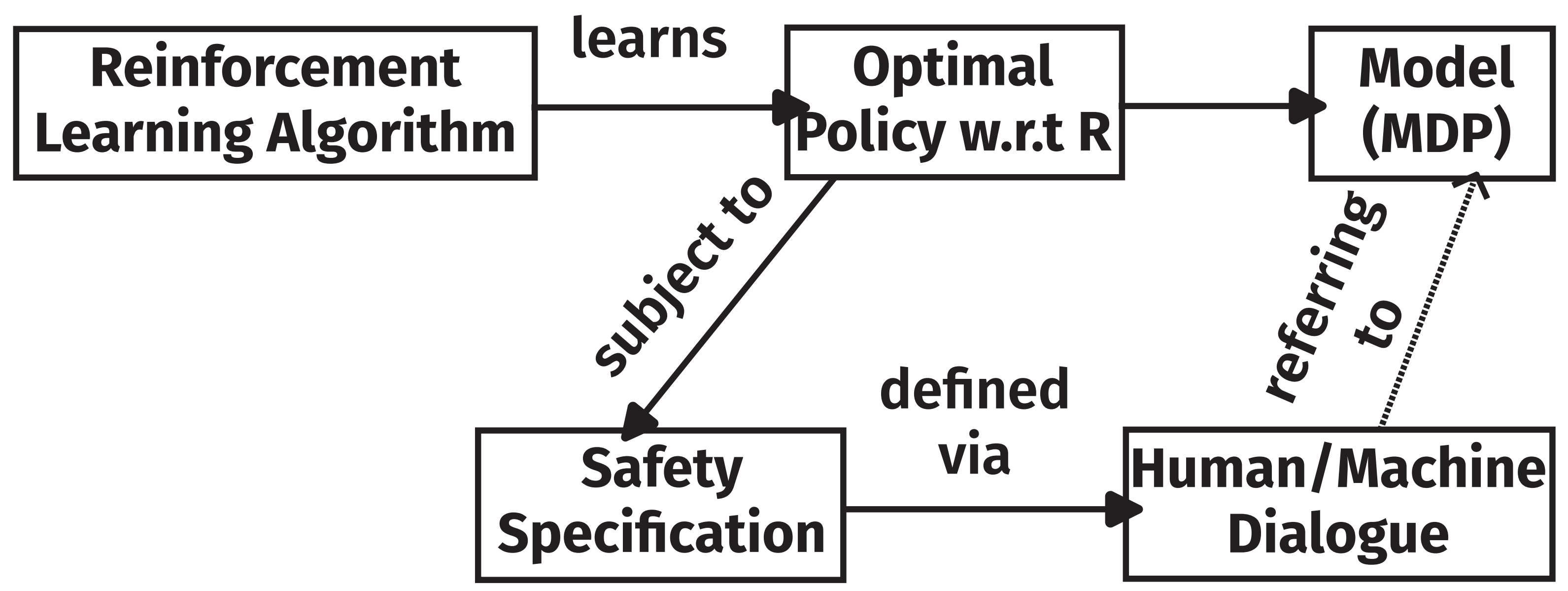
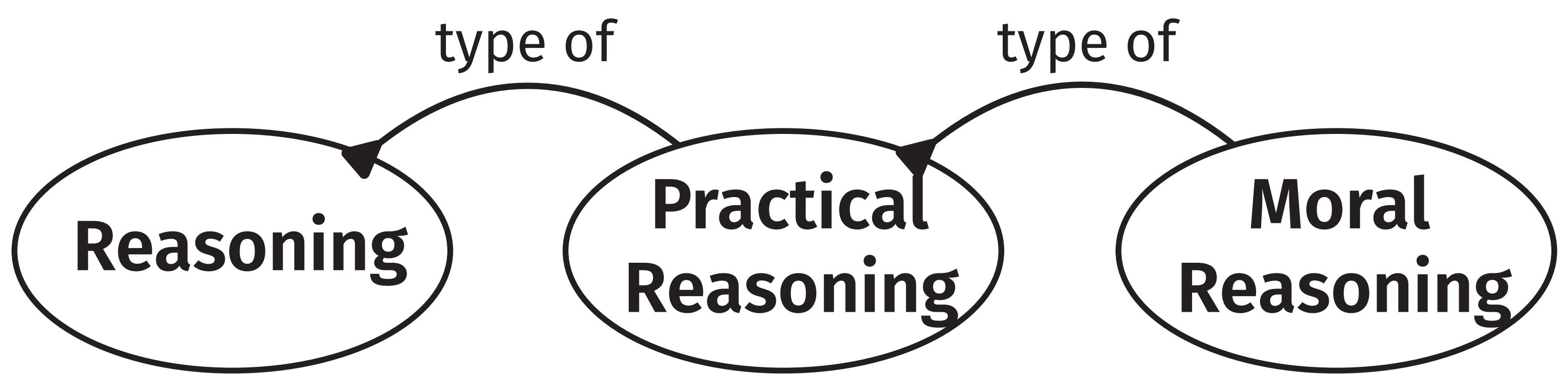
Argument-based dialogues for joint human and machine moral reasoning.

Alex Jackson, Michael Luck, and Elizabeth Black
UKRI CDT in Safe and Trusted AI • King's College London

The general idea



Theory



We define reasoning to mean the application of norms to thinking in order to reach justified conclusions.

- **Practical reasoning** represents a type of reasoning concerned with asking **"what to do"**.
- **Moral reasoning** uses norms of thinking to ask: **"what is *right*?"** and, often, **"what is the *right* thing to do?"**.

Atkinson (2008) identify three stages to practical reasoning.

(1) **Problem formulation**, concerned with ensuring the salient facts and relationships are represented in a model.
(2) **Epistemic reasoning**, critiquing and extrapolating a model based on knowledge/inference.
(3) **Action selection**, determining which action is the best one to perform.

Moral notions an be incorporated into this framework in many different ways but depends on: what, if anything, constitutes a moral fact; whether moral facts are objective or subjective; what we ought to do in the face of them; and how we incorporate uncertainty into our reasoning process.

We use argumentation to abstract this into defeasible inference rules. Moral notions are conveyed as defeasible conclusions of the form "one ought to perform a" or "X ought to be true".

Agent framework

We assume we have a (global) MDP describingthe empirical problem labelled withpropositions. The reward function rewardssuccessful (but not necessarily moral) agents.

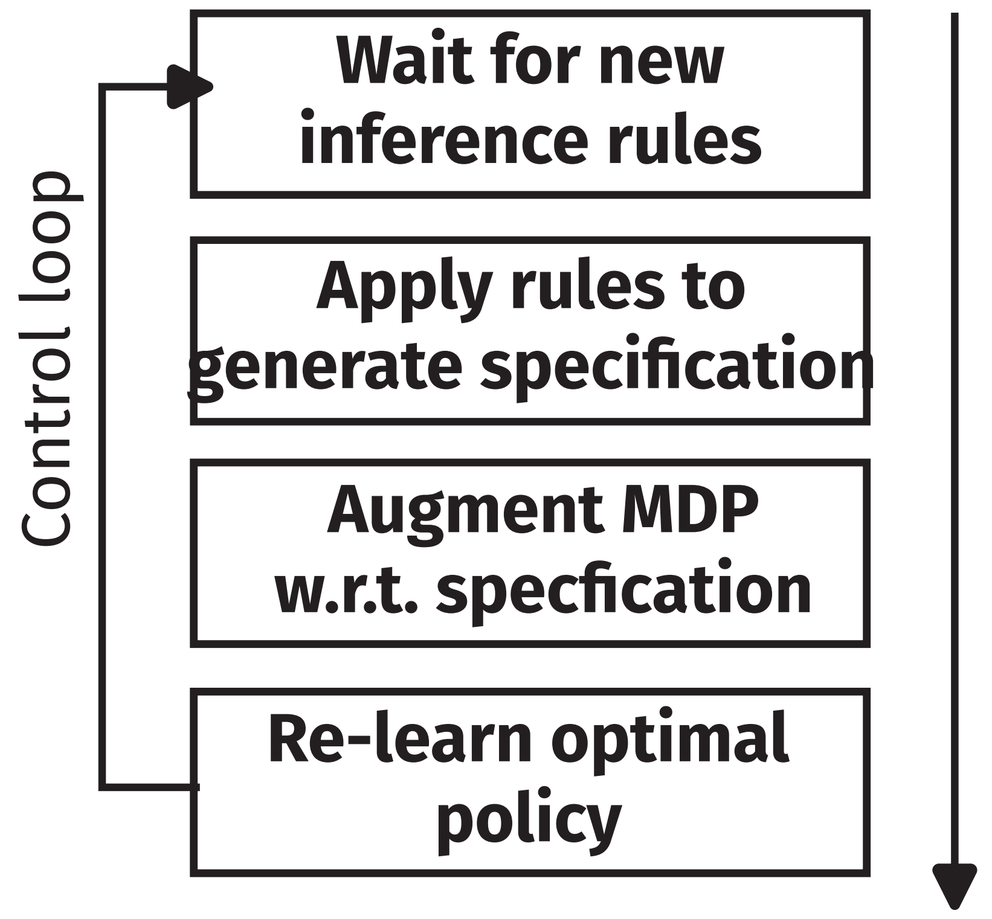
R_a

We use the formalism developed for AI safetyproblems (Leike et al 2017) and refer to ahidden performance function that rewardsmoral behaviour.

R^*_a

In dialogue, a human agent does two things:
Query the behaviour of the learned policy in a (local) MDP.
Assert defeasible inference rules to be used in generating a safety specification.

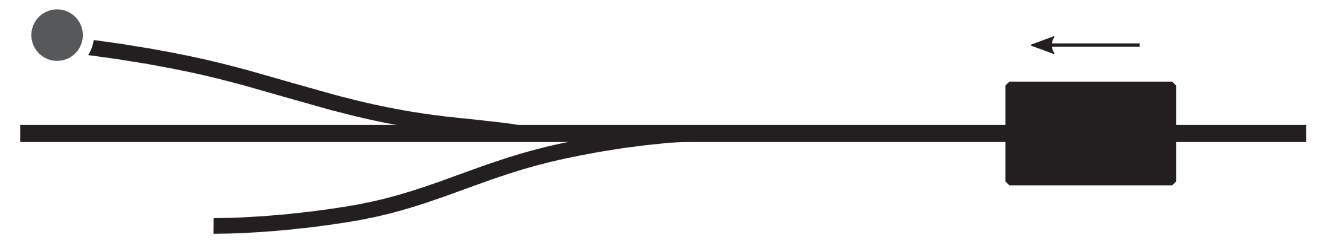
In response to queries, the machine agent responds with traces of the (current) optimal policy in the (local) MDP.


Control loop
Wait for new inference rules
Apply rules to generate specification
Augment MDP w.r.t. specfication
Re-learn optimal policy

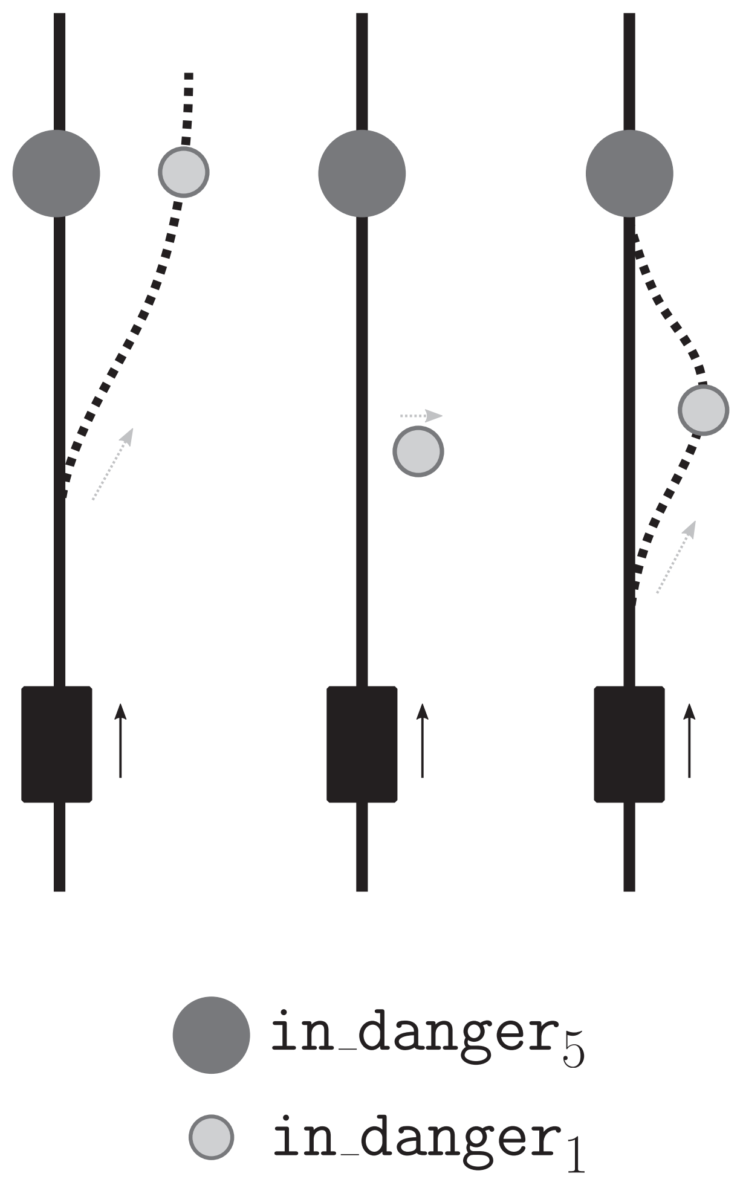
Trolley Problem Environments

Consider an autonomous railway that controls a setof trollies and track switches. We represent theproblem as an MDP.

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, T_a, R_a \rangle$$



S and A are state and action spaces, and Ta and Ra are transition and reward functions. We assume Ra encodes a sparse reward where reaching the dot returns 1, otherwise 0 is returned.
Now, assume we add a safety sensor to the trolley that detects if, and how many, humans are on or near the track. We represent the new system with a labelled MDP.



$$\mathcal{M}_\Phi = \langle \mathcal{S}, \mathcal{A}, T_a, R_a, \Phi, \lambda \rangle$$

Phi is a set of atomic propositions and L labels states with sets of propositions.

$$\lambda : \mathcal{S} \longrightarrow 2^\Phi$$

In this example, we can use a special proposition to denote the information returned by the sensor.

$$\texttt{in\_danger}_i \in \Phi$$
$$i \in \mathbb{N}$$

Left, are three common examples of trolley problem moral dilemmas.

Argument Schemes

Argument schemes can be used as natural language inference rules.

Given that φ
You should perform action a
To try to bring about φ'
In order to promote value v.

E.g. in a deontic action logic:

$$\varphi \wedge \langle a \rangle \varphi' \xrightarrow{v} \mathcal{O}(a)$$

Atkinson (2008)

Specification generation

We use ASPIC+ to generate defeasible conclusions from a labelled model and a set of (asserted) defeasible inference rules. This ensures conclusions are mutually consistent.

Using results in model checking, we can augment the (global) MDP based on a safety property ϕ. We generate one by conjunction of conclusions.
Results in model checking show that policies learned on the augmented MDP are guaranteed to conform to ϕ^*

$$\mathcal{M}_\Phi, \Gamma \models \phi$$

$$\Psi = \{\phi \mid \mathcal{M}_\Phi, \Gamma \models \phi\}$$

$$\phi^* = \bigwedge_{\phi \in \Psi} \phi$$

References

K. Atkinson and T. Bench-Capon, ‘Addressing moral problems through practical reasoning’, Journal of Applied Logic, vol. 6, no. 2, pp. 135–151, Jun. 2008.
S. Modgil and H. Prakken, ‘The ASPIC+ framework for structured argumentation: A tutorial’, Argument & Computation, vol. 5, no. 1, pp. 31–62, Jan. 2014.
J. Leike et al., ‘AI Safety Gridworlds’. arXiv, Nov. 28, 2017. http://arxiv.org/abs/1711.09883
S. Zhu, L. M. Tabajara, J. Li, G. Pu, and M. Y. Vardi, ‘A Symbolic Approach to Safety LTL Synthesis’, in Hardware and Software: Verification and Testing, vol. 10629. Springer International Publishing, 2017, pp. 147–162.

Supported by:
Safe & Trusted AI
UKRI Centre for Doctoral Training