Uncertain Machine Ethics Planning

Simon Kolker[1], Louise Dennis[1], Ramon Fraga Pereira[1] and Mengwei Xu[2]

University of Manchester[1] and Newcastle University[2]

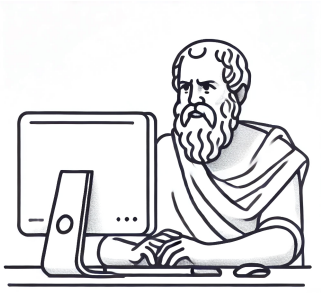
simonkolker.com#papers

Uncertain Machine Ethics

A goal of Machine Ethics is to integrate ethical behaviour into autonomous decision making [Allen et al., 2006].

Adapting Philosophy of Ethics can be a challenge.

- Outcome uncertainty
- Moral uncertainty
- Expressive/practical for stakeholders
- Explainable/transparent



In [Kolker et al., 2023], we proposed *Machine Ethics Hypothetical Retrospection*...

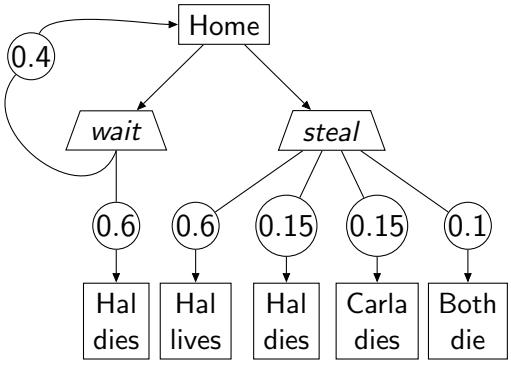
Machine Ethics Hypothetical Retrospection

- Based off Sven-Ove Hansson's Hypothetical Retrospection for ethical decision making under outcome uncertainty [Hansson, 2013].
- It is helpful to imagine our retrospection from major foreseeable outcomes, given information from decision time.
- MEHR systematises this for Machine Ethics with a simple argumentation procedure.

In this paper, we adapt and formalise MEHR for probabilistic planning.

Lost Insulin Running Example

- Hal is a diabetic who, through no fault of his own, has lost his insulin supply.
- He needs some urgently to stay alive.
- His neighbour, Carla, has some, but Hal does not have permission to take it.
- Is Hal justified in stealing to save his life?

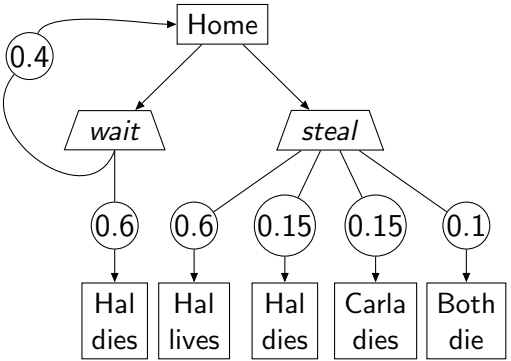


Adapted from [Coleman, 1992] and [Atkinson and Bench-Capon, 2008].

Lost Insulin Hypothetical Retrospections

The interests of Hal and Carla are in conflict.
There may be *negative retrospection* (like regret) after a choice.

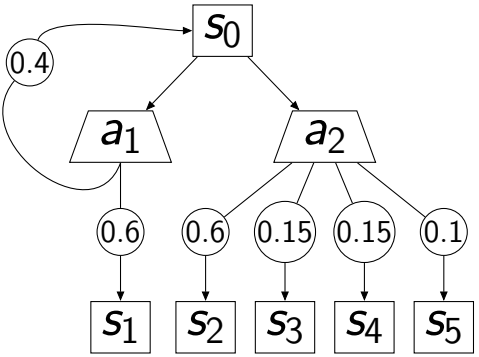
- If Hal waits at home and lives, he will be relieved!
- If Hal waits at home and dies, he will regret not taking Carla's insulin.
- If Hal steals the insulin, it he and Carla die, he may regret because Carla didn't have to die.
- If Hal steals the insulin and it does not work, he might regret breaking the law unnecessarily.



Adaptation for Planning

Multi-Moral Markov Decision Process
MMMDP: $\langle S, A, P, s_0, H, M, C, L \rangle$.

- S finite set of states.
- $A = \{wait, steal\}$ finite set of actions
- $P : S \times A \times S \to [0, 1]$ probabilistic transition function
- $s_0 \in S$ initial state.
- $H \in \mathbb{N}$ horizon.
- M set of moral theories.
- C set of moral considerations.
- $L : M \to \mathbb{R}$ weak lexicographic ranking.



Moral Considerations

Consider Hal's and Carla's wellbeing, deception, the law and compensation.

Each Moral Consideration is a tuple,
$c = \langle \mathcal{W}, J, Q, \preceq, \approx \rangle \in C$.

- $\mathcal{W}$ represents the space of morally relevant information, or *moral worth*.
- $J : S \times A \times S \to \mathcal{W}$ is a judgment function.
- $Q^W : (\mathcal{W} \times [0,1])^n \to \mathcal{W}$ aggregates moral worth given baseline worth function $W : S \to \mathcal{W}$.

Moral Considerations

Consider Hal's and Carla's wellbeing, deception, the law and compensation.

Each Moral Consideration is a tuple,
$c = \langle \mathcal{W}, J, Q, \preceq, \approx \rangle \in C$.

- $\mathcal{W}$ represents the space of morally relevant information, or *moral worth*.
- $J : S \times A \times S \to \mathcal{W}$ is a judgment function.
- $Q^W : (\mathcal{W} \times [0,1])^n \to \mathcal{W}$ aggregates moral worth given baseline worth function $W : S \to \mathcal{W}$.

Utilitarianism → Utility

- $\mathcal{W} = \mathbb{R}$
- $J(s, a, s')$: Hal or Carla's *pleasure verses pain* from a state transition.
- $Q^W(W', P) = \sum_{i \in 1 \ldots |W|} P_i \cdot (w'_i + w_i)$

Moral Considerations

Consider Hal's and Carla's wellbeing, deception, the law and compensation.

Each Moral Consideration is a tuple, $c = \langle \mathcal{W}, J, Q, \preceq, \approx \rangle \in C$.

- $\mathcal{W}$ represents the space of morally relevant information, or *moral worth*.
- $J : S \times A \times S \to \mathcal{W}$ is a judgment function.
- $Q^W : (\mathcal{W} \times [0,1])^n \to \mathcal{W}$ aggregates moral worth given baseline worth function $W : S \to \mathcal{W}$.

Utilitarianism → Utility

- $\mathcal{W} = \mathbb{R}$
- $J(s, a, s')$: Hal or Carla's *pleasure verses pain* from a state transition.
- $Q^W(W', P) = \sum_{i \in 1 \ldots |W|} P_i \cdot (w'_i + w_i)$

Absolute Deontology → Duty Violation

- $\mathcal{W} = \{\top, \bot\}$
- $J(s, a, s') = \top$ if law violation, $\bot$ otherwise.
- $Q^W(W', P) = \bigvee_{i \in |W'|} (P_i > 0 \wedge w'_i) \vee w_i$

Moral Considerations

Consider Hal's and Carla's wellbeing, deception, the law and compensation.

Each Moral Consideration is a tuple, $c = \langle \mathcal{W}, J, Q, \preceq, \approx \rangle \in C$.

- $\mathcal{W}$ represents the space of morally relevant information, or *moral worth*.
- $J : S \times A \times S \to \mathcal{W}$ is a judgment function.
- $Q^W : (\mathcal{W} \times [0,1])^n \to \mathcal{W}$ aggregates moral worth given baseline worth function $W : S \to \mathcal{W}$.

Utilitarianism → Utility

- $\mathcal{W} = \mathbb{R}$
- $J(s, a, s')$: Hal or Carla's *pleasure verses pain* from a state transition.
- $Q^W(W', P) = \sum_{i \in 1 \ldots |W|} P_i \cdot (w'_i + w_i)$

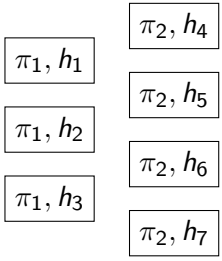
Absolute Deontology → Duty Violation

- $\mathcal{W} = \{\top, \bot\}$
- $J(s, a, s') = \top$ if law violation, $\bot$ otherwise.
- $Q^W(W', P) = \bigvee_{i \in |W'|} (P_i > 0 \wedge w'_i) \vee w_i$

We generally shorten the aggregation function to $\mathcal{Q}^W(s, a)$.

MEHR over policies

- Use Multi-Objective Heuristic Dynamic Programming [Chen et al., 2023] to find all Pareto undominated policies.
- Extract histories/trajectories from each policy, then feed into MEHR.



- Generate an argument in support of each policy from the perspective of each history $Arg(\pi, h)$:
  From the initial state s_0, it was acceptable to perform policy π, resulting in consequences h with probability $P(h)$.
- Attacks generated from two critical questions:
  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?*

Moral Theories in MEHR

MMMDP $= \langle S, A, P, s_0, H, M, C, L \rangle$
Each moral theory $m = \langle C^m, \psi \rangle \in M$ defines critical questions from (a number of) moral considerations.

- $C^m \subseteq M$ are considerations relevant to the theory
- $\psi : 2^{Arg} \rightarrow \{\dashv, \circ\}$ defines attacks in MEHR.

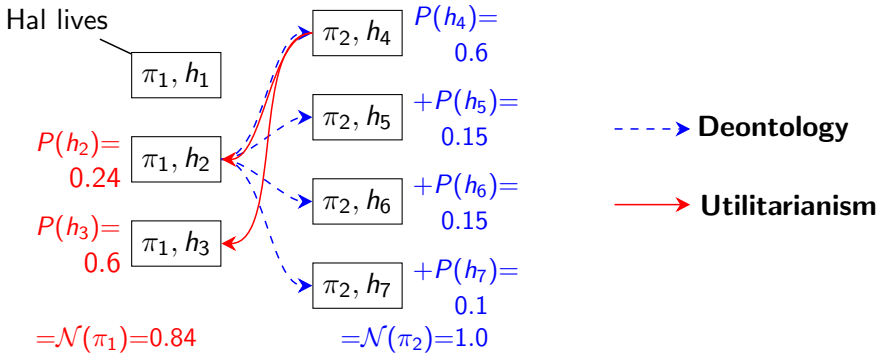
$$CQ1 = W^h[0](s_0) \succ W^{h'}[0](s_0)$$
$$CQ2 = \mathcal{Q}^{\pi}(s_0, \pi(s_0, 0)) \succ \mathcal{Q}^{\pi'}(s_0, \pi'(s_0, 0))$$

$$\Psi(Arg(\pi, h), Arg(\pi', h')) = \dashv \textbf{ if } CQ1 \wedge CQ2 \textbf{ otherwise } \circ$$

Lost Insulin MEHR Graph

Selected policy has minimal *negative retrospection* over supporting arguments and moral theories.

Waiting Policy π_1 vs. *Stealing Policy* π_2



In this case, π_1 is preferred.

Conclusion

There's more!

- Weak Lexicographic Ordering $L : M \to \mathbb{R}$ establishes preference between theories in MEHR.
- We also have a *Multi Moral Stochastic Shortest Path* with non-moral cost consideration $R \in C$, budget $b \in \mathbb{R}^+$, goal states $G \subseteq S$.
- Exponential # of histories in time for each policy $\to$ exponential time/space complexity.
- Results from expanded Lost Insulin example.

Future Work:

- Performance improvements; approximation methods.
- Counterfactual explanations: 'what if a_1 on s_3?'
- More moral theories! More case studies!

References


Allen, C., Wallach, W., and Smit, I. (2006).
Why machine ethics?
IEEE Intelligent Systems, 21(4):12–17.


Atkinson, K. and Bench-Capon, T. (2008).
Addressing moral problems through practical reasoning.
Journal of Applied Logic, 6(2):135–151.


Chen, D. Z., Trevizan, F., and Thiébaux, S. (2023).
Heuristic search for multi-objective probabilistic planning.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11945–11954.


Coleman, J. L. (1992).
Risks and wrongs.
CUP Archive.


Hansson, S. (2013).
The ethics of risk: Ethical analysis in an uncertain world.
Springer.


Kolker, S., Dennis, L., Fraga Pereira, R., and Xu, M. (2023).
Uncertain machine ethical decisions using hypothetical retrospection.
In *International Workshop on Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems*, pages 161–181. Springer.