Moral Planning Agents with LTL Values

Umberto Grandi, Emiliano Lorini, Timothy Parker

Moral Planning Agents with LTL Values

Umberto Grandi , Emiliano Lorini , Timothy Parker
IRIT, CNRS, University of Toulouse, Toulouse, France
{umberto.grandi,emiliano.lorini,timothy.parker}@irit.fr

Abstract

A moral planning agent (MPA) seeks to compare two plans or compute an optimal plan in an interactive setting with other agents, where relative ideality and optimality of plans are defined with respect to a prioritized value base. We model MPAs whose values are expressed by formulas of linear temporal logic (LTL) and define comparison for both joint plans and individual plans. We introduce different evaluation criteria for individual plans including an optimistic (risk-seeking) criterion, a pessimistic (risk-averse) one, and two criteria based on the use of anticipated responsibility. We provide complexity results for a variety of MPA problems.

1 Introduction

Evaluation is a core concept in cognitive theories of action [Gollwitzer, 1996], emotion [Moors *et al.*, 2013], knowledge and beliefs [Abelson, 1979], and in the connection between epistemic and motivational attitudes [Miceli and Castelfranchi, 2000]. It is the operation of comparing the goodness or desirability of options from a given set of alternatives in relation to a set of goals. It also important for ethics where the philosophical doctrine of pluralistic consequentialism [Sen, 1985; Sen, 1987] and recent theories of reason-based choice [Dietrich and List, 2013; Dietrich and List, 2017] have emphasized its crucial role in decision-making. According to pluralistic consequentialism, a moral agent has to weigh different and sometimes conflicting values to assess the relative *ideality* of different alternatives. Evaluation also plays a pivotal role is some existing computational models of ethical deliberation and planning in robotics [Arkin *et al.*, 2012; Vanderelst and Winfield, 2018] and in AI [Rodriguez-Soto *et al.*, 2020; Serramia *et al.*, 2018; Ajmeri *et al.*, 2020; Cranefield *et al.*, 2017]. More recently, it was also formalized in a multi-modal logic of values, knowledge and preferences [Lorini, 2021].

In this paper we study the role of evaluation in a novel multi-agent planning setting where agents share a set of ethical values expressed by formulas of linear temporal logic (LTL), ranked according to their priority. In particular, we focus on evaluation in the context of moral planning agents (MPAs). MPAs compare plans, which are either joint plans or individual plans, lexicographically with respect to their prioritized value base. A joint plan is a finite sequence of joint actions performed by a coalition of agents, while an individual plan is a finite sequence of individual actions performed by a single agent. The notion of joint plan is applicable whenever planning is delegated to a central planner which has to compute the optimal solution for all agents who will each execute their part of the joint plan on their own. Alternatively, the notion of individual plan is relevant for decentralized applications where agents do not know each other's plans.

In order for an agent to compare individual plans, it must consider all possible outcomes given the possible actions of the other agents. We study several evaluation criteria for individual plans: an optimistic (risk-seeking) criterion, a pessimistic (risk-averse) one, and criteria based on intrinsically ethical notions of responsibility. An optimistic agent compares individual plans by considering for each plan the best-possible history that could result. A pessimistic agent considers only the worst-possible histories. Finally, an agent sensitive to anticipated blameworthiness will be concerned by the possible violation of some ethical values that it brought about (anticipated active blameworthiness) or that it could have prevented (anticipated passive blameworthiness). These two notions of blameworthiness rely on more primitive concepts of active and passive responsibility. Similar notions of optimism and pessimism have been been studied in the domain of qualitative decision theory [Brafman and Tennenholtz, 2018], while concepts of responsibility have been formalized in logical settings [Lorini *et al.*, 2014], in causal models [Chockler and Halpern, 2004] and game-theoretic settings [Braham and van Hees, 2012; Lorini and Mühlenbernd, 2015; Lorini and Mühlenbernd, 2018]. However, to the best of our knowledge, the use of responsibility and blameworthiness in plan evaluation is novel.

The paper is organised as follows: Section 2 defines our model of multi-agent planning and introduces an illustrative example of our model in action. Section 3 describes how we evaluate joint plans and demonstrates this in our example. Section 4 focuses on individual plans, introducing and discussing optimistic and pessimistic comparison, as well as our notion of blameworthiness. Section 5 analyses the computational complexity of the comparison notions introduced in Section 4. Section 6 situates our paper in the wider fields of ethics in AI and planning, and compares our work to a

number of similar papers. Finally Section 7 summarises the paper and suggests directions for future work.

2 The Model

In this section we introduce and define our planning framework for Moral Planning Agents. Our model is grounded on the logical theory of ethical choice presented by Lorini [2015], which is in turn based on situation calculus [Reiter, 2001]. We use a compact representation of values as formulas of linear temporal logic, in line with the work of Bienvenu et al. [2006] on preference-based temporal planning.

2.1 Agents, Actions, and Histories

Let Agt be a set of agents, and let $Prop$ be a countable set of atomic propositions, defining a set of states $S = 2^{Prop}$, with elements $s, s', \ldots$. Let Act be a finite non-empty set of action names. Elements of $Prop$ are noted $p, q, \ldots$, while elements of Act are noted $a, b, \ldots$. We also assume the existence of a special action $skip$.

We define a k-history to be a pair $H = (H_{st}, H_{act})$ with $H_{st} : \{0, \ldots, k\} \longrightarrow S$ and $H_{act} : Agt \times \{0, \ldots, k-1\} \longrightarrow Act$. A history specifies the actual configuration of the environment (i.e., the state) at a certain time point and the actions executed by each of the agents that inform the next state. The set of k-histories is noted $Hist_k$. The set of all histories is $Hist = \bigcup_{k \in \mathbb{N}} Hist_k$.

2.2 Multi-Agent Action Theory

To represent actions' effects and preconditions compactly, we use a multi-agent action theory inspired by situation calculus [Reiter, 2001] and developed by Lorini [2015].

Let $\mathcal{L}_{\mathsf{PL}}$ be the standard propositional language built from atomic formulas p for $p \in Prop$ and $do(i, a)$ where $i \in Agt$ and $a \in Act$. We suppose actions in Act are described by an action theory $\gamma = (\gamma^+, \gamma^-)$, where γ^+ and γ^- are, respectively, the positive and negative effect precondition function $\gamma^+ : Agt \times Act \times Prop \longrightarrow \mathcal{L}_{\mathsf{PL}}$ and $\gamma^- : Agt \times Act \times Prop \longrightarrow \mathcal{L}_{\mathsf{PL}}$.

The fact $\gamma^+(i, a, p)$ guarantees that proposition p will be *true* in the next state when action a is executed by agent i (provided no other action interferes), while $\gamma^-(i, a, p)$ guarantees that proposition p will be *false* in the next state when action a is executed by i (without interference). In case of conflicts between actions, we use an intertial principle: if $\gamma^+(i, a, p)$ and $\gamma^-(j, b, p)$ are both true at a given state and actions a and b (which may be the same action) are executed by agents i and j (which may also be the same), then the truth value of p will not change in the next state.

Furthermore, in cases where not all actions are available to all agents we can simply set $\gamma^+(i, a, p) = \gamma^-(i, a, p) = \bot$ for all $p \in Prop$ to signal that action a is not available to agent i. We also the assume the existence of the special action $skip$, such that $\gamma^+(i, skip, p) = \gamma^-(i, skip, p) = \bot$ for all i and p. Equivalently, we could define $skip$ as follows: $\gamma^+(i, skip, p) = p$ and $\gamma^-(i, skip, p) = \neg p$ for all i and p.

Definition 1 (Action-compatible histories)**.** *Let $\gamma = (\gamma^+, \gamma^-)$ be an action theory and let $H = (H_{st}, H_{act})$ be a k-history. We say H is compatible with γ if the following condition holds for every $t \in \{0, \ldots, k-1\}$, $H_{st}(t+1) = (H_{st}(t) \setminus X) \cup Y$ where:*

$$
\begin{aligned}
X = \{&p \in Prop : \big(\exists i \in Agt, \exists a \in Act \text{ such that} \\
&H_{act}(i,t) = a \text{ and } H, t \models \gamma^-(i,a,p)\big) \text{ and} \\
&\big(\forall j \in Agt, \forall b \in Act \text{ if } H_{act}(j,t) = b \text{ then} \\
&H, t \models \neg\gamma^+(j,b,p)\big)\big\} \\
Y = \{&p \in Prop : \big(\exists i \in Agt, \exists a \in Act \text{ such that} \\
&H, t \models \gamma^+(i,a,p)\big) \text{ and } \big(\forall j \in Agt, \forall b \in Act \text{ if} \\
&H_{act}(j,t) = b \text{ then } H, t \models \neg\gamma^-(j,b,p)\big)\big\}.
\end{aligned}
$$

In words, a history H is compatible with the action theory $\gamma = (\gamma^+, \gamma^-)$ if its state transition respects the theory. This means that, all propositional facts for which the negative effect precondition of an executed action hold, while the positive effect preconditions of all executed actions do not hold, become false. In addition, all propositional facts for which the positive effect precondition of an executed action holds, while the negative effect preconditions of all executed actions do not hold, become true. All other propositional facts do not change their truth values.

2.3 Action Sequences and Joint Plans

Let us now move from the notion of action to the notion of plan. Given $k \in \mathbb{N}$, a k-action-sequence is a function

$$\pi : \{0, \ldots, k-1\} \longrightarrow Act.$$

The set of k-action-sequences is noted Seq_k. For a (non-empty) coalition of agents $J \in 2^{Agt} \setminus \emptyset$ we can define a joint k-plan as a function $\Pi : J \longrightarrow Seq_k$. If J is a singleton set then we call Π an individual plan. The set of joint k-plans for a coalition J is written $Plan_k^J$. The set of all joint plans for a non-empty coalition J is $Plan^J = \bigcup_{k \in \mathbb{N}} Plan_k^J$. Given a joint plan Π for coalition J and another sub-coalition $J' \subseteq J$, we can write the joint plan of coalition J' in Π as $\Pi^{J'}$, we can also write $\Pi^{-J'}$ for the joint plan of sub-coalition $J \setminus J'$.

Definition 2 (Plan Compatibility)**.** *Given $\Pi_1 \in Plan_{k_1}^J$ and $\Pi_2 \in Plan_{k_2}^{J'}$ for sub-coalition $J' \subseteq J$ and $k_2 \leq k_1$, we say that Π_1 is compatible with Π_2 if*

$$
\begin{aligned}
&\Pi_1(j)(t) = \Pi_2(j)(t)\ \forall t \in \{1, \ldots, k_2\}, j \in J' \text{ and} \\
&\Pi_1(j)(t) = skip\ \forall t \in \{k_2 + 1, \ldots, k_1\}, j \in J'
\end{aligned}
$$

Given two k-plans Π_1 and Π_2 for disjoint coalitions J_1, J_2, we write $\Pi_1 \cup \Pi_2$ (the combination of Π_1 and Π_2) for the shortest joint plan for $J_1 \cup J_2$ such that $(\Pi_1 \cup \Pi_2)^{J_1}$ is compatible with Π_1 and $(\Pi_1 \cup \Pi_2)^{J_2}$ is compatible with Π_2.

The following definition introduces the notion of history generated by a joint k-plan Π at an initial state s_0. It is the action-compatible k-history along which the agents jointly execute the plan Π starting at state s_0.

Definition 3 (History generated by a joint k-plan)**.** *Let $\gamma = (\gamma^+, \gamma^-)$ be an action theory, $s_0 \in S$ and $\Pi \in Plan_k^{Agt}$. Then, the history generated by Π from state s_0 in conformity*

with γ is the k-history $H^{\Pi,s_0,\gamma} = (H_{st}^{\Pi,s_0,\gamma}, H_{act}^{\Pi,s_0,\gamma})$ such that:

$i)$ $H^{\Pi,s_0,\gamma} \in Hist(\gamma)$,

$ii)$ $H_{st}^{\Pi,s_0,\gamma}(0) = s_0$,

$iii)$ $\forall t$ *s.t* $0 \leq t \leq k-1, \forall i \in Agt : H_{act}^{\Pi,s_0,\gamma}(i,t) = \Pi(i)(t)$.

Example 1 (Toy-sharing)**.** *Consider a childcare robot that shares a set of four toys between two children. We model this by supposing that there three are agents, called Adam, Beth and Rob (the robot). The available actions are the skip action and 24 actions*[1] *written in the form $Move(i,j,A)$, where the agent attempts to move toy A from agent i to agent j (where $i \neq j$). There are 12 atomic propositions, each written $Has(i,A)$ representing that agent i has toy A. The set of toys is written $toys$.*

The action theory γ_{toys} states that the action $Move(i,j,A)$ succeeds exactly when toy A is with agent i and no other agent attempts to take toy A from agent i. In case of such a conflict, the toy remains where it is. The full formalisation of γ_{toys} is in the supplementary material.

Consider an initial state with only two toys, both of which are held by Rob, that is $s_0 = \{Has(Rob,1), Has(Rob,2)\}$. Consider the following joint 2-plan Π_1:

Rob: $[0 \mapsto$ *Move(Rob,Adam,1)*$, 1 \mapsto$ *Move(Rob,Beth,2)*$]$,
Adam: $[0 \mapsto$ *skip*$, 1 \mapsto$ *skip*$]$, *Beth:* $[0 \mapsto$ *skip*$, 1 \mapsto$ *skip*$]$

In this plan Rob gives one toy to each child in turn, meaning that the final state will be $s_2 = \{Has(Adam,1), Has(Beth,2)\}$.

2.4 Linear Temporal Logic

In order to represent agents' values in a temporal planning situation, we introduce the language of LTL_f (Linear Temporal Logic over Finite Traces) [Pnueli, 1977; De Giacomo and Vardi, 2013; De Giacomo and Vardi, 2015], which we denote $\mathcal{L}_{\mathsf{LTL}_f}$, defined by the following grammar:

$$\varphi \quad ::= \quad p \mid do(i,a) \mid \neg\varphi \mid \varphi \wedge \varphi \mid \mathsf{X}\varphi \mid \varphi \,\mathsf{U}\, \varphi,$$

with $p \in Prop$, $i \in Agt$ and $a \in Act$. Atomic formulas in this language are those that consist of a single proposition p or a single instance of $do(i,a)$. X and U are the operators "next" and "until" of LTL_f. Operators "henceforth" (G) and "eventually" (F) are defined in the usual way: $\mathsf{G}\varphi \stackrel{\text{def}}{=} \neg(\top \,\mathsf{U}\, \varphi)$ and $\mathsf{F}\varphi \stackrel{\text{def}}{=} \neg\mathsf{G}\neg\varphi$.

Semantic interpretation of formulas in $\mathcal{L}_{\mathsf{LTL}_f}$ relative to a k-history $H \in Hist$ and a time point $t \in \{0,\ldots,k\}$ goes as follows (we omit boolean cases which are defined as usual):

$$\begin{aligned}
H,t \models & p && \iff p \in H_{st}(t),\\
H,t \models & do(i,a) && \iff t < k \text{ AND } H_{act}(i,t) = a\\
H,t \models & \mathsf{X}\varphi && \iff t < k \text{ AND } H,t+1 \models \varphi,\\
H,t \models & \varphi_1 \,\mathsf{U}\, \varphi_2 && \iff \exists t' \geq t : t' \leq k \text{ AND } H,t' \models \varphi_2\\
& && \quad \text{AND } \forall t'' \geq t : \text{ IF } t'' < t' \text{ THEN}\\
& && \quad H,t'' \models \varphi_1.
\end{aligned}$$

Given a set of $\mathcal{L}_{\mathsf{LTL}_f}$-formulas Σ, we define $Sat(\Sigma,H) = \{\varphi \in \Sigma : H,0 \models \varphi\}$ to be the set of formulas from Σ that are true in history H. Similarly, $Sat(\Sigma,\Pi,s_0,\gamma) = Sat(\Sigma,H^{\Pi,s_0,\gamma})$ for joint plan Π starting from state s_0 under the action theory γ.

2.5 Planning with Moral Agents

Moral values are not always consistent and occasionally conflict with each other [McConnell, 2022]. Furthermore, it is more serious to violate some values than others (murder is worse than lying). Therefore define the concept of a value base as an ordered sequence of sets of values (written as $\mathcal{L}_{\mathsf{LTL}_f}$-formulas) $\Omega_1, ..., \Omega_n$, with Ω_1 containing the most important values and Ω_n the least.

Definition 4 (Moral Planning Agent Problem)**.** *A moral planning agent problem (MPAP) is a tuple $\Delta = (\gamma, s_0, \overline{\Omega})$ where $\gamma = (\gamma^+, \gamma^-)$ is an action theory, s_0 is an initial state, and $\overline{\Omega} = (\Omega_1, \ldots, \Omega_m)$ is a value base with $\Omega_k \subseteq \mathcal{L}_{\mathsf{LTL}_f}$ for every $1 \leq k \leq m$.*

We assume that all agents have a single, shared value base. This value base is used to compute the *relative ideality* of histories, namely, whether a history H_1 is at least as ideal as another history H_2. Following Lorini [2021], we call *evaluation* the operation of computing an ideality ordering over plans from a value base.

Inspired by work in preference representation languages [Lang, 2004] and preference-based temporal planning [Bienvenu *et al.*, 2006; Grandi *et al.*, 2022], we define the following qualitative criterion of evaluation, noted $\preceq_\Delta^{qual}$, which compares two histories lexicographically on the basis of inclusion between sets of satisfied values.

Definition 5 (Qualitative ordering of histories)**.** *Let $\Delta = (\gamma, s_0, \overline{\Omega})$ be an MPAP with value base $\overline{\Omega} = (\Omega_1, \ldots, \Omega_m)$ and $H_1, H_2 \in Hist_k$ two γ-compatible histories. Then, $H_1 \preceq_\Delta^{qual} H_2$ if and only if:*

$i)$ $\exists n$ *s.t* $1 \leq n \leq m$ *and* $Sat(\Omega_n,H_1) \subset Sat(\Omega_n,H_2)$ *and*
$\quad \forall n'$ *if* $1 \leq n' < n$ *then* $Sat(\Omega_{n'},H_1) = Sat(\Omega_{n'},H_2)$; *or*
$ii)$ $\forall n$ *if* $1 \leq n \leq m$ *then* $Sat(\Omega_n,H_1) = Sat(\Omega_n,H_2)$

This method of evaluation compares histories by checking if either history satisfies a strict subset of the values satisfied by the other history at priority level 1. If they satisfy exactly the same set of values then we compare instead according to values of priority level 2, and so on. If at some point both histories satisfy different sets of values and neither is a subset of the other, then the two histories are incomparable according to qualitative ordering.

Since incomparability is not always desirable we can also define a notion of quantitative comparison $\preceq_\Delta^{quant}$. This is equivalent to qualitative comparison except that we compare the number of values satisfied by each history at each priority level. This ensures that any two histories will always be comparable.

2.6 From Values to LTL Formulas

So far we have said very little about either the source of these values forming a value base (whether they should be imposed from above or learned by the agent) or their nature (whether

they should be consequentialist, deontological, or something else). This is a deliberate choice, as we want our model to be able to accommodate a wide variety of different moral agents and their values (some of which may be practical or social rather than moral).

While defining a general method for formalising moral values in LTL_f would be outside the scope of this paper, we will give an example of how this could be done in practice, by formalising the value of equality.

More specifically, when confronted with situations where a finite set of objects It are distributed amongst a finite group of agents J, equality requires an equal distribution (at the end of the plan). Assuming the existence of some general ownership atom $Has(i,A)$, we can formalise such a value as $\textsc{Equality}(It,J) = \mathsf{FG}\, EqShare(It,J)$, where:

$$EqShare(It,J) \stackrel{\text{def}}{=} \bigvee_{n\in|It|} \bigwedge_{j\in J} AtLeast(It,j,n) \wedge \neg AtLeast(It,j,n+1)$$

$$AtLeast(It,j,n) \stackrel{\text{def}}{=} \bigvee_{I\subseteq It:|I|=n} \bigwedge_{A\in I} Has(j,A)$$

Example 2 (Toy sharing - continued)**.** *We suppose that the shared value base $\overline{\Omega}_{toys}$ contains three values. Firstly, both children must be given at least one toy (*subsistence*). Secondly, nobody should take a toy away from someone else, though they may give away their own toys (*property*). Finally, both children should have the same number of toys (*equality*). The formalisation of these values is as follows (where Ω_1 is* subsistence*, Ω_2 is* property *and Ω_3 is* equality*):*

$$\Omega_1 = \big\{\mathsf{FG}(AtLeast(toys, Adam, 1)), \mathsf{FG}(AtLeast(toys, Beth, 1))\big\}$$

$$\Omega_2 = \big\{\mathsf{G}(\bigwedge_{i\neq j, j\neq l, A\in[1,4]} \neg do(i, Move(j,l,A)))\big\}$$

$$\Omega_3 = \big\{\mathsf{FG}\, EqShare(toys, \{Adam, Beth\})\big\}$$

Note the different temporal characteristics of these values. Substistence *and* equality *both describe a state that must be satisfied at the end of the plan, whereas* property *must be satisfied at all points in the plan. Note also how our model is able to handle both consequentialist values (*subsistence *and* equality *both describe states of affairs that must hold at some point) and deontological values (*property *forbids agents from taking certain kinds of actions).*

3 Evaluating Joint Plans

Once we fix a moral agent planning problem Δ, every joint plan Π induces a single generated history $H^{\Pi,s_0,\gamma}$ (recall Definition 3). Thus, we can lift Definition 5 of comparison among histories to comparison among joint plans.

Definition 6. *Let $\Delta = (\gamma, s_0, \overline{\Omega})$ be an MPAP and let $\Pi_1, \Pi_2 \in Plan^{Agt}$ be two joint plans for Agt. Then, $\Pi_1 \preceq^{qual}_{\Delta} \Pi_2$ if and only if $H^{\Pi_1,s_0,\gamma} \preceq^{qual}_{\Delta} H^{\Pi_2,s_0,\gamma}$. Similarly, $\Pi_1 \preceq^{quant}_{\Delta} \Pi_2$ if and only if $H^{\Pi_1,s_0,\gamma} \preceq^{quant}_{\Delta} H^{\Pi_2,s_0,\gamma}$.*

Given an MPAP $\Delta = (\gamma, s_0, \overline{\Omega})$ and a joint plan $\Pi \in Plan^{Agt}$ we say that Π is *k-non-dominated* for Δ if there is no plan Π' of length at most k such that $\Pi \preceq^{qual}_{\Delta} \Pi'$ and $\Pi' \not\preceq^{qual}_{\Delta} \Pi$ (an equivalent definition can be given using $\preceq^{quant}_{\Delta}$).

In conclusion, comparing joint plans with reference to a value base is conceptually equivalent to the problem of comparing two histories. Thus, we can import the analysis and the computational complexity results of Grandi et al. [2022], who showed that the problem of comparing single-agent plans (and hence histories) can be solved in polynomial time, while the problem of testing whether a given plan or history is non-dominated is PSPACE-complete.

Example 3 (Toy sharing - continued)**.** *Consider again the initial state with only two toys, both of which are held by Rob, $s_0 = \{Has(Rob,1), Has(Rob,2)\}$. In this instance there are several joint plans that satisfy all values (and therefore are, by necessity, non-dominated). The first is the plan that we introduced before:*

$$\textit{Rob: } [0 \mapsto \textit{Move(Rob,Adam,1)}, 1 \mapsto \textit{Move(Rob,Beth,2)}],$$
$$\textit{Adam: } [0 \mapsto \textit{skip}, 1 \mapsto \textit{skip}], \textit{Beth: } [0 \mapsto \textit{skip}, 1 \mapsto \textit{skip}]$$

It is straightforward to see that this plan satisfies the values of subsistence*,* property *and* equality*.*

Consider now a different initial state with only one toy, held by Rob ($s_0 = \{Has(Rob,1)\}$). In this case there is no plan satisfying all values, as our subsistence *values conflict with our* equality *value. A non-dominated plan will therefore be one that prioritises values according to the lexicographic ordering. In this case, the best Rob can do is to satisfy the* property *value and one of the* subsistence *values:*

$$\Pi_2 = (\textit{Rob: } [0 \mapsto \textit{Move(Rob,Beth,1)}), \textit{Adam: } [0 \mapsto \textit{skip}], \textit{Beth: } [0 \mapsto \textit{skip}])$$

However, if we change the value base by swapping the contents of Ω_1 and Ω_3, then Π_2 is dominated by the following plan, which would itself be non-dominated (as now the best option for Rob is to keep the toy and satisfy equality*):*

$$\textit{Rob: } [0 \mapsto \textit{skip}], \textit{Adam: } [0 \mapsto \textit{skip}], \textit{Beth: } [0 \mapsto \textit{skip}]$$

4 Evaluating Individual Plans

Evaluating joint plans is helpful whenever agents are able to coordinate, for example in presence of a central planner. Otherwise, agents may not have any knowledge about the other agents' plans. Meaning that their individual plans may be part of many possible joint plans, meaning many possible histories and many possible sets of satisfied values.

4.1 Optimistic and Pessimistic Comparison

Two intuitive ways of comparing individual plans under complete ignorance of the other agents' plans are by comparing according to the best-case outcome of that individual plan (risk-seeking) or the worst-case outcome of that individual plan (risk-averse). In line with Lorini [2015], we call these methods of comparison "optimistic" and "pessimistic".

Definition 7 (Optimistic k-comparison)**.** *Let $\Delta = (\gamma, s_0, \overline{\Omega})$ be an MPAP, $i \in Agt$, and Π_1, Π_2 be individual plans. Let k be an integer greater than or equal to the length of both Π_1 and Π_2. Then, $\Pi_1 \leq^{opt}_{i,k} \Pi_2$ iff $\exists\, \Pi'_2 \in Plan^{Agt}_k$ s.t. Π'_2 is*

compatible with Π_2 *and* $\forall\, \Pi_1' \in Plan_k^{Agt}$ *s.t* Π_1' *is compatible with* Π_1 *we have* $\Pi_1' \preceq_\Delta^{quant} \Pi_2'$.

Definition 8 (Pessimistic k-comparison)**.** *Let* $\Delta = (\gamma, s_0, \overline{\Omega})$ *be an MPAP,* $i \in Agt$*, and* Π_1, Π_2 *be individual plans. Let* k *be an integer greater than or equal to the length of both* Π_1 *and* Π_2*. Then,* $\Pi_1 \leq_{i,k}^{pess} \Pi_2$ *iff* $\exists \Pi_1' \in Plan_k^{Agt}$ *s.t* Π_1' *is compatible with* Π_1 *and* $\forall \Pi_2' \in Plan_k^{Agt}$ *s.t* Π_2' *is compatible with* Π_2 *we have* $\Pi_1' \preceq_\Delta^{quant} \Pi_2'$.

The reading of the plan comparison statement $\Pi_1 \leq_{i,k}^{opt} \Pi_2$/$\Pi_1 \leq_{i,k}^{pess} \Pi_2$ is "agent i's k-plan Π_2 is *optimistically/pessimistically* at least as ideal as agent i's k-plan Π_1". We use the standard notation of writing $\Pi_1 <_{i,k}^{pess} \Pi_2$ for $\Pi_1 \leq_{i,k}^{pess} \Pi_2$ and $\Pi_2 \not\leq_{i,k}^{pess} \Pi_1$, and $\Pi_1 \approx_{i,k}^{pess} \Pi_2$ if $\Pi_1 \leq_{i,k}^{pess} \Pi_2$ and $\Pi_2 \leq_{i,k}^{pess} \Pi_1$. The notion of plan comparison allows us to define the corresponding notion of non-dominated plan.

Definition 9 (Optimistic k-non-dominance)**.** *Let* $\Delta = (\gamma, s_0, \overline{\Omega})$ *be an MPAP,* $i \in Agt$*, and* $\Pi_1 \in Plan_k$*. Then, we say that agent* i*'s plan* Π_1 *is* k*-non dominated iff* $\nexists \Pi_2 \in Plan_k$ *such that* $\Pi_1 <_{i,k}^{opt} \Pi_2$.

Given an agent $i \in Agt$, we say that i is *optimistic-rational* if i only ever selects plans that are optimistic non-dominated. We define the notions of "pessimistic k-non-dominance" and "pessimistic-rationality" in an equivalent way.

Note that we have used quantitative comparison in our definition in order to simplify later proofs, but we could alternatively use qualitative comparison. Qualitative comparison does not guarantee the existence of a single history that is weakly preferred to all other histories, hence the slightly more-complex-than-usual phrasing of optimistic and pessimistic comparison.

Example 4 (Toy Sharing - continued)**.** *Consider the plans available to Rob using pessimistic comparison. While there are many joint plans that satisfy all values, guaranteeing the satisfaction of values in the individual planning case is much harder. This is because if we fix the actions of Rob, Adam and Beth can always "coordinate" their actions to cause a worst-possible outcome. This can be done by having Adam "block" all of Rob's moves by always making conflicting moves, while Beth moves toys in a way that violates as many values as possible. This means pessimistic comparison does not discriminate in this planning domain.*

On the other hand, optimistic comparison does discriminate between plans, since if Rob chooses a plan that does not violate property*,* property *will not be violated in the best-case outcome, but if Rob chooses a plan that does violate* property*, then* property *will be violated even in the best-case outcome. Moreover, it can be seen that if all agents are optimistic-rational,* property *will not be violated.*

The example above suggests that optimistic comparison generally discriminates more than pessimistic comparison. However, we now define two classes of MPAP where both notions cannot discriminate among plans.

Definition 10 (k-Resilient Domains)**.** *Let* $\Delta = (\gamma, s_0, \overline{\Omega})$ *be an MPAP, we say that* Δ *is* k-resilient *iff* $\forall i \in Agt, \Pi \in Plan_k^{\{i\}}$, $\exists \Pi' \in Plan_k^{Agt}$ *where* Π' *is compatible with* Π *and* Π' *is a* k*-non-dominated joint plan according to* $\preceq_\Delta^{quant}$.

Definition 11 (k-Fragile Domains)**.** *Given an MPAP* $\Delta = (\gamma, s_0, \overline{\Omega})$*, we say that* Δ *is* k-fragile *iff* $\forall i \in Agt, \Pi \in Plan_k^{\{i\}}$, $\exists \Pi' \in Plan_k^{Agt}$ *where* Π' *is compatible with* Π *and* Π' *is* k*-maximally-dominated according to* $\preceq_\Delta^{quant}$ *(there is no plan* Π'' *such that* $\Pi'' \prec_\Delta^{quant} \Pi'$*).*

The following two theorems show that optimistic comparison cannot discriminate between plans in a resilient domain, and pessimistic comparison cannot discriminate between plans in a fragile domain. Full proofs can be found in the supplementary material.

Theorem 1. *Given a* k*-resilient MPAP* $\Delta = (\gamma, s_0, \overline{\Omega})$*, an agent* $i \in Agt$ *and two individual plans* Π_1 *and* Π_2*, we have that* $\Pi_1 \approx_{i,k}^{opt} \Pi_2$.

Theorem 2. *Given a* k*-fragile MPAP* $\Delta = (\gamma, s_0, \overline{\Omega})$*, an agent* $i \in Agt$ *and two individual plans* Π_1 *and* Π_2*, we have that* $\Pi_1 \approx_{i,k}^{pess} \Pi_2$.

The intuition behind these proofs is that in a resilient domain, no individual plan will optimistic-dominate any other because all plans are optimistic non-dominated (since all best-case outcomes are equally good). An equivalent issue occurs for pessimistic-rational agent in a fragile domain. This is also why pessimistic comparison is indifferent in Example 4, as Δ_{toys} is a k-fragile domain for any $k \in \mathbb{N}$ in any initial state where all toys belong to Rob.

4.2 Anticipating Blameworthiness

The previous section illustrates the need for another method of plan comparison to complement optimistic or pessimistic comparison. A moral agent would care not just about values being satisfied, but also that they are not *blameworthy* for values being violated. This is something that we aim to introduce to our model, by defining the following notion of blameworthiness. First, given a set of $\mathcal{L}_{\mathsf{LTL}_f}$-formulas Σ, let us denote with $Viol(\Sigma,\Pi,s_0,\gamma)$ the set $\Sigma \setminus Sat(\Sigma,\Pi,s_0,\gamma)$ of those formulas in Σ that are not satisfied by the history generated by plan Π from state s_0 according to action theory γ.

Definition 12 (Anticipated Passive Blameworthiness with k-Horizon)**.** *Let* k *be a positive integer. Let* $MPAP = (\gamma, s_0, \overline{\Omega})$ *be an MPP,* $i \in Agt$ *an agent, and* $\Pi \in Plan_{\{i\}}^{l}$ *an individual plan where* $l \leq k$*. Let* $\omega \in \overline{\Omega}$*. Then, we say that* i *is P-*k*-blameworthy for* ω *in* Π *if there exists some joint plan* $\Pi_1 \in Plan_{Agt\setminus\{i\}}^{l_1}$ *such that* $l \leq l_1 \leq k$*,* $H^{\Pi \cup \Pi_1, s_0, \gamma} \not\models \omega$*, and there exists some joint plan* $\Pi_2 \in Plan_{\{i\}}^{l_2}$ *such that* $l_1 \leq l_2 \leq k$ *and* $H^{\Pi_2 \cup \Pi_1, s_0, \gamma} \models \omega$.

Definition 13 (Anticipated Active Blameworthiness with k-Horizon)**.** *Let* k *be a positive integer. Let* $MPAP = (\gamma, s_0, \overline{\Omega})$ *be an MPP,* $i \in Agt$ *an agent, and* $\Pi \in Plan_l^i$ *an individual plan where* $l \leq k$*. Let* $\omega \in \overline{\Omega}$*. Then, we say that* i *is A-*k*-blameworthy for* ω *in* Π *if there exists some joint plan* $\Pi_1 \in Plan_{l_1}^{Agt}$ *such that* $l_1 \leq k$*,* $H^{\Pi_1, s_0, \gamma} \models \omega$*, and for all joint plans* $\Pi_2 \in Plan_{l_2}^{Agt}$ *such that* $l \leq l_2 \leq k$ *and* Π_2 *is compatible with* Π*,* $H^{\Pi_2, s_0, \gamma} \not\models \omega$.

In line with previous work [Lorini and Mühlenbernd, 2018; Lorini *et al.*, 2014], we assume blameworthiness has two components: the causal responsibility component and the value violation component. This means that an agent is blameworthy for any value that it is causally responsible for violating. The notions of causal responsibility we consider are both the *passive* sense (i.e. given the other agents' choices, the agent could have prevented the value from being violated by making a different choice) and the *active* sense (i.e. given the agent's choices, the other agents could not prevent the value from being violated by making different choices). For more on the distinction between *passive* and *active* responsibility we refer to Lorini et al. [2014]. Note that our definition is for anticipated responsibility, meaning that i *may* be responsible if they perform plan Π. This only really affects passive responsibility, since if an agent is actively responsible the actions of all other agents are irrelevant.

We can also define equivalent notions of active and passive praiseworthiness, meaning that an agent is actively praiseworthy for any value that they guaranteed the satisfaction of, and passively praiseworthy for any value that was satisfied but which they could have violated.

Given an MPAP $\Delta = (\gamma, s_0, \overline{\Omega})$, an agent $i \in Agt$ and an individual plan Π and an integer k greater than or equal to the length of Π, we define $ABW(i, \Pi, k, \Delta)$ as the set of all values $\omega \in \overline{\Omega}$ such that agent i is A-blameworthy for ω in Π (where Δ is obvious from context, it is omitted). We define $PBW(i, \Pi, k, \Delta)$ as composed of all values $\omega \in \overline{\Omega}$ such that agent i is P-blameworthy for ω in Π. Equivalently, we can define $APW(i, \Pi, k, \Delta)$ and $PPW(i, \Pi, k, \Delta)$ for praiseworthiness.

Example 5 (Uses of Responsibility)**.** *A responsibility-conscious agent could allow us to improve on both optimistic and pessimistic comparison. For optimistic comparison, consider a simple example with multiple agents with the goal of emptying a bin. This is acheived by some agent performing the* $emptybin$ *action at some point in their plan (let plan* Π_1 *be a sequence of skip actions and* Π_2 *the same but with an* $emptybin$ *action at the beginning). This is then a* k*-resilient domain (for all* k*) since the best-case outcome for all individual plans is the same. Therefore an optimistic-rational agent could not decide between* Π_1 *and* Π_2*.*

	Passive	*Active*
Praiseworthy	Π_2	Π_2
Blameworthy	Π_1	

However, the table above shows that by emptying the bin, the agent will be actively and passively praiseworthy for the bin being emptied, whereas by not emptying it, they will be passively blameworthy for not doing so. Therefore a responsibility-conscious agent would choose to empty the bin. For pessimistic comparison, consider again Rob from the toy-sharing example. Let Π_1 *be some plan containing an action that violates* property*, and* Π_2 *a plan that does not contain such an action. Then consider the following table, illustrating responsibility for the* property *value:*

	Passive	*Active*
Praiseworthy	Π_2	
Blameworthy	Π_1	Π_1

Therefore by choosing plan Π_2 *Rob can ensure that they will not be blameworthy for the violation of* property *(even if the value is violated by another agent).*

5 Complexity Results

This section provides an analysis of the computational complexity of optimistic and pessimistic comparison, as well as the notions of blameworthiness introduced in Section 4.2. These tasks are PSPACE-complete, in line with the complexity of classical problems in planning or the model checking of $\mathcal{L}_{\mathsf{LTL}_f}$-formulas. Full proofs of our results can be found in the supplementary material.

Define OPT-k-COMPARISON (respectively PESS-k-COMPARISON) as the decision problem that takes as input an MPAP $\Delta = (\gamma, s_0, \overline{\Omega})$, an integer k, and two individual plans Π_1 and Π_2 for some agent i, and asks whether $\Pi_1 \leq^{opt}_{i,k} \Pi_2$ (resp. $\Pi_1 \leq^{pess}_{i,k} \Pi_2$). Membership in PSPACE is straightforward, given that we can generate and compare joint plans two-by-two. With a reduction from the classical PSPACE-complete problem PLANMIN [Bylander, 1994] we can show the following:

Theorem 3. OPT-K-COMPARISON *and* PESS-K-COMPARISON *are both PSPACE-complete.*

For responsibility, we consider the problem of recognising whether an agent is blameworthy for a given value. Let A-BLAME (respectively P-BLAME) be the decision problem where given an individual plan Π for some agent i, k greater than or equal to the length of Π and a formula $\varphi \in \cup\overline{\Omega}$ in an MPAP $\Delta = (\gamma, s_0, \overline{\Omega})$, we must decide if $\varphi \in ABW(i, \Pi, k, \Delta)$ (respectively $\varphi \in PBW(i, \Pi, k, \Delta)$). Again, membership to PSPACE is straightforward and a reduction from PLANMIN shows the following:

Theorem 4. A-BLAME *and* P-BLAME *are both PSPACE-complete.*

While we did not introduce any notion of comparison based on anticipated responsibility, our Theorem 4 shows that any reasonable such definition will be PSPACE-hard to use.

6 Related Work

Our work is a contribution to the field of ethical planning in AI [Dennis *et al.*, 2016; Lindner *et al.*, 2019], and also to the larger body of research on planning with preferences [Baier and McIlraith, 2008; Juma *et al.*, 2012] and temporally extended goals [Bienvenu *et al.*, 2006; De Giacomo and Vardi, 2015; Camacho *et al.*, 2017].

Of these, the work of Bienvenu et al. [2006] and Dennis et al. [2016] are the two most closely related to our work. Bienvenu et al. [2006] present a model for planning with temporal preferences, represented using various combinations of LTL formulas. While this model does allow lexicographic combinations of values, it cannot efficiently model our concept of a lexicographic value base. Also, unlike our work, this model does not allow for incomparability between plans, which can be useful when dealing with ethical dilemmas. Dennis et al. [2016] focus instead on the verification of autonomous agents in relation to their capacity to select ethical plans. This paper

models ethical values in propositional logic, and uses a total (not necessarily strict) order on values that is functionally equivalent to our notion of quantitative comparison. To evaluate plans, this model uses ethical rules which map specific actions (or their absence) to the violation of specific values in a given context. While our representation of value bases is equivalent, their model allows the tracking of repeated violations of values, however it is limited to propositional values that can be attributed to single actions, whereas we have the full expressivity of LTL_f. Finally, both papers are limited to single-agent planning applications.

However, some approaches to multi-value LTL planning do consider multiple agents. For example, Guo and Dimarogonas [2015] present a model for motion planning in a cooperative setting, where each agent has an independent goal represented in LTL as a series of hard and soft constraints. This approach differs from ours in that it is specialised towards motion planning. While it can model potentially conflicting values through its use of soft constraints, it has no way of indicating the relative priority of constraints. On the other hand, it includes partial information and communication between agents, which suggests useful additions to our work.

A popular approach to managing relative levels of importance in multi-value planning is to assign numerical weights to values. Juma et al. [2012] explore solving Preference Based Planning problems (a form of multi-objective planning) via partial weighted MaxSAT techniques. Partial weighted MaxSAT has both hard constraints and soft constraints (which have weighted penalties for not satisfying). This approach can easily handle multiple conflicting goals, but we chose to avoid it in favour of a lexicographic approach as we wanted to guarantee that our most important values (such as safety) could not be overridden by a multitude of lesser values (such as politeness). Furthermore, this paper is only able to express preferences as partial sets of assignments to variables, rather than our model which uses LTL_f.

Similar approaches to our work can also be found in the field of agent programming. Agent programming is similar to planning in that agents have a set of available actions and a goal to achieve. The difference is that rather than finding a plan to achieve the goal, agents follow the rules of some program. This is computationally much more tractable than planning, but does not guarantee that the agents actions will be ideal, which is why we did not use this approach. Hindriks and Birna van Riemsdijk [2008] use (limited) look-ahead and soft and hard constraints to further restrict available actions. The soft constraints are strictly ordered in a lexicographic fashion (unlike our model where the ordering can be non-strict) and hard constraints simply must not be violated.

The complexity analysis we perform is in line with the work of Lindner et al. [2019] which reaches similar conclusions concerning consequentialist plan comparison. In this work, the authors evaluate the permissibility or impermissibility of various single-agent single-goal plans according to various ethical principles. They evaluate various ethical theories including deontology, consequentialism and the principle of double effect under a numerical representation of values on actions and properties of the environment.

Our work on responsibility is part of the larger field of Responsible AI, as outlined by Dignum [2020]. She identifies three main branches of responsible AI: developing legislation for determining responsibility in cases involving AI, ensuring AI compliance with human ethics and values, and ensuring widespread accessibility and participation in the development and use of AI. Our work focuses on the second branch.

The idea of combining notions responsibility and blameworthiness in multi-agent planning has also been explored in a recent work by Alechina et al. [2017], which grounds on previous work by [Chockler and Halpern, 2004] on the formalization of responsibility (see also [Halpern and Kleiman-Weiner, 2018]). Rather than using LTL, as in our approach, this work uses structural equation modeling (SEM). Also, while our work focuses on responsibility as a method of plan evaluation (forward-looking), this paper considers responsibility retrospectively, and assigns responsibility for the failure of a goal (backward-looking). This model can assign different degrees of responsibility to different agents, which ours cannot, but it is limited to planning problems with a single goal, and there is no connection to ethics.

Our model is grounded on previous work in ethical planning and moral reasoning. First, our representation of values and their interpretation on histories is based on the work of Grandi et al. [2022]. Second, our notions of dominance between multi-agent plans is inspired by Lorini [2015], who however uses a numerical representation of values.

7 Conclusion

In this paper we outlined our model of moral planning agents with temporal values, equipped with novel criteria to evaluate their plans. We showed, with the help of a running example, that optimistic and pessimistic comparison can be indecisive in ethically sensitive situations. We introduced new criteria based on responsibility, and showed that they overcome this limitation. Finally, we showed that comparing plans in our model is PSPACE-complete.

Our intention is to create a model that can be used by autonomous moral agents to produce ethically-acceptable plans. To this end, we would like to expand the expressiveness of our model and drop some of the simplifying assumptions made in this paper. At present, our model assumes that all agents have the same value base, it also assumes perfect information and fully deterministic actions, which will not always be a practical assumption. In line with Dietrich and List [2017], we intend to relax these assumptions in future work.

Furthermore, while we have modelled a causal form of responsibility (where the actions of agents causally contribute to an outcome) it would be interesting to consider indirect notions of responsibility as studied by Lorini and Sartor [2016]. We would also like to extend LTL_f to include epistemic operators for applications to AI such as ChatGPT, whose ethical violations (misleading users, revealing sensitive information) are mostly epistemic in nature.

Last but not least, it would also be interesting to refine our methods for plan comparison, including developing methods that make use of dominance such as those proposed by Horty [2001]. This could be used to create real-world implementations of our model to demonstrate its capability.

Ethical Statement

There are no ethical issues.

Acknowledgments

Support from the ANR project CoPains "Cognitive Planning in Persuasive Multimodal Communication" (grant number ANR-18-CE33-0012) and the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANITI) is gratefully acknowledged.

References

[Abelson, 1979] R. Abelson. Difference between belief and knowledge systems. *Cognitive Science*, 3, 1979.

[Ajmeri *et al.*, 2020] N. Ajmeri, H. Guo, P. Murukannaiah, and M. Singh. Elessar: Ethics in norm-aware agents. In *Proceedings of the 19th Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2020.

[Alechina *et al.*, 2017] N. Alechina, J. Halpern, and B. Logan. Causality, responsibility and blame in team plans. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, (AAMAS)*, 2017.

[Arkin *et al.*, 2012] R. Arkin, P. Ulam, and A. Wagner. Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE*, 100(3):571–589, 2012.

[Baier and McIlraith, 2008] J. Baier and S. McIlraith. Planning with preferences. *AI Mag.*, 29, 2008.

[Bienvenu *et al.*, 2006] M. Bienvenu, C. Fritz, and S. McIlraith. Planning with qualitative temporal preferences. In *Proceedings of the 10th Conference on Principles of Knowledge Representation and Reasoning, (KR)*, 2006.

[Brafman and Tennenholtz, 2018] R. Brafman and M. Tennenholtz. An axiomatic treatment of three qualitative decision criteria. *Journal of the ACM*, 47(3):452–482, 2018.

[Braham and van Hees, 2012] M. Braham and M. van Hees. An anatomy of moral responsibility. *Mind*, 121(483):601–634, 2012.

[Bylander, 1994] T. Bylander. The computational complexity of propositional STRIPS planning. *Artificial Intelligence*, 69(1-2):165–204, 1994.

[Camacho *et al.*, 2017] A. Camacho, E. Triantafillou, C. Muise, J. Baier, and S. McIlraith. Non-deterministic planning with temporally extended goals: LTL over finite and infinite traces. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017.

[Chockler and Halpern, 2004] H. Chockler and J. Halpern. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22:93–115, 2004.

[Cranefield *et al.*, 2017] S. Cranefield, M. Winikoff, V. Dignum, and F. Dignum. No pizza for you: Value-based plan selection in BDI agents. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[De Giacomo and Vardi, 2013] G. De Giacomo and M. Vardi. Linear temporal logic and linear dynamic logic on finite traces. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

[De Giacomo and Vardi, 2015] G. De Giacomo and M. Vardi. Synthesis for LTL and LDL on finite traces. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

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

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

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

[Dignum, 2020] V. Dignum. Responsibility and Artificial Intelligence. In *The Oxford Handbook of Ethics of AI*. Oxford University Press, 2020.

[Gollwitzer, 1996] P. Gollwitzer. The volitional benefits of planning. In *The psychology of action*. Guilford Press, 1996.

[Grandi *et al.*, 2022] Umberto Grandi, Emiliano Lorini, Timothy Parker, and Rachid Alami. Logic-based ethical planning. In *Proceedings of the 21st International Conference of the Italian Association for Artificial Intelligence (AIxIA)*, 2022.

[Guo and Dimarogonas, 2015] M. Guo and D. V. Dimarogonas. Multi-agent plan reconfiguration under local LTL specifications. *The International Journal of Robotics Research*, 34(2):218–235, 2015.

[Halpern and Kleiman-Weiner, 2018] J. Halpern and M. Kleiman-Weiner. Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

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

[Juma *et al.*, 2012] F. Juma, E. Hsu, and S. McIlraith. Preference-based planning via maxsat. In *Proceedings of the 25th Canadian Conference on Artificial Intelligence*, 2012.

[Lang, 2004] J. Lang. Logical preference representation and combinatorial vote. *Annals of Mathematics and Artificial Intelligence*, 42(1-3):37–71, 2004.

[Lindner *et al.*, 2019] F. Lindner, R. Mattmüller, and B. Nebel. Moral permissibility of action plans. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 2019.

[Lorini and Mühlenbernd, 2015] E. Lorini and R. Mühlenbernd. The long-term benefits of following fairness norms: A game-theoretic analysis. In *Proceedings of the 18th International Conference on Principles and Practice of Multi-Agent Systems (PRIMA)*, 2015.

[Lorini and Mühlenbernd, 2018] E. Lorini and R. Mühlenbernd. The long-term benefits of following fairness norms under dynamics of learning and evolution. *Fundamenta Informaticae*, 158(1-3):121–148, 2018.

[Lorini and Sartor, 2016] E. Lorini and G. Sartor. A STIT logic for reasoning about social influence. *Studia Logica*, 104(4):773–812, 2016.

[Lorini *et al.*, 2014] E. Lorini, D. Longin, and E. Mayor. A logical analysis of responsibility attribution: emotions, individuals and collectives. *Journal of Logic and Computation*, 24(6):1313–1339, 2014.

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

[Lorini, 2021] E. Lorini. A logic of evaluation. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2021.

[McConnell, 2022] T. McConnell. Moral Dilemmas. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2022.

[Miceli and Castelfranchi, 2000] M. Miceli and C. Castelfranchi. The role of evaluation in cognition and social interaction. In *Advances in consciousness research. Human cognition and social agent technology*. John Benjamins Publishing Company, 2000.

[Moors *et al.*, 2013] A. Moors, P. Ellsworth, K. Scherer, and N. Frijda. Appraisal theories of emotion: State of the art and future development. *Emotion Review*, 5(2):119–124, 2013.

[Pnueli, 1977] A. Pnueli. The temporal logic of programs. In *Proceedings of the 18th Annual Symposium on Foundations of Computer Science (FOCS)*, 1977.

[Reiter, 2001] R. Reiter. *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*. The MIT Press, 2001.

[Rodriguez-Soto *et al.*, 2020] M. Rodriguez-Soto, M. López-Sánchez, and J. Rodríguez-Aguilar. A structural solution to sequential moral dilemmas. In *Proceedings of the 19th Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2020.

[Sen, 1985] A. Sen. Well-being, agency and freedom: The dewey lectures 1984. *The Journal of Philosophy*, 82(4):169–221, 1985.

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

[Serramia *et al.*, 2018] M. Serramia, M. López-Sánchez, J. Rodríguez-Aguilar, M. Rodríguez, M. Wooldridge, J. Morales, and C. Ansótegui. Moral values in norm decision making. In *Proceedings of the 17th Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2018.

[van Hindriks and van Riemsdijk, 2008] K. van Hindriks and M. Birna van Riemsdijk. Using temporal logic to integrate goals and qualitative preferences into agent programming. In *Declarative Agent Languages and Technologies, 6th International Workshop, DALT*, 2008.

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