relative to models is a direct corollary of Theorem 5.2 and Theorem 5.4.

Corollary 5.5. *Let $X \subseteq \{\boldsymbol{Real}_{\mathsf{V}_i}, \boldsymbol{Cons}_{\mathsf{V}_i}\}$. Then, the logic* LEV_X *is sound and complete for the class* $\mathbf{M}_{\{f_c(x):x\in X\}}$.

6 VALUE EXPANSION

So far, we have only considered the static aspects of evaluation. In this section, we look at the dynamic aspects by exploring the connection between value change and preference change. We extend the language $\mathcal{L}$ introduced in Section 3 by operators of the form $[+_J\varphi]$, which are used to describe the consequences of a value expansion operation by all agents in the coalition J. We call $\mathcal{L}^+(Atm, Agt)$ the resulting language and define it by the following grammar:

$$\varphi \quad ::= \quad p \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \mathsf{K}_i\varphi \mid \mathsf{V}_i\varphi \mid [\preceq_i]\varphi \mid [+_J\psi]\varphi,$$

where p ranges over Atm, i ranges over Agt and J ranges over the set of coalitions $2^{Agt*} = 2^{Agt} \setminus \{\emptyset\}$. The new formula $[+_J\psi]\varphi$ is meant to stand for "φ holds, after every agent in J has expanded her set of values with ψ". We assume that value expansion operations are public, i.e., if an agent expands her set of values with ψ, then this is common knowledge among all agents. This assumption could be relaxed by using a variant of event models [6] for our semantics, which would allow us to model private and semi-private value change operations. The following definition provides truth conditions for this new type of formulas.

Definition 6.1 (Satisfaction relation (cont.)). Let $M = (W, (\equiv_i)_{i\in Agt}, N, (\preceq_i)_{i\in Agt}, V)$ be an EM and let $w \in W$. Then:

$$M, w \models [+_J\psi]\varphi \iff M^{+_J\psi}, w \models \varphi,$$

where $M^{+_J\psi} = (W^{+_J\psi}, (\equiv_i^{+_J\psi})_{i\in Agt}, N^{+_J\psi}, (\preceq_i^{+_J\psi})_{i\in Agt}, V^{+_J\psi})$ such that $W^{+_J\psi} = W$, and for every $i \in Agt$ and $w \in W$:

$$\equiv_i^{+_J\psi} = \equiv_i,$$

$$N^{+_J\psi}(i,w) = \begin{cases} N(i,w) \cup \{||\psi||^M_{i,w}\} \text{ if } i \in J, \\ N(i,w) \text{ if } i \notin J, \end{cases}$$

$$\preceq_i^{+_J\psi} = \begin{cases} \preceq_i \setminus \{(w,v) : M, w \models \psi \text{ and } M, v \models \neg\psi\} \text{ if } i \in J, \\ \preceq_i^{+_J\psi} = \preceq_i \text{ if } i \notin J, \end{cases}$$

$$V^{+_J\psi}(w) = V(w) \setminus \{\mathsf{id}_i \in Atm^+ : w \notin ||\psi||^M_{i,w} \text{ and } i \in J\}.$$

The value expansion operation $+_J\psi$ only affects the mental attitudes of the agents in J, while keeping unchanged the mental attitudes of the agents outside J. Specifically, for every agent i in J, (i) it extends i's value set with ψ, (ii) it upgrades i's preference relation by removing from i's information set all preferences of a ψ-world over a $\neg\psi$-world and, finally, (iii) it shrinks i's set of ideal worlds to ψ-worlds. This update operation is well-defined since it preserves the properties of the model the class $\mathbf{M}$.

Proposition 6.2. *Let $\psi \in \mathcal{L}^+$ and $J \in 2^{Agt*}$. Then, if $M \in \mathbf{M}$ then $M^{+_J\psi} \in \mathbf{M}$.*

We leave for future work the definition of variants of the value expansion operation for the model classes $\mathbf{M}_{VR}$ and $\mathbf{M}_{VC}$. Just notice that, in order to preserve value realism, it would be necessary to make the execution of the update operation dependent on the satisfaction of the condition $\widehat{\mathsf{K}}_i\psi$. In other words, an agent in J will expand her values with ψ, only if ψ is consistent with her knowledge. In order to preserve value consistency, the update operation should be made subject to the satisfaction of the condition $\widehat{\mathsf{K}}_i(\psi \wedge \mathsf{id}_i)$. In other words, for an agent in J to expand her set of values with ψ, ψ has to be consistent with i's set of (envisaged) ideal situations.

It is time to define the the logic DLEV (Dynamic LEV) which extends the logic LEV by the dynamic operators $[+_J\psi]$.

Definition 6.3. We define DLEV to be the extension of the logic LEV of Definition 5.3 generated by the following reduction axioms for the dynamic operators $[+_J\psi]$:

$$[+_J\psi]p \leftrightarrow p \qquad (\mathbf{Red}_{+_J\psi,p})$$

$$[+_J\psi]\neg\varphi \leftrightarrow \neg[+_J\psi]\varphi \qquad (\mathbf{Red}_{+_J\psi,\neg})$$

$$[+_J\psi](\varphi_1 \wedge \varphi_2) \leftrightarrow ([+_J\psi]\varphi_1 \wedge [+_J\psi]\varphi_2) \qquad (\mathbf{Red}_{+_J\psi,\wedge})$$

$$[+_J\psi]\mathsf{K}_i\varphi \leftrightarrow \mathsf{K}_i[+_J\psi]\varphi \qquad (\mathbf{Red}_{+_J\psi,\mathsf{K}_i})$$

$$[+_J\psi]\mathsf{V}_i\varphi \leftrightarrow \big(\mathsf{V}_i[+_J\psi]\varphi \vee \mathsf{K}_i(\psi \leftrightarrow [+_J\psi]\varphi)\big) \text{ if } i \in J \qquad (\mathbf{Red1}_{+_J\psi,\mathsf{V}_i})$$

$$[+_J\psi]\mathsf{V}_i\varphi \leftrightarrow \mathsf{V}_i[+_J\psi]\varphi \text{ if } i \notin J \qquad (\mathbf{Red2}_{+_J\psi,\mathsf{V}_i})$$

$$[+_J\psi][\preceq_i]\varphi \leftrightarrow \Big((\psi \to [\preceq_i](\psi \to [+_J\psi]\varphi)) \wedge (\neg\psi \to [\preceq_i][+_J\psi]\varphi)\Big) \text{ if } i \in J \qquad (\mathbf{Red1}_{+_J\psi,[\preceq_i]})$$

$$[+_J\psi][\preceq_i]\varphi \leftrightarrow [\preceq_i][+_J\psi]\varphi \text{ if } i \notin J \qquad (\mathbf{Red2}_{+_J\psi,[\preceq_i]})$$

$$[+_J\psi]\mathsf{id}_i \leftrightarrow (\mathsf{id}_i \wedge \psi) \text{ if } i \in J \qquad (\mathbf{Red1}_{+_J\psi,\mathsf{id}_i})$$

$$[+_J\psi]\mathsf{id}_i \leftrightarrow \mathsf{id}_i \text{ if } i \notin J \qquad (\mathbf{Red2}_{+_J\psi,\mathsf{id}_i})$$

and the following rule of inference:

$$\frac{\psi_1 \leftrightarrow \psi_2}{\varphi \leftrightarrow \varphi[\psi_1/\psi_2]} \qquad (\mathbf{RRE})$$

It is routine exercise to verify that the equivalences in Definition 6.3 are valid for class $\mathbf{M}$ and that Rule **RRE** preserves validity. The completeness of DLEV for this class of models follows from Theorem 5.5, in view of the fact that the reduction axioms and the rule of replacement of proved equivalents can be used to find, for any $\mathcal{L}^+$-formula, a provably equivalent $\mathcal{L}$-formula.

Theorem 6.4. *The logic* DLEV *is sound and complete for the class* $\mathbf{M}$.

7 CONCLUSION

We have presented a logical analysis of both the static and the dynamic aspects of evaluation with the support of a multimodal language for knowledge, values and preferences. Moreover, we have provided sound and complete axiomatics for a family of logics of evaluation.

The evaluation criterion we have studied is purely qualitative. It is based on the idea that, for a world v to be at least as good another world w, the set of values satisfied at w should be included in the set of values satisfied at v. Future work will be devoted to study a variant of the logic LEV based on the following quantitative criterion:

$$w \preceq_i v \text{ if and only if } |Sat_M(i,w)| \leq |Sat_M(i,v)|.$$

This alternative criterion makes the preference ordering $\preceq_i$ complete and only works under the assumptions that an agent's value set is countable and that the agent's values have all the same weight. Future work will also be devoted to study further types of value change operation having an indirect influence on preferences including value forgetting — which removes a value from an agent's value set —, and value revision — which expands the agent's value set first and then restores global consistency of the agent's value set —. This latter operation is particularly relevant for the model class $\mathbf{M}_{VC}$.

A more long-term objective is to introduce a variant of the logic of evaluation in which values are graded (i.e., an agent can have values with different weight or priority) and the preference ordering over the worlds in built by taking the values' priorities into account.

We also plan to compare the notion of choice in the *seeing to it that* (STIT) logic [7, 24, 27] with the notion we defined in Section 4. While in STIT choices are given as primitives, in our approach they are computed from values.

Last but not least, we plan to study decidability and complexity of satisfiability checking for the three logics LEV, $\text{LEV}_{\{\mathbf{Real}_{V_i}\}}$ and $\text{LEV}_{\{\mathbf{Cons}_{V_i}\}}$. Concerning decidability, we believe we can use filtration techniques to prove it. The limitation of filtration is that, by using it for satisfiability checking, one has to guess a model exponential in the size of the formula to be checked. Thus, by filtration argument, we could only prove that satisfiability checking is in NEXPTIME. Our conjecture is that satisfiability checking for our logics is EXPTIME-complete. It is certainly EXPTIME-hard since, given Constraint C1 in Definition 3.1, the S5-modality K_i plays the role of the universal modality with respect to the S4-modality $[\preceq_i]$. As shown in [23], adding the universal modality to a multimodal logic with independent modalities such as $[\preceq_i]$ causes EXPTIME-hardness.

A SELECTED PROOFS

A.1 Proof of Theorem 5.2

Proof. The left-to-right direction is trivial since every EM is also a quasi-EM.

In order to prove the right-to-left direction, we start with a quasi-EM satisfying φ and transform it to obtain an EM which satisfies the same formulas as the original model. Let $M = (W, (\equiv_i)_{i\in Agt}, N, (\preceq_i)_{i\in Agt}, V)$ be a quasi-EM and $w \in W$ such that $M, w \models \varphi$.

We define the structure $M' = (W', (\equiv'_i)_{i\in Agt}, N', (\preceq'_i)_{i\in Agt}, V')$ as follows:

- $W' = \{w_1 : w \in W\} \cup \{w_2 : w \in W\}$;
- for every $x, y \in \{1,2\}$ and $w_x, v_y \in W'$, $w_x \equiv'_i v_y$ if and only if $w \equiv iv$;
- for every $x, y \in \{1,2\}$ and $w_x, v_y \in W'$, $w_x \preceq'_i v_y$ if and only if:
  - $x = y$ and $w \preceq_i v$, or
  - $v \in I_M(i,w)$;
- for every $x \in \{1,2\}$ and $w_x \in W'$, $V'(w_x) = V(w)$;
- for every $x \in \{1,2\}$ and $w_x \in W'$,

$$N'(i,w_x) = \Big\{\{w_1 : w \in X\} \cup \{w_2 : w \in X\} : X \in N(i,w)\Big\} \cup \bigcup_{v_y \in \equiv_i(w_x) \text{ and } v_y \notin I_{M'}(i,w_x)} \big\{\preceq'(v_y) \cup I_{M'}(i,w_x)\big\}.$$

The first thing to verify is that M' is an EM. It is routine exercise to check that every $\equiv'_i$ is an equivalence relation and that every $\preceq'_i$ is a partial preorder. Clearly, it satisfies Constraints C1, C2 and C3 in Definition 3.1 since M satisfies them too. Let us prove that it satisfies Constraint C4. We distinguish two cases: (Case 1) $w_x \notin I_{M'}(i,w_x)$, and (Case 2) $w_x \in I_{M'}(i,w_x)$.

Case 1. We first prove the left-to-right direction. Suppose $w_x \preceq'_i v_y$. By definition of $\preceq'_i$, we have (i) ($x = y$ and $w \preceq_i v$) or (ii) $v \in I_M(i,w)$. Since M is a quasi-EM which satisfies Constraints C3, C6 and C7, if $v \in I_M(i,w)$ then $Sat_M(i,v) = N(i,v) = N(i,w)$, and if $w \preceq_i v$ then $Sat_M(i,w) \subseteq Sat_M(i,v)$. Thus, $Sat_M(i,w) \subseteq Sat_M(i,v)$. By construction of M', the fact that $w_x \preceq'_i v_y$, the transitivity of $\preceq'$ and the fact that $Sat_M(i,w) \subseteq Sat_M(i,v)$, both (i) and (ii) imply $Sat_{M'}(i,w_x) \subseteq Sat_{M'}(i,v_y)$.

As for the right-to-left direction, suppose $v_y \notin \preceq'_i (w_x)$. By definitions of $\preceq_i$ and V', the latter implies that $v_y \notin I_{M'}(i,w_x)$. Thus, $v_y \notin \big(\preceq'_i (w_x) \cup I_{M'}(i,w_x)\big)$. Since $w_x \notin I_{M'}(i,w_x)$, by definition of N', we have $\big(\preceq'_i(w_x) \cup I'(i,w_x)\big) \in N'(w_x)$. Moreover, $w_x \in \preceq'_i (w_x)$. Therefore, $Sat_{M'}(i,w_x) \nsubseteq Sat_{M'}(i,v_y)$. Consequently, $Sat_{M'}(i,w_x) \subseteq Sat_{M'}(i,v_y)$ implies $w_x \preceq_i v_y$.

Case 2. As for the left-to-right direction, suppose $w_x \preceq'_i v_y$. The latter means that (i) $x = y$ and $w \preceq_i v$, or (ii) $v \in I_M(i,w)$. Suppose (i). Since M is a quasi-EM which satisfies Constraint C9, by definition of V' and the fact that $w_x \in I_{M'}(i,w_x)$, we have $v \in I_M(i,w)$. Thus, both (i) and (ii) imply $w \in I_M(i,w)$ and $v \in I_M(i,w)$. Hence, by construction of M' and the fact that M is a quasi-EM which satisfies Constraints C3 and C7, we have $Sat_{M'}(i,w_x) = Sat_{M'}(i,v_y)$.

As for the right-to-left direction, suppose $v_1 \notin \preceq'_i (w_x)$. By definition of $\preceq'_i$, $v \notin I_M(i,w)$. Therefore, $v_1 \notin I_{M'}(i,w_x)$. Let $X = \big(\preceq'_i (v_1) \cup I_{M'}(i,w_x)\big)$ and $X' = \big(\preceq'_i (v_2) \cup I_{M'}(i,w_x)\big)$. Because of $v_1 \notin \preceq'_i (w_x)$ and $v_1 \notin I'(i,w_x)$, by definitions of N' and $\preceq'_i$, we have $X, X' \in N'(i,w_x)$, $X \neq X'$, $v_1 \in X$ and $v_1 \notin X'$. Since $w_x \in I_{M'}(i,w_x)$, by construction of M', we have $Sat_{M'}(i,w_x) = N'(i,w_x)$. It follows that $Sat_{M'}(i,w_x) \nsubseteq Sat_{M'}(i,v_1)$. We can conclude that $Sat_{M'}(i,w_x) \subseteq Sat_{M'}(i,v_1)$ implies $w_x \preceq_i v_1$. In an analogous way, we can prove that $Sat_M(i,w_x) \subseteq Sat_M(i,v_2)$ implies $w_x \preceq_i v_2$.

In order to prove that M' satisfies Constraint C5, we first observe that, by construction of M', if $v_y \in I_{M'}(i,w_x)$ then $Sat_{M'}(i,v_y) = N'(i,v_y)$.

Now, suppose $v_1 \notin I_{M'}(i, w_x)$ and $v_1 \in \equiv_i(w_x)$. Thus, by definition of V', $v \notin I(i, w)$ and $v_2 \notin I'(i, w_x)$. Therefore, by definition of $\preceq'_i$ and N', $v_1 \notin \left(\preceq'_i (v_2) \cup I'(i, w_x)\right)$ and $\left(\preceq'_i (v_2) \cup I'(i, w_x)\right) \in N'(i, w_x)$. Thus, $Sat_{M'}(i, v_1) \neq N'(i, v_1)$. We can conclude that $Sat_{M'}(i, v_1) = N'(i, v_1)$ and $v_1 \in \equiv_i (w_x)$ imply $v_1 \in I'(i, w_x)$. The proof that $Sat_{M'}(i, v_2) = N'(i, w_x)$ and $v_2 \in \equiv_i (w_x)$ imply $v_2 \in I'(i, w_x)$ is analogous.

It is easy to show that, for every $X \subseteq \{VR, VC\}$, if M satisfies every property in X, then M' satisfies them too.

By induction on the structure of the formula φ, we can prove that $M, w \models \varphi$ if and only if $M', w_x \models \varphi$, for every $x \in \{1, 2\}$ and $w \in W$. □

A.2 Proof of Theorem 5.4

Proof. It is routine to check that the axioms of LEV are all valid relative to the class of quasi-EMs and that the rule of inference $\mathbf{Nec}_{\mathsf{K}_i}$ preserves validity with respect to this class.

To prove completeness, we use a canonical model argument.

We consider maximally consistent sets of formulas in $\mathcal{L}$ (MCSs). The following proposition specifies some usual properties of MCSs.

Proposition A.1. *Let Γ be a MCS and let $\varphi, \psi \in \mathcal{L}$. Then:*

- *if $\varphi, \varphi \rightarrow \psi \in \Gamma$ then $\psi \in \Gamma$;*
- *$\varphi \in \Gamma$ or $\neg\varphi \in \Gamma$;*
- *$\varphi \vee \psi \in \Gamma$ iff $\varphi \in \Gamma$ or $\psi \in \Gamma$.*

The following is the Lindenbaum's lemma for our logic. As the proof is standard (cf. [10, Lemma 4.17]) we omit it here.

Lemma A.2. *Let Δ be a* LEV*-consistent set of formulas. Then, there exists a MCS Γ such that $\Delta \subseteq \Gamma$.*

Let the canonical quasi-EM model be the tuple $M^c = (W^c, (\equiv_i^c)_{i \in Agt}, N^c, (\preceq_i^c)_{i \in Agt}, V^c)$ such that:

- W^c is set of all MCSs;
- for all $w, v \in W^c$ and $i \in Agt$, $w \equiv_i^c v$ iff, for all $\varphi \in \mathcal{L}$, if $\mathsf{K}_i\varphi \in w$ then $\varphi \in v$;
- for all $w, v \in W^c$ and $i \in Agt$, $w \preceq_i^c v$ iff, for all $\varphi \in \mathcal{L}$, if $[\preceq_i]\varphi \in w$ then $\varphi \in v$;
- for all $w \in W^c$ and $i \in Agt$, $N^c(i, w) = \{A_\varphi(i, w) : \mathsf{V}_i\varphi \in w\}$;
- for all $w \in W^c$ and $p \in Atm$, $p \in V^c(w)$ iff $p \in w$;

with $A_\varphi(i, w) = \left\{v \in \equiv_i^c(w) : \varphi \in v\right\}$.

We have to prove that M^c is a quasi-EM by showing that, for every $i \in Agt$, $\equiv_i^c$ is an equivalence relation and $\preceq_i$ is partial preorder, and that Conditions C1, C2, C3, C6, C7, C8 and C9 in Definitions 3.1 and 5.1 are satisfied. The proof uses Proposition A.1: Axiom $\mathbf{T}_{\mathsf{K}_i}$ guarantees that $\equiv_i^c$ is reflexive, Axiom $\mathbf{4}_{\mathsf{K}_i}$ guarantees that it is transitive, and Axiom $\mathbf{B}_{\mathsf{K}_i}$ guarantees that it is symmetric; Axiom $\mathbf{T}_{[\preceq_i]}$ guarantees that $\preceq_i^c$ is reflexive, while Axiom $\mathbf{4}_{[\preceq_i]}$ guarantees that it is transitive. We are going to show that M^c satisfies Constraints C1, C2, C3 and C6 as an example. We leave to the reader the task of proving that it satisfies Constraints C7, C8 and C9 as well.

As for C1, suppose $w \preceq_i^v v$ and $w \not\equiv_i^c v$. By definition of $\equiv_i^c$, $\mathsf{K}_i\varphi \in w$ and $\varphi \notin v$ for some φ. By Axiom $\mathbf{Mix}_{\mathsf{K}_i,[\preceq_i]}$ and Proposition A.1, $[\preceq_i]\varphi \in w$. Hence, by definition of $\preceq_i^c$, we have $\varphi \in v$ which leads to a contradiction.

As for C2, suppose $X \in N^c(i, w)$. The latter means that $X = \{u \in \equiv_i^c (w) : \varphi \in u\}$ and $\mathsf{V}_i\varphi \in w$ for some φ. Thus, clearly, $X \subseteq \equiv_i^c(w)$.

As for C3, suppose $w \equiv_i^c v$, $X \in N^c(i, w)$ and $X \notin N^c(i, v)$. By definition of N^c, $X = A_\varphi(i, w)$ for some φ such that $\mathsf{V}_i\varphi \in w$ and $\mathsf{V}_i\varphi \notin v$. Hence, by Axiom $\mathbf{Mix2}_{\mathsf{K}_i,\mathsf{V}_i}$ and Proposition A.1, $\mathsf{K}_i\mathsf{V}_i\varphi \in w$. Because of $w \equiv_i^c v$, the latter implies $\mathsf{V}_i\varphi \in v$ which leads to a contradiction. Since the relation $w \equiv_i^c v$ is symmetric, in an analogous way we can prove that $w \equiv_i^c v$, $X \notin N^c(i, w)$ and $X \in N^c(i, v)$ also leads to a contradiction.

As for C6, suppose $w \preceq_i^c v$ and $w \in X$ for some $X \in N_i^c(i, w)$. We are going to prove that $v \in X$ and $X \in N_i^c(i, v)$. Suppose not. Thus, either $v \notin X$ or $X \notin N_i^c(i, v)$. $X \in N_i^c(i, w)$ means that $X = A_\varphi(i, w)$ for some $\mathsf{V}_i\varphi \in w$. By Axiom $\mathbf{Mix2}_{\mathsf{K}_i,\mathsf{V}_i}$ and Proposition A.1, the latter implies that $X = A_\varphi(i, w)$ and $\mathsf{K}_i\mathsf{V}_i\varphi \in w$. Thus, by definition of $\equiv_i^c$, $\mathsf{V}_i\varphi \in v$. Consequently, since $A_\varphi(i, w) = A_\varphi(i, v)$, $X = A_\varphi(i, w) \in N_i^c(i, v)$ which is in contradiction with $X \notin N_i^c(i, v)$. We only need to show that $v \notin X$ also leads to a contradiction. From $w \in X$ and $X \in N_i^c(i, w)$, we have $w \in X$ and $X = A_\varphi(i, w)$ for some $\mathsf{V}_i\varphi \in w$. Thus, by definition of $A_\varphi(i, w)$, $\varphi \in w$ and $\mathsf{V}_i\varphi \in w$. By Axiom $\mathbf{Mix}_{\mathsf{V}_i,[\preceq_i]}$ and Proposition A.1, the latter implies that $[\preceq_i]\varphi \in w$. Thus, since $w \preceq_i^c v$, $\varphi \in v$. Hence, since $w \equiv_i^c v$, we have $v \in X$ which is in contradiction with the initial assumption.

The next step in the proof is the following existence lemma. The proof is again standard (cf. [10, Lemma 4.20]) and we omit it.

Lemma A.3. *Let $\varphi \in \mathcal{L}$ and $w \in W^c$. If $\widehat{\mathsf{K}}_i\varphi \in w$, then there exists $v \in W^c$ such that $w \equiv_i^c v$ and $\varphi \in v$.*

Finally, we can prove the following truth lemma.

Lemma A.4. *Let $\varphi \in \mathcal{L}$ and $w \in W^c$. Then, $M^c, w \models \varphi$ iff $\varphi \in w$.*

Proof. The proof is by induction on the structure of the formula. The cases with φ atomic, Boolean, and the form $\mathsf{K}_i\psi$ are provable in the standard way (cf. [10, Lemma 4.21]).

The proof for the case $\varphi = \mathsf{V}_i\psi$ goes as follows.

($\Rightarrow$) Suppose $M^c, w \models \mathsf{V}_i\psi$. Thus, $\{u \in \equiv_i^c(w) : M^c, u \models \psi\} \in N^c(i, w)$. Hence, by definition of N^c, there exists χ such that $\mathsf{V}_i\chi \in w$ and $\{u \in \equiv_i^c(w) : \chi \in u\} = \{u \in \equiv_i^c(w) : M^c, u \models \psi\}$. Thus, by induction hypothesis, $\{u \in \equiv_i^c(w) : \chi \in u\} = \{u \in \equiv_i^c(w) : \psi \in u\}$. Now, suppose that $\mathsf{K}_i(\chi \leftrightarrow \psi) \notin w$. By Proposition A.1, it follows that $\neg\mathsf{K}_i(\chi \leftrightarrow \psi) \in w$. This means that $\widehat{\mathsf{K}}_i((\chi \wedge \neg\psi) \vee (\neg\chi \wedge \psi)) \in w$. By Lemma A.3, the latter implies that there exists $v \in W^c$ such that $w \equiv_i^c v$ and $((\chi \wedge \neg\psi) \vee (\neg\chi \wedge \psi)) \in v$ which is in contradiction with $\{u \in \equiv_i^c (w) : \chi \in u\} = \{u \in \equiv_i^c (w) : \psi \in u\}$. Thus, we have $\mathsf{K}_i(\chi \leftrightarrow \psi) \in w$. From $\mathsf{V}_i\chi \in w$ and $\mathsf{K}_i(\chi \leftrightarrow \psi) \in w$, by Proposition A.1 and Axiom $\mathbf{Mix1}_{\mathsf{K}_i,\mathsf{V}_i}$, it follows that $\mathsf{V}_i\psi \in w$.

($\Leftarrow$) Suppose $\mathsf{V}_i\psi \in w$. Thus, by definition of N^c, $A_\psi(i, w) = \{v \in \equiv_i^c(w) : \psi \in v\} \in N^c(i, w)$. Hence, by induction hypothesis, $\{v \in \equiv_i^c(w) : M^c, v \models \psi\} \in N^c(i, w)$. Therefore, $M^c, w \models \mathsf{V}_i\psi$. □

To conclude the proof, suppose φ is a LEV-consistent formula in $\mathcal{L}$. By Lemma A.2, there exists $w \in W^c$ such that $\varphi \in w$. Hence, by Lemma A.4, there exists $w \in W^c$ such that $M^c, w \models \varphi$. □

ACKNOWLEDGMENTS

Support from the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANITI) is gratefully acknowledged.

REFERENCES

[1] R. P. Abelson. 1979. Difference between belief and knowledge systems. *Cognitive Science* 3 (1979), 355–366.

[2] N. Ajmeri, H. Guo, P. K. Murukannaiah, and M. P. Singh. 2020. Elessar: Ethics in Norm-Aware Agents. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020)*. IFAAMAS, 16–24.

[3] C. Allen, G. Varner, and J. Zinser. 2000. Prolegomena to any future artificial moral agent. *Journal of Experimental and Theoretical Artificial Intelligence* 12, 3 (2000), 251–261.

[4] R. C. Arkin, P. Ulam, and A. R. Wagner. 2012. Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception. *Proc. IEEE* 100, 3 (2012), 571–589.

[5] P. Balbiani, D. Fernández-Duque, and E. Lorini. 2016. A Logical Theory of Belief Dynamics for Resource-Bounded Agents. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (AAMAS 2016)*. ACM, 644–652.

[6] A. Baltag, L. Moss, and S. Solecki. 1998. The Logic of Public Announcements, Common Knowledge and Private Suspicions. In *Proceedings of the Seventh Conference on Theoretical Aspects of Rationality and Knowledge (TARK'98)*, Itzhak Gilboa (Ed.). Morgan Kaufmann, San Francisco, CA, 43–56.

[7] N. Belnap, M. Perloff, and M. Xu. 2001. *Facing the future: agents and choices in our indeterminist world*. Oxford University Press, New York.

[8] C. Benzmüller, X. Parent, and L. W. N. van der Torre. 2020. Designing normative theories for ethical and legal reasoning: LogiKEy framework, methodology, and tool support. *Artificial Intelligence* 287 (2020).

[9] F. Berreby, G. Bourgne, and J.-G. Ganascia. 2017. A Declarative Modular Framework for Representing and Applying Ethical Principles. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2017)*. ACM, 96–104.

[10] P. Blackburn, M. de Rijke, and Y. Venema. 2001. *Modal Logic*. Cambridge University Press, Cambridge.

[11] C. Castelfranchi and F. Paglieri. 2007. The role of beliefs in goal dynamics: prolegomena to a constructive theory of intentions. *Synthese* 155, 2 (2007), 237–263.

[12] S. Cranefield, M. Winikoff, V. Dignum, and F. Dignum. 2017. No Pizza for You: Value-based Plan Selection in BDI Agents. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)*. ijcai.org, 178–184.

[13] L. A. Dennis, M. Fisher, M. Slavkovik, and M. Webster. 2016. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems* 77 (2016), 1–14.

[14] J. Dewey. 1939. *Theory of Valuation*. University of Chicago Press.

[15] F. Dietrich and C. List. 2013. A reason-based theory of rational choice. *Noûs* 47, 1 (2013), 104–134.

[16] F. Dietrich and C. List. 2017. What matters and how it matters: A choice-theoretic representation of moral theories. *Philosophical Review* 126, 4 (2017), 421–479.

[17] V. Dignum. 2017. Responsible Autonomy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)*. ijcai.org, 4698–4704.

[18] A. Etzioni and O. Etzioni. 2017. Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics* 21 (2017), 403–418.

[19] R. Fagin, J. Halpern, Y. Moses, and M. Vardi. 1995. *Reasoning about Knowledge*. MIT Press, Cambridge.

[20] P. M. Gollwitzer. 1996. The volitional benefits of planning. In *The psychology of action*. Guilford Press, 287–312.

[21] S. O. Hansson. 2001. *The Structure of Values and Norms*. Cambridge University Press.

[22] J. Harsanyi. 1982. Utilitarianism and Beyond. In *Morality and the theory of rational behaviour*, A. K. Sen and B. Williams (Eds.). Cambridge University Press, Cambridge.

[23] E. Hemaspaandra. 1996. The Price of Universality. *Notre Dame Journal of Formal Logic* 37, 2 (1996), 174–203.

[24] J. F. Horty. 2001. *Agency and Deontic Logic*. Oxford University Press, Oxford.

[25] I. Levi. 1990. *Hard Choices: Decision Making Under Unresolved Conflict*. Cambridge University Press.

[26] F. Liu. 2011. *Reasoning about Preference Dynamics*. Springer.

[27] E. Lorini. 2013. Temporal STIT logic and its application to normative reasoning. *Journal of Applied Non Classical Logics* 23, 4 (2013), 372–399.

[28] E. Lorini. 2015. A logic for reasoning about moral agents. *Logique & Analyse* 58, 230 (2015), 177–218.

[29] M. Miceli and C. Castelfranchi. 2000. The role of evaluation in cognition and social interaction. In *Advances in consciousness research. Human cognition and social agent technology*, K. Dautenhahn (Ed.). John Benjamins Publishing Company, 225–261.

[30] A. Moors, P. C. Ellsworth, K. Scherer, and N. Frijda. 2013. Appraisal Theories of Emotion: State of the Art and Future Development. *Emotion Review* 5, 2 (2013), 119–124.

[31] P. K. Murukannaiah, N. Ajmeri, C. M. Jonker, and M. P. Singh. 2020. New Foundations of Ethical Multiagent Systems. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020)*. IFAAMAS, 1706–1710.

[32] T. Nagel. 1980. The Limits of Objectivity. In *The Tanner Lectures on Human Values*. Cambridge University Press, 77–139.

[33] M. Rodriguez-Soto, M. López-Sánchez, and J. A. Rodríguez-Aguilar. 2020. A Structural Solution to Sequential Moral Dilemmas. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2018)*. IFAAMAS, 1152–1160.

[34] J. Searle. 2001. *Rationality in Action*. MIT Press, Cambridge.

[35] A. Sen. 1985. Well-Being, Agency and Freedom: The Dewey Lectures 1984. *The Journal of Philosophy* 82, 4 (1985), 169–221.

[36] A. Sen. 1987. *On Ethics and Economics*. Basil Blackwell.

[37] M. Serramia, M. López-Sánchez, J. A. Rodríguez-Aguilar, M. Rodríguez, M. J. Wooldridge, J. Morales, and C. Ansótegui. 2018. Moral Values in Norm Decision Making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2018)*. IFAAMAS, 1294–1302.

[38] J. van Benthem. 2016. Tracking Information. In *J. Michael Dunn on Information Based Logics*, K. Bimbó (Ed.). Springer, 363–389.

[39] J. van Benthem, D. Fernández-Duque, and E. Pacuit. 2014. Evidence and plausibility in neighborhood structures. *Annals of Pure and Applied Logic* 165, 1 (2014), 106–133.

[40] J. van Benthem and F. Liu. 2007. Dynamic logic of preference upgrade. *Journal of Applied Non Classical Logics* 17, 2 (2007), 157–182.

[41] D. Vanderelst and A. F. T. Winfield. 2018. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research* 48 (2018), 56–66.

[42] W. Wallach and C. Allen. 2008. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.

[43] A. F. T. Winfield, C. Blum, and W. Liu. 2014. Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection. In *Proceedings of the 15th Annual Conference on Advances in Autonomous Robotics Systems (TAROS 2014) (LNCS, Vol. 8717)*. Springer, 85–96.