Uncertain Machine Ethics Planning

Simon Kolker
University of Manchester
Manchester, United Kingdom
simon.kolker@manchester.ac.uk

Louise A. Dennis
University of Manchester
Manchester, United Kingdom
louise.dennis@manchester.ac.uk

Ramon Fraga Pereira
University of Manchester
Manchester, United Kingdom
ramon.fragapereira@manchester.ac.uk

Mengwei Xu
Newcastle University
Newcastle, United Kingdom
mengwei.xu@newcastle.ac.uk

ABSTRACT

Machine Ethics decisions should consider the implications of uncertainty over decisions. Decisions should be made over sequences of actions to reach preferable outcomes long term. The evaluation of outcomes, however, may invoke one or more moral theories, which might have conflicting judgements. Each theory will require differing representations of the ethical situation. For example, *Utilitarianism* measures numerical values, *Deontology* analyses duties, and *Virtue Ethics* emphasises moral character. While balancing potentially conflicting moral considerations, decisions may need to be made, for example, to achieve morally neutral goals with minimal costs. In this paper, we formalise the problem as a *Multi-Moral Markov Decision Process* and a *Multi-Moral Stochastic Shortest Path Problem*. We develop a heuristic algorithm based on *Multi-Objective AO**, utilising Sven-Ove Hansson's *Hypothetical Retrospection* procedure for ethical reasoning under uncertainty. Our approach is validated by a case study from Machine Ethics literature: the problem of whether to steal insulin for someone who needs it.

KEYWORDS

Machine Ethics; Multi-Objective Planning; Uncertainty

1 INTRODUCTION

As our lives are increasingly impacted by the decisions of machines [39], one objective of *Machine Ethics* is to develop computational techniques to incorporate ethical behaviour with decision-making, preserving human values into the future [1]. A major concern for Machine Ethics comes from situations with uncertainty over outcomes, commonplace in the real world. Most existing approaches either assume actions are deterministic or they use expected utility maximisation [38]. This can yield disagreeable results (see Section 3 and [10]), demanding a more sophisticated handling
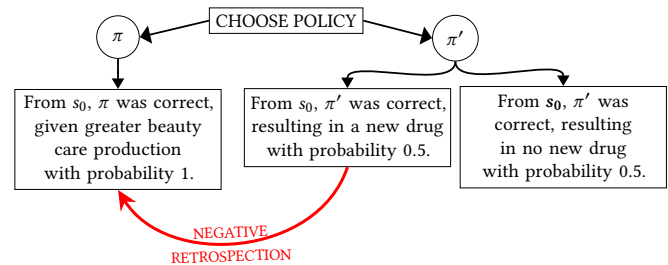
Figure 1: A company may produce a beauty care product or research a drug with 50% probability of saving lives. Beauty care has negative retrospection (like regret) for missing the chance to save lives; if research is unsuccessful, there is no negative retrospection since risk was accepted at decision-time. Arguments are generated from each outcome. Directed edges are attacks indicating negative retrospection.

of *outcome uncertainty*. Decisions are further complicated by *moral uncertainty* [18]. Philosophy has no universal consensus over which decisions are correct in all situations: moral theories may conflict, leading to *moral dilemmas*. While we defer the choice of theories to *stakeholders*, who must approve a system before deployment (owners, regulators, or affected individuals), there are likely to be conflicts or uncertainty between stakeholder approved theories. Implementations must handle these situations gracefully.

Another issue is the integration of ethical behaviour into existing systems with non-moral goals. For example, an autonomous vehicle must balance risk to other road users with reaching its destination.

As the *Automated Planning* community demonstrates, many environments require sequences of actions (action plans or policies) to achieve desired goals and the most rewarding outcomes [19]. Machine Ethics implementations, therefore, should consider future decisions, rather than selecting immediately preferred actions greedily as is the case in most existing systems [38].

Hypothetical Retrospection is an ethical procedure for handing outcome uncertainty [22]. It suggests we judge decisions by the circumstances under which they were made. In [25], *Machine Ethics Hypothetical Retrospection* (MEHR) is formalised as a framework for single action decisions with outcome uncertainty and multiple moral theories under moral uncertainty. Arguments are generated in support of each action from the hypothetical perspective of their outcomes. An argument is attacked if there is negative retrospection

for missing an ethically preferred outcome. An action with minimal attacks on its supporting arguments is selected. An example is in Figure 1, details in Section 2.

In this paper, we formalise uncertain ethical planning problems as a variant of the *Markov Decision Process* (MDP). Our *Multi-Moral Markov Decision Process* (MMMDP) may contain a number of potentially conflicting moral theories. Stakeholders may express preference or moral uncertainty over theories with a *weak lexicographic ordering*. A theory's space of morally relevant information is general and inclusive to many ethical perspectives. We extend our formalism to the *Multi-Moral Stochastic Shortest Path Problem* (MMSSP) for non-moral situations. Preferable solutions act ethically subject to goals and costs. We produce an offline heuristic planning algorithm to solve these problems based on AO* [31], Multi-Objective heuristic planning [11] and MEHR [25]. We evaluate the algorithm with a case study from Ethics and Machine Ethics literature: the problem of whether to steal insulin for someone who needs it [2].

2 BACKGROUND

Markov Decision Process. Uncertain decision making problems can be described as a *Finite Horizon Markov Decision Process* (MDP). We represent them as a tuple $\langle S, A, P, R, H \rangle$, where S is a finite set of states, and A is a finite set of actions. The state space is traversed through discrete sequential transitions. The probabilistic transition function $P : S \times A \times S \rightarrow [0, 1]$ defines the probability one state transitions to another via an action. Transitions are valued by a reward function $R : S \times A \times S \rightarrow \mathbb{R}$. An optimal solution is a non-stationary policy $\pi : S \times \{0, ..., H - 1\} \rightarrow A$ that maps a state and timestamp to an action with maximum expected reward. The horizon $H \in \mathbb{N}$ is the number of state transitions before the problem terminates. Policies are evaluated with a non-stationary value function defined recursively: $V^{\pi}(s, t) = \sum_{s' \in S} P(s, \pi(s, t), s')[R(s, \pi(s, t), s') + V^{\pi}(s', t + 1)]$. By Bellman's optimality principle [5], if a policy's value can be measured by its expected additive utility, then there exists a policy π^* optimal at all sates and it satisfies the *Bellman optimality equation* $V^{\pi^*}(s, t) = \max_{a \in A} \sum_{s' \in S} P(s, a, s')[R(s, a, s') + V^{\pi^*}(s', t+1)]$. *Stochastic Shortest Path* (SSP) problems are MDPs with goal states $G \subseteq S$ that are terminal, meaning there are no transitions away from them: $\forall_{s \in G, a \in A}, P(s, a, s) = 1 \land R(s, a, s) = 0$; all other rewards are strictly negative. SSPs also have an initial state s_0. Optimal policies are *proper* if there is 1.0 probability of reaching a goal from s_0. Improper policies have an expected cost of ∞.

MDP Algorithms. MDPs can be solved using *Dynamic Programming* algorithms, e.g. *Policy Iteration* [7, 23] or *Value Iteration* (VI) [7]. VI refines an arbitrary value function V_0 with *Bellman backups*: iterative applications of the Bellman optimality equation [5]. Values propagate until the maximum absolute difference between values in V_i and V_{i+1} (the *residual*) is below a constant $\epsilon > 0$ (ϵ-consistency). This requires *full sweeps* across the state space. *Heuristic* algorithms (RTDP [4], AO* [31], LAO* [21] and LRTDP [9]) improve performance when there is an initial state s_0 by only iterating on reachable, high value states. They also use a heuristic function to initialise V_0 to an estimate of each state's optimal reward.

Multi-Objective MDP. Many decision scenarios cannot be modelled by a single reward function. A *Multi-Objective Markov Decision Process* (MOMDP) generalises MDP rewards into vectors $\boldsymbol{R} : S \times A \times S \rightarrow \mathbb{R}^n$ for a number of objectives $n > 0$. A given policy may be optimal with respect to one objective, but not the others. Thus, we are interested in the set of *undominated* policies: policies where no objective can be improved with no detriment to another objective. For notation, we treat value functions interchangeably with ordered sets: $v_i = V(s_i)$ for $s_i \in S$. Thus, a single-objective function V dominates U if $\forall i \in [1, \ldots, n], v_i < u_i$, written $V \prec U$. A set of vectors $\boldsymbol{V}$ dominates $\boldsymbol{U}$ if $\forall V \in \boldsymbol{V}$ there exists $U \in \boldsymbol{U}$ such that $U \prec V$. Finally, a policy π dominates π' if $\boldsymbol{V}^{\pi'} \prec \boldsymbol{V}^{\pi}$. The set of undominated policies is the maximal set Π^* where $\forall \pi \in \Pi^*, \pi' \notin \Pi^* : \boldsymbol{V}^{\pi} \nprec \boldsymbol{V}^{\pi'}$.

MDP algorithms are adapted for MOMDPs using two distinct approaches [33]. Outer-loop algorithms repeatedly run single-objective solvers on scalarised versions of the problem. Inner-loop approaches generalise the backup operation for multiple objectives e.g., the value function generalises neatly with vector addition $\boldsymbol{V}^{\pi}(s, t) = \sum_{s' \in S} P(s, \pi(s, t), s')[\boldsymbol{R}(s, \pi(s, t), s') \oplus \boldsymbol{V}^{\pi}(s', t + 1)]$. This paper adapts an inner-loop approach for heuristic planning [11].

Hypothetical Retrospection. Hansson's Hypothetical Retrospection procedure systematises human ethical decision-making in scenarios with uncertain outcomes [22]. Hansson suggests we extend our perspective with future perceptions of actions. He argues to view the ethics of decisions from perspectives up to the endpoint of all major foreseeable outcomes. In other words, we imagine the moral reaction to our decision in every eventuality. By hypothetically retrospecting on potential branches of future development, we can build ethical arguments about what to do in the present.

Kolker et al. adapted this procedure into *Machine Ethics Hypothetical Retrospection* (MEHR) [25], described in Figure 1. In summary, an *Ethical Decision Problem* contains a set of states, actions and potential branches of future development. Based on a Value-Based Argumentation Framework [3, 17], arguments in favour of each action are generated from the perspective of their branches of development using an *argument scheme*:

> "From initial state I, it was acceptable to perform action **a**, resulting in consequences **b** with probability $\mathbb{P}$."

Attacks are generated between arguments in support of different actions by two *critical questions*. For every moral theory under consideration, the argument from branch $\mathbf{b_1}$ in support of action $\mathbf{a_1}$ attacks the argument from branch $\mathbf{b_2}$ in support of action $\mathbf{a_2}$ when the following are affirmative:

> **CQ1:** *Did* $\mathbf{b_2}$ *violate a moral theory that* $\mathbf{b_1}$ *did not?*
> **CQ2:** *Did* $\mathbf{a_2}$ *hold a greater probability of violating the moral theory than* $\mathbf{a_1}$?

Here, the moral theory is a free variable, intentionally non-specific to generalise across theories. MEHR has been demonstrated with *Utilitarianism*, assigning numeric values to outcomes and *Deontology*, assigning binary duties to actions. Other theories should be compatible [25]. Argument attacks can be generated from multiple, potentially conflicting moral theories, meaning MEHR extends Hansson's procedure for problems with moral uncertainty.

When an argument is attacked, it represents negative retrospection: another action could have avoided the violation and this was foreseeable. This indicates the decision-making was flawed and so there is negative retrospection. Ideally, an action is selected with no negative retrospection on any of its potential outcomes. If no such action exists, the action with minimal probability of resulting in negative retrospection is selected. This probability is referred to as *non-acceptability*. We note that MEHR simplifies hypothetical retrospection by only considering the perspective of outcome endpoints without intermediate perspectives.

3 RELATED WORK

There are different approaches to *Machine Ethics* planning.

Often, a quantitative approach is used with Act-Utilitarianism [16] and decision theory, i.e., to maximise expected utility. The Utilibot Project is an architecture for autonomous robots in the home [12]. It integrates Bayesian Networks representing uncertain knowledge with MDP actions and utilities, using Policy Iteration to maximise expected utility. Utilitarianism as a sole decision mechanism is problematic. It assumes all outcomes can be appraised in terms of a single number[1] without a clear or agreed methodology. Brundage discusses other critiques of Consequentialism in [10].

Alternatively, there are qualitative approaches. Jedwabny et al. [24] create a linearization from a ranking over ethical rules. For cognitive agents, Dennis et al. derive a substantive order over plans from a ranking of formal ethical principles [14]. Lindner et al. develop techniques to evaluate plans using ethical theories [28, 29]; Dennis et al. apply this to cognitive agents [15], However these approaches assume determinism. Grandi et al. evaluate plans in a Multi Agent System with ranked logical formulae. Uncertainty over other agents' actions is handled optimistically, pessimistically, or to avoid blame [20], all without probability as in our approach.

Rodriguez-Soto et al. use a Multi-Agent MDP with norms and an action praiseworthiness function integrated with existing, non-moral rewards. They adjust penalties and praiseworthiness to ensure agents learn ethically aligned policies in Reinforcement Learning [32]. Outcome uncertainty is handled with utility expectation.

Similarly, the Ethically Compliant Autonomous System (ECAS) framework represents MDPs with a non-moral reward function and moral principles [30, 37]. Agents have conflicting value functions which determine the truth of a moral principle function. The MDP is solved as a linear program, maximising expected reward subject to a single moral principle constraint. ECAS may find there is no moral solution, unlike MEHR which finds the best available policy with minimal non-acceptability.

For problems with moral uncertainty between moral theories, Ecoffet and Lehman apply a voting system in Reinforcement Learning [18]. Moral theories are trained to effectively express their views at each time step using a limited voting budget. Preference between theories is represented by some credence $c \in (0, 1]$ instead of the weak lexicographic order in our paper. Theory preferences are expressed by utilities only and outcome uncertainty by expectation.

Stojanovski et al. define an MDP with positive/negative values that state-action transitions can promote/demote and repair (treating previous actions). They build an ethical reward function that lexicographically avoids harm then, then maximises good [36]. They suggest a hierarchy over MDP values and relative strength in promotion/demotion for future work.

4 MULTI-MORAL MARKOV DECISION PROCESS

We model our planning problem as a *Multi-Moral Markov Decision Process* (MMMDP) based on a MOMDP.

DEFINITION 1 (MULTI-MORAL MARKOV DECISION PROCESS). *A tuple $\langle S, A, P, s_0, H, M, C, L \rangle$ with a finite set of states S, a finite set of actions A and a probabilistic transition function $P : S \times A \times S \rightarrow [0, 1]$. There is an initial state $s_0 \in S$ and $H \in \mathbb{N}$ is the finite horizon.*

M is a finite set of moral theories. C is a finite set of moral considerations and $L : M \rightarrow \mathbb{Q}$ is a weak lexicographic ranking.

States and dynamics are identical to ordinary finite-horizon MDPs (see Section 2). Briefly, moral theories in M are internally consistent moral perspectives, such as Act-Utilitarianism or Deontology [6]. They define an attack relation for MEHR argumentation, detailed in Section 5. Judgements are based on some moral considerations in C, such as utility functions or duties. Preferences between theories are defined with the weak lexicographic order L. In the following two sections we will explain each component in detail with a running example of The Lost Insulin.

States, Actions and Dynamics

We use the ethical dilemma discussed by Coleman [13], adapted to Machine Ethics by Atkinson and Bench-Capon [2]. Hal is a diabetic who, through no fault of his own, has lost his insulin supply and needs some urgently to stay alive. He knows his neighbour Carla has some, but does not have permission to enter her house and take it. Is Hal justified in stealing the insulin to save his life?

We model the problem as a MMMDP visualised on the left of Figure 2. In the initial state s_0 Hal has no insulin. With action $a_1 \in A$ Hal waits and there is a $P(s_0, a_1, s_1) = 0.6$ probability of transitioning to s_1 where Hal dies; there is $P(s_0, a_1, s_0) = 0.4$ probability that Hal lives and faces the choice again. In action $a_2 \in A$, Hal goes to Carla's house and takes the insulin. In state s_2, Carla does not need the insulin and they both live at probability 0.6; in state s_3, the stolen insulin is insufficient and Hal dies with probability 0.15; in state s_4, Carla dies without her insulin at probability 0.15; finally, in state s_5 they both die with probability 0.1. We let the horizon $H = 2$ represent the hours until Hal receives his regular insulin delivery and the dilemma no longer applies. Thus, while there is a $P(s_0, a_0, s_0) = 0.4$ chance Hal survives by waiting through the first hour, the chance Hal lives until his insulin resupply is $0.4^2 = 0.16$. If Hal dies before the horizon, all transitions loop to the same state.

A Moral Consideration

Meaning is extracted from the problem using moral considerations C which track sources of morally relevant information or *moral worth*. We associate states in the MMMDP with worth using a *worth function* $W : S \rightarrow \mathcal{W}$. For notation, we may interpret the worth
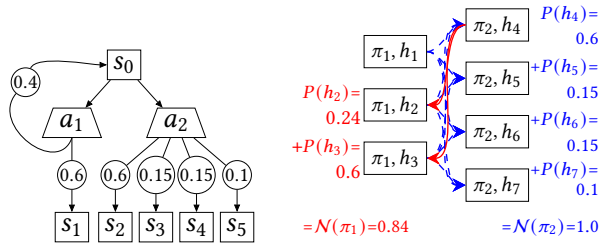
---

[1]An outcome where a person is injured may have -10 utility and an outcome where they lose their life may have -100 utility. Thus, 10 injuries is equal to a loss of life, which does not match intuition.

Figure 2: Left: in s_0 Hal has no insulin; s_1 Hal dies waiting; s_2 both live; s_3 Hal dies; s_4 Carla dies; s_5 both die. Right: MEHR argumentation graph. Boxes are arguments from policy-history pairs, $Arg(\pi, h)$. Directed edges are attacks, meaning $\Psi_m(Arg(\pi', h'), Arg(\pi', h'))$. Red/solid edge for utility attack; blue/dashed line for no-stealing attack. Probabilities on attacked arguments sum to policy's non-acceptability $\mathcal{N}(\pi)$.

function as a sequence $W(s_i) = W_i$ for $s_i \in S$. We define a moral consideration formally below.

DEFINITION 2 (MORAL CONSIDERATION). *A Moral Consideration is a tuple $c = \langle \mathcal{W}, J, Q, \preceq, \approx \rangle$.*

- *$\mathcal{W}$ represents the space of morally information required for decisions. It is referred to as moral worth.*
- *$J : S \times A \times S \to \mathcal{W}$ is a judgement function that extracts the morally relevant information from a state-action transition.*
- *$Q^W : (\mathcal{W} \times [0,1])^n \to \mathcal{W}$ aggregates morally relevant information with probabilities into a single estimate of future moral worth. It uses a baseline of existing morally relevant information from a worth function $W : S \to \mathcal{W}$, subject to Assumption 1 (additivity over moral worth).*
- *$\preceq$ is a transitive preference relation over $\mathcal{W}$.*
- *$\approx: \mathcal{W} \times \mathcal{W} \to \{\top, \bot\}$ maps to $\top$ when moral worth is consistent, $\bot$ otherwise.*

A moral consideration represents a source of morally relevant information in the planning domain. It may be considered a generalisation of an MDP's reward function. This is clearest when formalising the moral consideration of utility associated with Act-Utilitarianism: $\mathcal{W}$ becomes the space of real numbers $\mathbb{R}$ representing utility gained/lost through state transitions $J(s, a, s')$. The aggregation function $Q^W(W', P) = \sum_{i \in 1...|W'|} P_i \cdot (w'_i + w_i)$ calculates expected utility where W' is a set of state utilities, P are associated probabilities, and W are baseline utilities[2]. Relation $\preceq$ is the standard for $\mathbb{R}$. Consistency relation $w \approx w'$ is $\top$ when $|w' - w| < \epsilon$ for some small constant $\epsilon > 0$, $\bot$ otherwise.

For the Lost Insulin, suppose Act-Utilitarianism appraises death with -10 utility, i.e., $J(s_0, a_1, s_0) = 0, J(s_0, a_1, s_1) = -10, J(s_0, a_2, s_5) = -20$. The expected utility of stealing a_2 at s_0 is expressed by aggregation function $Q^W(\{J(s_0, a_2, s') : \forall_{s' \in S}\}, \{P(s_0, a_2, s') : \forall_{s' \in S}\})$. This is $0.6(0 + W(s_2)) + 0.15(-10 + W(s_3)) + 0.15(-10 + W(s_4)) + 0.1(-20 + W(s_5))$ and with baseline W initialised to 0, this is $0.6(0 + 0) + 0.15(-10 + 0) + 0.15(-10 + 0) + 0.1(-20 + 0) = -5$ utility[3].

---

[2]We assume for simplicity that W, P and W' have equal magnitude.
[3]Successor states from a_2 with 0 probability are omitted from the equations for space.

For syntactic sugar, we abstract state-action aggregation as follows: $Q^W(s, a) = Q^W(\{J(s, a, s') : \forall_{s' \in S}\}, \{P(s, a, s') : \forall_{s' \in S}\})$. And so, $Q^W(s_0, a_2) = -5$. For reasons in Section 5, alternate sets of states and probabilities may be passed as follows

$$Q^W(s, a, S', P') = Q^{W'}(\{J(s, a, s') : \forall_{s' \in S'}\}, P')$$

where W' contains worth for states in S': $W' = \{W(s) : s \in S'\}$.

We are interested in non-stationary, deterministic policies $\pi : S \times \{0, \ldots, H-1\} \to A$. We exclude stochastic policies for simplicity and because, arguably, randomness is not acceptable in ethical decision making. A moral consideration evaluates a policy with respect to every state iteratively using a *non-stationary worth function*. We interpret this as an $H + 1 \times |S|$ matrix where $W[t]$ is a worth function for all states at time step $t \in \{0, ..., H\}$; $W[t](s)$ is the expected worth at time t, state s. The worth function is defined as

$$W^\pi[t](s) = Q^{W^\pi[t+1]}(s, \pi(s, t)), \forall s \in S, t \in \{0, \ldots, H-1\}$$

The vector $W[H]$ contains null values in $\mathcal{W}$ so as to not affect aggregation in Q. One worth function is preferred to another if it is preferable at every state and time. A *consideration-optimal worth function* is the most preferable worth achievable at each state and time. This is calculated like the Bellman Optimality Equation [5]:

$$W^*[t](s) = \max_{a \in A}^{\succeq} Q^{W^*[t+1]}(s, a), \forall s \in S, t \in \{0, \ldots, H-1\}$$

This assumes moral worth is additive through Q.

ASSUMPTION 1 (ADDITIVE MORAL WORTH). *Let $W'^+, W'^- \in \mathcal{W}^n$ be worth functions, for any n, and let P'^+, P'^- be sets of n probabilities such that (W'^+, P'^+) is preferred to (W'^-, P'^-), i.e., $Q^W(W'^+, P'^+) \prec Q^W(W'^-, P'^-)$ for any baseline $W \in \mathcal{W}^n$.*

Furthermore, let $W^+, W^- \in \mathcal{W}^n$ be worth functions such that W^+ is preferred to W^-, i.e., $Q^{W^+}(W', P') \prec Q^{W^-}(W', P')$ for any worth function $W' \in \mathcal{W}^n$ and any probabilities $P' \in [0,1]^n$.

It must be the case that $Q^{W^-}(W'^-, P'^-) \prec Q^{W^+}(W'^+, P'^+)$.

Many Moral Considerations

MMMDPs may have plenty of moral considerations. Suppose an absolute moral theory where any probability of breaking the law is wrong. A moral consideration's worth could be represented by $\mathcal{W} = \{\top, \bot\}$ where $\top$ violates the law and $\top \prec \bot$. In the example, $J(s_0, a_2, s_2) = \top$ and $J(s_0, a_1, s_1) = \bot$. Worth $\mathcal{W}$ aggregates with $Q^W(W', P) = \bigvee_{i \in |W'|}(P_i > 0 \wedge w'_i) \vee w_i$. Worth is consistent only if it is equal, $w \approx w' = w == w'$. Thus, if $W[t](s) = \bot$ for all $t \in \{0, \ldots, H\}$ and $s \in S$, then $Q^W(s_0, a_1) = Q^W((\bot \vee \bot) \vee \ldots \vee (\bot \vee \bot)) = \bot$ (no law violated) and $Q^W(s_0, a_2) = \top$ (law violated).

For notation, we distinguish moral considerations with subscripts: consideration $c_i \in C$ is the tuple $\langle \mathcal{W}_i, J_i, Q_i^W, \preceq_i, \approx_i \rangle$. To hold the worth for all considerations in C, we use a *multi-worth function* $\mathbf{W} = \{W_1, \ldots, W_{|C|}\}$ where $W_i \in \mathbf{W}$ is a worth function for consideration $c_i \in C$. We hold a single element of worth from all moral considerations using a *worth-vector* $\vec{w} \in 2^{\{\mathcal{W}_1, \ldots, \mathcal{W}_{|C|}\}}$. We generalise Q for multiple considerations as $\mathbf{Q}$, which returns expectations as a worth-vector. We bend notation to apply the worth-vector across a multi-worth function at state s and time t: $\mathbf{W}^\pi[t](s) = \mathbf{Q}^\pi(s, \pi(s, t)) = \{Q_i^\pi(s, \pi(s, t)) : \forall i \in 1, ..., |C|\}$. We demonstrate in Figure 3.

$$W_{abs} = \begin{bmatrix} \top & \bot & \bot \\ \top & \top & \bot \\ \top & \bot & \bot \end{bmatrix} \quad W_{util} = \begin{bmatrix} 1 & 2 & 0 \\ 3 & 4 & 0 \\ 5 & 6 & 0 \end{bmatrix} \begin{matrix} s_0 \\ s_1 \\ s_2 \end{matrix}$$

$$t = 0 \ 1 \ 2 \qquad W = \{W_{abs}, W_{util}\}$$

$$W[1](s_1) = \vec{w} = \{\top, 4\}$$

$$\vec{W} = \begin{bmatrix} \{\{\top, 1\}\} & \dots & \{\{\bot, 0\}\} \\ \{\{\top, 3\}\} & \dots & \{\{\bot, 0\}\} \\ \{\{\bot, 5\}\} & \dots & \{\{\bot, 0\}\} \end{bmatrix}$$

$$W_{abs}[1] = (\bot, \top, \bot) \quad W_{util}[1] = (2, 4, 6)$$

$$W_{abs}[1](s_1) = \top \quad W_{util}[1](s_1) = 4$$

$$\vec{W}[1](s_1) = \{\{\top, 4\}\}$$

Figure 3: Shows notation. Worth function W maps state-time to worth. Multi-worth function W holds many worth functions. Worth-vectors $\vec{w}$ store worth across considerations. From Section 7, $\vec{W}$ maps state-time to a set of worth-vectors.

Moral Theories

A moral theory is a domain-dependent representation of an ethical perspective. It defines the nature of hypothetical retrospection based on worth from moral considerations; a moral theory decides how policies and branches of development are compared.

DEFINITION 3 (MORAL THEORY). *A moral theory $m \in M$ is a tuple $m = \langle C^m, \Psi_m \rangle$ where $C^m \subseteq C$ is a subset of moral considerations from a MMMDP. Function $\Psi_m : 2^{Arg} \to \{\div, \circ\}$ determines if arguments should attack each other according to this theory (see Section 5).*

In this paper, we only consider moral theories with a single moral consideration, though the formalism supports many. In that case, the attack relation would define how considerations are balanced.

Moral theories may conflict; stakeholders may believe some theories more than others. This is expressed with a weak lexicographic order. For theories $m, m' \in M$, if $L(m) < L(m')$, any worth from considerations C^m is prioritised over $C^{m'}$. In the running example, a Utilitarian-optimal worth function has $W^*_{util}[0](s_0) = -5$ by stealing insulin. However, the absolutist optimal worth is $W^*_{abs}[0](s_0) = \bot$ by waiting. The same policy cannot optimise both. If $L(m_{util}) < L(m_{abs})$, the Utilitarian's optimal policy would be preferable; on the other hand if $L(m_{util}) > L(m_{abs})$, then the absolute no-stealing policy is preferred. If $L(m_{util}) = L(m_{abs})$, the stakeholder has no preference or moral uncertainty. Another mechanism is required to differentiate the policies such as MEHR.

5 MACHINE ETHICS HYPOTHETICAL RETROSPECTION OVER POLICIES

Machine Ethics Hypothetical Retrospection (MEHR) [25] is an argumentation procedure [17] for Machine Ethics decision making with outcome and moral uncertainty. We contribute a novel formalism for probabilistic planning. It creates an argument in favour of each policy from the perspective their potential branches of future development and models negative retrospection as argument attacks. We model a branch of development as a history/trajectory in the MMMDP $h = \langle S^h, A^h \rangle$ where S^h is a sequence of states and A^h is a sequence of actions. S_1^h is the start state, action A_1^h transitions to state S_2^h etc., i.e., $h = \{S_1^h, A_1^h, S_2^h, A_2^h, \dots, S_{H+1}^h\}$. The probability of a history is $P(h) = \prod_{i=1}^{|A^h|} P(S_i^h, A_i^h, S_{i+1}^h)$. A history's worth is

---

Algorithm 1 Extract history worth and probabilities from policy

Input: Policy π, multi-worth function W
Output: Mapping $\phi : \{W_c : \forall c \in C\} \to [0, 1]$

1: **procedure** EXTRACTHISTORIES($s, t, \vec{w}, pr, \phi$)
2:     **if** $t \geq H$ **or** $\{s' \in S : P(s, a, s') > 0\}$ **is** $\emptyset$ **then**
3:         $\phi(\vec{w}) \leftarrow \phi(\vec{w}) + pr$
4:         **return**
5:     **for** $\{s' \in S, a \in A : P(s, a, s') > 0\}$ **do**
6:         $W' \leftarrow \{\vec{w}_c : c \in C\}$ // Set baseline to prev. worth
7:         $\vec{w}' \leftarrow Q^{W'}(s, a, \{s'\}, \{1.0\})$
8:         $ExtractHistories(s', t+1, \vec{w}', pr \cdot P(s, a, s'), \phi)$
9: $\phi(\vec{w}) \leftarrow 0 : \forall \vec{w} \in 2^{\{W_c : \forall c \in C\}}$ // Set default to map 0.
10: $ExtractHistories(s_0, 0, \{null \ W_c : \forall_{c \in C}\}, 1.0, \phi)$
11: **return** ϕ

---

found using the same aggregation function as with policies:

$$W^h[t](s_i) = Q^{W^h[t+1]}(s_i, a_i, \{s_{i+1}\}, \{1.0\})$$

A single successor state is specified and 1.0 probability is passed to find a history's worth. For a policy π, the worth and probability of all its histories can be found using Depth-First-Search. The algorithm exploits Assumption 1 by aggregating worth while traversing from s_0 to the horizon. For transparency, there is pseudocode in Algorithm 1; we do not examine it since we do not consider it a notable contribution. For the Lost Insulin running example, the 'waiting policy' always selects a_1 and has three potential histories: $h_1 \equiv \{s_0, a_0, s_0, a_0, s_0\}$ (waiting two hours), $h_2 \equiv \{s_0, a_0, s_0, a_0, s_1\}$ (waiting two hours, then collapse), $h_3 \equiv \{s_0, a_0, s_1, a_0, s_1\}$ (waiting one hour, then collapse).

We take the expected worth of a policy as its worth at the initial state s_0 and time $t = 0$, $Q^\pi(s_0, \pi(s_0, 0))$. Thus, from a selection of alternate policies, we use MEHR argumentation to select an ethical policy [25]. The procedure translates naturally. An argument Arg is generated from the perspective of each history endpoint in support of the preceding policy. Arguments are generated by an *Argument Scheme*. We define $Arg(\pi, h)$ as

> "From the initial state s_0, it was right to use policy π, resulting in the history $h = \{S_1^h, A_1^h...A_{H-1}^h S_H^h\}$ with probability $P(h)$."

This may be interpreted as the default hypothetical retrospective argument, though it may be invalidated by other retrospections. There may be negative retrospection for a policy after an outcome if an alternate policy is preferred. We define this as an attack relation C_m for each moral theory $m \in M$. An argument $Arg(\pi, h)$, supporting policy π from history h, attacks opposing argument $Arg(\pi', h')$ if it causes negative retrospection: reveals a decision making flaw in choosing π'. For each moral theory, we define the attack relation in terms of Critical Questions [2, 25]. There is an attack when $\Psi(Arg(\pi, h), Arg(\pi', h')) = \div$ and the following are affirmative:

> **CQ1:** Did h' violate the moral theory and h did not?
> **CQ2:** Was there greater expectation that π' would violate the moral theory than π?

This represents that an alternative history was preferable and expected to be preferable, resulting in negative retrospection and

undermining the target argument. Each moral theory determines the meaning of a violation and violation expectation. For Act-Utilitarianism, the questions are **CQ1:** Did h have greater utility than h'? **CQ2:** Did π have greater expected utility than π'? For each theory the attack is defined as $\Psi(Arg(\pi, h), Arg(\pi', h')) = $ ⊢ **if** $CQ1 \wedge CQ2$ **otherwise** ○ when

$$CQ1 = W^h[0](s_0) \succ W^{h'}[0](s_0)$$

$$CQ2 = Q^\pi(s_0, \pi(s_0, 0)) \succ Q^{\pi'}(s_0, \pi'(s_0, 0))$$

Thus, there is a supportive argument generated from the outcome of every policy; moral theories suggest attacks between arguments that support different policies. For each argument, we define its attackers as a set of argument and moral theory pairs:

$$\mathcal{X}(Arg(\pi, h)) =$$
$$\{(m, Arg(\pi', h')) : \forall m \in M,$$
$$\Psi_m(Arg(\pi', h'), Arg(\pi, h)) = \vdash \wedge \nexists m' \in M, L(m') < L(m) \wedge$$
$$Q^{\pi'}_{m'}(s_0, \pi'(s_0, 0)) \prec Q^\pi_{m'}(s_0, \pi(s_0, 0))\}$$

In words, there is a potential attack on an argument for every other policy π', each of their histories h', and every moral theory m. MEHR attacks are based the moral theories' attack relation Ψ_m, but an attack is blocked if there is a lexicographic preferred theory such that π is preferred. Note, there can be multiple attacks from one argument to another, unlike traditional argumentation [17].

The non-acceptability of a policy is its probability of reaching an outcome with negative retrospection for each moral theory. Equivalently, it is the number of attacks on each supporting argument multiplied by the argument history's probability:

$$\mathcal{N}(\pi) = \sum_{h \text{ from } \pi} P(h) \cdot |\mathcal{X}(Arg(\pi, h))|$$

A policy with 0 non-acceptability should be selected if one exists. Otherwise, there is a 'generalised case of a moral dilemma' [22] and we choose the policy with minimal non-acceptability instead. We define the *morally acceptable policies* as those minimising non-acceptability $\Pi^* = \{\pi : \mathcal{N}(\pi) = \min_{\pi'} \mathcal{N}(\pi')\}$. We visualise MEHR with The Lost Insulin example on the right of Figure 2.

We note this formalism's philosophical assumptions. First, that an action's moral worth can be aggregated and combined with probabilities–that an aggregation function Q can be expressed for moral theories. As in MEHR, we assume the perspective of the endpoint of action outcomes is sufficient for moral decision-making. Hansson's original Hypothetical Retrospection argues relevant information can be described in terms of consequences, though Hansson considers all perspectives up to an action's endpoint, rather than the endpoint perspective alone. MEHR also differs by allowing conflicting moral theories under moral uncertainty.

6 A NON-MORAL EXTENSION

We make our formalism compatible with non-moral costs and goals using ideas from the Stochastic Shortest Path (SSP) Problem [5].

DEFINITION 4 (MULTI-MORAL STOCHASTIC SHORTEST PATH PROBLEM). *A Multi-Moral Stochastic Shortest Path Problem (MMSSP) is a tuple $\langle S, A, P, s_0, H, M, C, L, R, b, G \rangle$ with components identical to a MMMDP. Additionally: a non-moral cost consideration $R \in C$,*

defined as a special type of moral consideration; a positive budget $b \in \mathbb{R}^+$; *and a set of goal states $G \subseteq S$ that are absorbing:* $P(s, a, s') = 0, \forall s \in G, a \in A, s' \in S\backslash G$ *and* $P(s, a, s) = 1, \forall s \in G, a \in A$.

A MMSSP has competing moral theories and quantified costs over state transitions. For simplicity, costs are a moral consideration, almost identical to the utility consideration discussed in Section 4. This consideration is not associated with any moral theory. Also, state transitions must cost more than 0, i.e., $J_R(s, a, s') > 0, \forall s \in S\backslash G, s' \in S, a \in A$. This is a standard assumption for SSPs [8]. There is a budget $b \in \mathbb{R}^+$ which we interpret as the cost an agent is willing to spend on morality before reaching a goal state $s \in G$.

One potential issue when integrating non-moral objectives and moral considerations is the presence of *positive moral loops*. Take the example of an autonomous system managing investments for a private business. A policy ought to invest funds ethically while achieving goals, otherwise stakeholders will not adopt it. A truly ethical policy may get caught in a positive moral loop and never achieve its goals (donating to charity until bankruptcy). Thus, we prune policies over budget b. Stationary policies $\pi : S \to A$, typical for SSPs, would prune moral loops entirely once they exceed the budget (no money for charity), which is why we chose non-stationary solutions $\pi : S \times \{0, \ldots, H-1\} \to A$. Our solutions follow the moral loop for some number of cycles while leaving budget to pursue the goal afterwards. The budget is similar to the bound vector used for Multi-Objective SSPs [11], though we bound solutions with respect to just one consideration instead of all objectives.

Solutions to infinite horizon SSPs are assumed to be proper, meaning the probability of reaching a goal from s_0 is 1.0. In finite horizon problems, stochasticity may prevent the policy from reaching a goal within the horizon. Thus, we say proper policies reach a goal from s_0 under the budget b with non-zero probability.

ASSUMPTION 2 (WEAK BUDGETED IMPROPER POLICIES). *For a MMSSP $\langle S, A, P, s_0, H, M, C, L, R, b, G \rangle$ all improper policies π have an expected cost that exceeds the budget:* $Q^\pi_R(s_0, \pi(s_0, 0)) > b$. *Also, there exists a proper policy π with an expected cost under the budget:* $Q^\pi_R(s_0, \pi(s_0, 0)) \leq b$.

We define *optimal moral non-stationary policies* for a MMSSP:

$$\Pi^b = \{\pi \in \Pi : Q^\pi_R(s_0, \pi(s_0, 0)) \leq b\},$$

$$\Pi^{b*} = \{\pi : \mathcal{N}(\pi) = \min_{\pi' \in \Pi^b} \mathcal{N}(\pi')\}, \pi^* = \arg\min_{\pi \in \Pi^{b*}} Q^\pi_R(s_0, \pi(s_0, 0))$$

The set Π^b are all the *proper policies* whose expected cost is under the budget b. The set of *moral proper policies* Π^{b*} are those proper policies with minimal non-acceptability. The *optimal moral policy* is the moral proper policy with minimal expected cost.

7 MORAL PLANNER

In the previous sections we construct a Multi-Moral MDP/SSP like a Multi-Objective MDP, with added lexicographic preferences and greater freedom over types. We presented a mechanism to evaluate solution policies ethically using MEHR. The following presents a heuristic moral planning algorithm with two stages. First, an AO* Multi-Objective heuristic planning algorithm is adapted from [11] to find the set of proper *Pareto undominated* policies. Second, the undominated policies are evaluated using MEHR argumentation. A

Algorithm 2 MPlan Algorithm

Input: Multi-Moral SSP $\langle S, A, P, s_0, H, M, C, L, R, b, G \rangle$, heuristic η
Output: Optimal moral policy, π

1:   $\vec{W} \leftarrow \{\{\eta_c(s,t) : \forall c \in C\}\}; F \leftarrow \{(s_0, 0)\}; I \leftarrow \emptyset;$
2:   $\alpha(s,t) \leftarrow \emptyset : \forall_{s \in S, t \in \{0,...,H-1\}}; \alpha(s_0, t) \leftarrow A$
3:   **do**
4:      $\vec{W}' \leftarrow \vec{W}; Z \leftarrow \text{postOrderDFS}(I \cup F, \alpha)$
5:      **for** $s, t \in Z$ **do**
6:         $\vec{W}[t](s) \leftarrow \{\boldsymbol{Q}^{W[t+1]}(s,a) : \forall a \in A, \boldsymbol{W} \text{ from } \vec{W}\}$
7:         $\vec{W}[t](s) \leftarrow \{\vec{w} \in \vec{W}[t](s) : \vec{w}_R[t](s) \leq b\}$
8:         $\vec{W}[t](s) \leftarrow \text{pprune}(\vec{W}[t](s))$
9:         $\alpha[t](s) \leftarrow \{a \in A : \boldsymbol{Q}^{W[t+1]}(s,a) \in \vec{W}[t](s)\}$
10:       $F \leftarrow (F \backslash \{(s,t)\}); I \leftarrow I \cup \{(s,t)\}$
11:       $F \leftarrow F \cup \{(s', t+1) : t+1 < H \wedge \exists_{a \in A} P(s, a, s') > 0\} \backslash I$
12:   **while** $((F \cap Z) \backslash G = \emptyset) \wedge (\vec{W} \approx \vec{W}')$
13:   $\Pi \leftarrow$ Extract policies π from $\vec{W}$
14:   $\phi^\pi \leftarrow$ Extract Histories from $\pi, \forall \pi \in \Pi$
15:   **return** $\pi \leftarrow \text{MEHR}(\Pi, \phi)$

---

policy with minimal non-acceptability (and then minimal expected cost) is output. We give pseudocode in Algorithm 2. The algorithm explores the MMMDP as a graph whose nodes are state-time pairs. It maintains an explicit sub-graph, represented by interior I, with successors inside the sub-graph, and a fringe F with successors outside. The explicit sub-graph grows as the algorithm explores the MMMDP. A node's worth across moral considerations is grouped by a *worth vector*, $\vec{w} \in 2^{\mathcal{W}_1 \times ... \times \mathcal{W}_{|C|}}$ where $\vec{w}_i$ is the expected worth for moral consideration C_i (see Figure 3). A worth vector *Pareto dominates* another if it is preferable by one moral consideration (including the cost function) and at least as preferable by all others. This ignores the weak lexicographic order.

DEFINITION 5 (WORTH VECTOR PARETO DOMINANCE). *For two worth vectors* $\vec{w}$ *and* $\vec{w}'$, *then* $\vec{w}$ *Pareto dominates* $\vec{w}'$ *if and only if* $\exists i \in \{1, \ldots, |C|\}, \vec{w}_i \succ_i \vec{w}'_i$ *and* $\forall j \in \{1, ..., |C|\}, \vec{w}_j \succeq_j \vec{w}'_j$.

Each state-time maps to a set of worth vectors, $\vec{W} : S \times \{0, ..., H-1\} \rightarrow 2^{\{\mathcal{W}_1 \times ... \times \mathcal{W}_{|C|}\}}$ (see Figure 3). Each worth vector $\vec{w} \in \vec{W}[t](s)$ represents the expectations of a potential undominated solution at time t, state s. On line 1, $\vec{W}$ gets one worth vector for each node from a heuristic estimate of the consideration optimal worth function $W_c^*[t](s)$ (Section 4), $\eta_c : S \times [0, ..., H-1] \rightarrow \mathcal{W}_c$. Initially, frontier F has initial state s_0 at time 0 and interior I is empty (line 1). The sub-graph will grow to contain state-time nodes reachable in undominated solutions. Specifically, $\alpha : S \times \{0, \ldots, H-1\} \rightarrow 2^A$ maps a state-time to their mapped actions from all potential undominated solutions, Initially, any solution could be undominated, so α maps to all actions for state s_0, time 0.

The algorithm performs iterative *back-ups* on state-time nodes in the sub-graph. This informs the action mappings of potential solutions α. The sub-graph is backed up via post-order DFS or descending time order (line 4-5) to maximise each back-up's effect.

A back-up on a state-time collects the multi-worth function $\boldsymbol{W}$ (Section 4) from each permutation of expected worth in $\vec{W}$, expressed in pseudocode as '$\forall \boldsymbol{W}$ from $\vec{W}$' (line 6). In an informal example, say $\vec{W}[t](s)$ contains two worth vectors, estimating the

worth of two potential undominated solutions for state s, time t, $\vec{W}[t](s) = \{\vec{w}_1, \vec{w}_2\}$. Given one successor with $\vec{W}[t+1](s') = \{\vec{w}'_1, \vec{w}'_2\}$, 4 multi-worth functions are constructed ($\vec{w}_1 + \vec{w}'_1, \vec{w}_1 + \vec{w}'_2, \vec{w}_2 + \vec{w}'_1, \vec{w}_2 + \vec{w}'_2$), placing estimate permutations for state s, time t and state s', time $t+1$ into $\boldsymbol{W}$. In a back-up, these multi-worth functions are passed to the worth aggregation function for every action. This finds the worth vector expectation of applying each action to a potential undominated solution. For MMSSPs, worth vectors over the budget are pruned away (line 7). The *pprune* algorithm [34] prunes all Pareto dominated worth vectors (line 8).

The actions that generated undominated worth vectors are added to $\alpha[t](s)$ (line 9). After a back-up, fringe nodes are expanded, moved from the fringe to the interior and replaced by successors (lines 10-11). The algorithm reiterates over the expanded sub-graph (line 3), backing-up state-time nodes reachable through actions in α, thus, nodes that are in-budget and promising by some moral consideration. Other nodes are pruned. This focused search avoids full sweeps across the state space, as required by planning algorithms like (MO) Value Iteration [5, 40].

Iteration continues until there are no unexpanded state-time nodes that could be in an undominated solution (line 12). Additionally, the expectations of solutions must be consistent between iterations. To find this, we generalise the $\approx$ operator from the moral considerations to $\vec{W}$ and a copy before each iteration, $\vec{W}'$. We say $\vec{W} \approx \vec{W}'$ if and only if $\forall s \in S, t \in \{0, \ldots, H\}, \exists \vec{w} \in \vec{W}$ and $\vec{w}' \in \vec{W}'$ such that $\forall c \in C, \vec{w}_c \approx \vec{w}'_c$. In other words, each worth vector from $\vec{W}$ must correspond to an equivalent in $\vec{W}'$.

After a set of Pareto undominated worth vectors are found, all policies are extracted on line 13 (algorithm omitted for space). The worth and probability of policy histories is found on line 16 using Algorithm 1. These values are passed to the MEHR argumentation procedure to find the most preferable policies. This is implemented with the algorithm described in [25], with actions as policies, potential branches of future development as histories, and moral theories as described in Section 5.

8 EXPERIMENT

We evaluate our algorithm with experiments based on The Lost Insulin example. Implementation, MDPs and results are available on https://github.com/sameysimon/MoralPlanner. All experiments ran on an 2020, M1 MacBook Air with 8GB of memory.

We expand the scenario for the experiments. Hal urgently needs insulin. There is a scheduled resupply in 200 minutes. Each time step represents a 10 minute increment. At each step, there is a 0.6 probability Hal dies without insulin. Hal can wait or go to Carla's house. There is a 0.2 chance he is caught, arrested and forced to wait. If Hal gets to Carla's, he will begin searching, though he can go home and wait. While searching, he may compensate Carla: he can leave a lot of money with a 0.7 chance of covering the insulin cost; he can leave a small amount, with a 0.1 chance of covering the cost. Once finding the insulin, he can steal it or leave and go home. Without insulin, Carla has 0.1 chance of dying at each time step. We assume Hal replaces Carla's insulin after the delivery.

We discern a few moral theories. As before, a collective Act-Utilitarianism measures Hal and Carla's well-being. This can divide

Table 2: Table showing number of state-time expansions (with % of 286 state-time space), backups and iterations performed by the Multi-Moral AO* algorithm.

Config.	Expanded	Backups	Iterations
Hal & Carla	243 (84%)	2472	23
Hal & Carla & Steal	243 (84%)	2472	23
Hal & Carla & Steal_Comp	243 (84%)	2472	23
Non-Moral Hal & Carla	271 (95%)	2822	21

Table 1: Moral Theory for Hal's Ethical Egoism denoted by H; Carla's Altruism by C; absolute no stealing by S; stealing allowed with compensation by SC; non-moral cost function by R. The superscript denotes lexicographic rank. Shows config non-acceptability and execution time to 3 s.f.

Config.	Moral Planner					
	$Q^{W[1]}(s_0, \pi(s_0, 0))$			Non-	No. of	Time
	Hal	Carla	Steal	acc.	solns.	(ms)
C^0, H^0	-8.9312	-1.1007	N/A	0.109	8	44.5
C^0, H^1	-10.211	0	N/A	0	8	44.7
C^1, H^0	-8.9312	-1.1007	N/A	0	8	44.0
C^0, H^0, S^0	-8.9312	-1.1007	$\top$	0.237	8	59.2
C^1, H^0, SC^0	-8.9312	-1.1007	$\top$	0.147	8	57.7
C^0, R	-18.387	-0.7705	N/A	0	4	43.6
C^0, S^0, R	-18.387	-0.7705	$\top$	0	4	58.0

into *Ethical Egoism*, measuring only Hal's utility [35], and *Altruism* [27], measuring only Carla's well-being. There is an absolute theory against stealing and another that allows stealing if there is compensation. For MMSSPs, states where Hal has insulin are goals and all transitions cost 1. We set the budget b to 18.5.

We use domain dependent heuristics for $\eta(s, t)$. We estimate Hal's utility with -0.4 when he waits and 0 otherwise; Carla's utility to -0.1 when she loses her insulin and 0 otherwise. For absolute theories, we estimate $\bot$ since Hal can always avoid stealing.

The resulting MDP has 286 state-time pairs. We ran configurations with different moral theories and lexicographic ranks 5 times each and averaged the total run time. Results are in Table 1. When Hal and Carla's theories have equal rank, a policy for stealing the insulin is selected, matching intuition since Hal's life is at greater risk. The non-acceptability is greater than 0 meaning there is a moral dilemma. This is likely because it is not guaranteed that Hal's life will be saved. From the perspective of the outcomes where Hal steals the insulin but dies anyway, there would be negative retrospection. High and low compensation policies yield the same non-acceptability as these theories are ambivalent. When Hal's life is lexicographic preferred, any chance of saving Hal is worth risking Carla. When Carla's life is lexicographic preferred, Hal must wait at home since any risk to Carla is not permissible.

Introducing the law against stealing, the policy where Hal steals insulin is still preferable though non-acceptability has risen. This

is because more outcomes are indefensible. With the compensation theory, Hal leaves the maximum amount and non-acceptability somewhat recovers. Considering the MMSSP examples, in both cases, a middle ground that is best for Carla while remaining in budget is selected. We see a trend across the results that a greater number of moral considerations leads to a greater running time.

In Table 2, we show total back-ups, iterations and state expansions. Absolute moral theories back propagate easily due the simplicity of the moral consideration. The non-moral cost function causes more backups, but reduces the number of iterations. This may be because search is directed towards policies that go to Carla's house, which contains more states, but makes convergence faster.

9 DISCUSSION

We have formalised moral theories for ethical planning with uncertainty over decision outcomes and moral theories. While the theories in this paper are simple (expected utility maximisation and probabilistic absolutism), we hope our formalism is expressive. One theory could back-propagate all state-probability information in its worth type $\mathcal{W}$, instead of reducing to a 'single expected' worth. This represents a great deal at the cost of exponential blow-up. In future, we will explore the balance of tractable expression.

Our approach is compatible with non-moral goals and costs with MMSSPs. We showed solutions must be non-stationary to account for positive moral loops. Future work may explore eliminating these loops (like traps [26]) to alleviate this. Our algorithm finds a proper policy if one exists, unlike the work by Svegliato et al. which modifies ethical polices towards non-moral goals. If this can not be done under a threshold, no policy is returned [37]. In future, we could try to apply this approach to our algorithm.

In some MMMDPs, there is a large number of undominated policies. Some multi-objective algorithms find the Pareto Front or the Convex Coverage Set (CCS) which is generally smaller [33]. Unfortunately, these may not include the MEHR optimal policy. Additionally, policies can have an exponential number of histories and since MEHR compares each policy's arguments, we expect time+space complexity to be exponential. Developing an n-step lookahead algorithm would help and we may try Monte Carlo history sampling to limit the number of arguments. We plan to search for optimisations and recommend our current algorithm for small scale problems which require more ethical scrutiny.

We hope our novel formalism helps the bridge between probabilistic planning and the expressive realm of Ethics. We integrated ethical heuristics to guide our search, which we believe is novel for Machine Ethics. The approach can express moral uncertainty, shown by experiments sensitive to configurations of moral theories. Our approach has potential for explainability with MEHR decisions based on arguments that are reproducible following an inquiry.

ACKNOWLEDGMENTS

REFERENCES

[1] Colin Allen, Wendell Wallach, and Iva Smit. 2006. Why Machine Ethics? *IEEE Intelligent Systems* 21, 4 (2006), 12–17.

[2] Katie Atkinson and Trevor Bench-Capon. 2008. Addressing moral problems through practical reasoning. *Journal of Applied Logic* 6, 2 (2008), 135–151.

[3] Katie Atkinson and Trevor Bench-Capon. 2016. Value based reasoning and the actions of others. In *ECAI 2016*. IOS Press, 680–688.

[4] Andrew G. Barto, Steven J. Bradtke, and Satinder P. Singh. 1995. Learning to Act Using Real-Time Dynamic Programming. *Artificial Intelligence* 72, 1-2 (1995), 81–138.

[5] Richard Bellman. 1957. *Dynamic Programming.* Dover Publications.

[6] Trevor JM Bench-Capon. 2020. Ethical approaches and autonomous systems. *Artificial Intelligence* 281 (2020), 103239.

[7] Dimitri Bertsekas. 1995. *Dynamic Programming and Optimal Control.* Athena Scientific, Belmont, MA.

[8] Dimitri P Bertsekas and John N Tsitsiklis. 1991. An analysis of stochastic shortest path problems. *Mathematics of Operations Research* 16, 3 (1991), 580–595.

[9] Blai Bonet and Hector Geffner. 2003. Labeled RTDP: Improving the Convergence of Real-Time Dynamic Programming. In *ICAPS*.

[10] Miles Brundage. 2014. Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence* 26, 3 (2014), 355–372.

[11] Dillon Z Chen, Felipe Trevizan, and Sylvie Thiébaux. 2023. Heuristic search for multi-objective probabilistic planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 11945–11954.

[12] Christopher Cloos. 2005. The Utilibot Project: An Autonomous Mobile Robot Based on Utilitarianism. *AAAI Fall Symposium - Technical Report* (01 2005).

[13] Jules L Coleman. 1992. *Risks and wrongs.* CUP Archive.

[14] Louise Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. 2014. Ethical choice in unforeseen circumstances. In *Towards Autonomous Robotic Systems (TAROS): Annual Conference.*

[15] Louise A. Dennis, Martin Mose Bentzen, Felix Lindner, and Michael Fisher. 2021. Verifiable Machine Ethics in Changing Contexts. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 13 (May 2021), 11470–11478. https://ojs.aaai.org/index.php/AAAI/article/view/17366

[16] Julia Driver. 2022. The History of Utilitarianism. In *The Stanford Encyclopedia of Philosophy* (Winter 2022 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

[17] Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence* 77, 2 (1995), 321–357.

[18] Adrien Ecoffet and Joel Lehman. 2021. Reinforcement learning under moral uncertainty. In *International conference on machine learning*. PMLR, 2926–2936.

[19] Hector Geffner and Blai Bonet. 2013. *A Concise Introduction to Models and Methods for Automated Planning.* Morgan & Claypool Publishers.

[20] Umberto Grandi, Emiliano Lorini, and Timothy Parker. 2023. Moral planning agents with ltl values. In *International Joint Conference on Artificial Intelligence (IJCAI).*

[21] Eric A. Hansen and Shlomo Zilberstein. 2001. LAO* A heuristic search algorithm that finds solutions with loops. *Artificial Intelligence* 129, 1-2 (2001), 35–62.

[22] S Hansson. 2013. *The ethics of risk: Ethical analysis in an uncertain world.* Springer.

[23] Ronald A. Howard. 1960. *Dynamic Programming and Markov Processes.* MIT Press.

[24] Martin Jedwabny, Pierre Bisquert, and Madalina Croitoru. 2021. Generating preferred plans with ethical features. In *Florida Artificial Intelligence Research Society*, Vol. 34.

[25] Simon Kolker, Louise Dennis, Ramon Fraga Pereira, and Mengwei Xu. 2023. Uncertain Machine Ethical Decisions Using Hypothetical Retrospection. *arXiv preprint arXiv:2305.01424* (2023).

[26] Andrey Kolobov, Mausam Mausam, Daniel Weld, and Hector Geffner. 2011. Heuristic search for generalized stochastic shortest path MDPs. In *Proceedings of the International Conference on Automated Planning and Scheduling*, Vol. 21. 130–137.

[27] Richard Kraut. 2020. Altruism. In *The Stanford Encyclopedia of Philosophy* (Fall 2020 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[28] Felix Lindner, Martin Mose Bentzen, and Bernhard Nebel. 2017. The HERA approach to morally competent robots. In *International Conference on Intelligent Robots and Systems (IROS).*

[29] Felix Lindner, Robert Mattmüller, and Bernhard Nebel. 2020. Evaluation of the moral permissibility of action plans. *Artificial Intelligence* 287 (2020), 103350.

[30] Samer Nashed, Justin Svegliato, and Shlomo Zilberstein. 2021. Ethically compliant planning within moral communities. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* 188–198.

[31] Nils J. Nilsson. 1980. Chapter 3 - Search Strategies for Decomposable Production Systems. In *Principles of Artificial Intelligence*, Nils J. Nilsson (Ed.). Morgan Kaufmann, San Francisco (CA), 99–129. https://doi.org/10.1016/B978-0-934613-10-1.50010-4

[32] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A Rodriguez-Aguilar. 2020. A structural solution to sequential moral dilemmas. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems.* 1152–1160.

[33] Diederik M Roijers, Shimon Whiteson, Ronald Brachman, and Peter Stone. 2017. *Multi-objective decision making.* Springer.

[34] Diederik M Roijers, Shimon Whiteson, Ronald Brachman, and Peter Stone. 2017. *Multi-objective decision making.* Springer. 39–41 pages.

[35] Robert Shaver. 2023. Egoism. In *The Stanford Encyclopedia of Philosophy* (Spring 2023 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

[36] Mihail Stojanovski, Nadjet Bourdache, Grégory Bonnet, and Abdel-Illah Mouaddib. 2024. Ethical Markov Decision Processes with Moral Worth as Rewards. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems.* 2486–2488.

[37] Justin Svegliato, Samer B Nashed, and Shlomo Zilberstein. 2021. Ethically compliant sequential decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11657–11665.

[38] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. 2020. Implementations in machine ethics: A survey. *ACM Computing Surveys (CSUR)* 53, 6 (2020), 1–38.

[39] Wendell Arnold Wallach. 2014. Moral Machines and Human Ethics. In *Sociable Robots and the Future of Social Relations (Frontiers in Artificial Intelligence and Applications, Vol. 273)*, Johanna Seibt, Raul Hakli, and Marco Nørskov (Eds.). IOS Press, 19–20.

[40] DJ677750 White. 1982. Multi-objective infinite-horizon discounted Markov decision processes. *Journal of mathematical analysis and applications* 89, 2 (1982), 639–647.