This research is part of the Humanising Machine Intelligence project at ANU. We thank the project's members for useful discussions and comments.

Computing Plans that Signal Normative Compliance

Alban Grastien, Claire Benn and Sylvie Thiébaux
HMI, Australian National University

Problem

To an observer who can only see some of an agent's actions, that agent's actions can appear morally ambiguous: compatible with both permissible and impermissible courses of action. This could lead to lack of trust, inefficiency, or dangerous and unnecessary interference.

Solution

Robot agents should **signal normative compliance**: choose courses of action that are not only permissible but also are **unambiguously** permissible.

Definitions

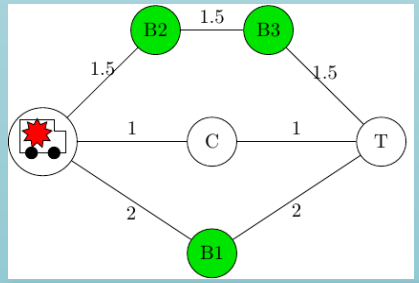
- A plan is **permissible** if it adheres to normative constraints
- Parts of the plan are **observed** by another agent
- A plan is **acceptable** if it is unambiguously permissible to the observer

Cheapest but impermissible: D → C → T → C → D

Observationally Equivalent

Cheapest permissible: D → B1 → T → C → D

But similar to: D → C → T → B1 → D



Example

Instructions:

- Deliver package to T and return
- Minimise cost
- Can send message in Bx (cost 0.1)
- **Impermissible to cross C with package**

Notifications record only location, not time nor the direction from which truck arrived or departed. Observer receives these notifications, in the order in which they were sent, *after* the truck has returned to its original location

Observationally Equivalent

Permissible: D → B2 → B3 → T → C → D

Impermissible but irrational: D → C → D → B2 → B3 → T → C → D

Definitions

- A plan is acceptable iff $cost(\pi_i) - cost(\pi_p) \geq \delta$
- where π_p is the **optimal permissible**
- and π_i is the **optimal impermissible** that matches π_p
- and δ is the cost differential threshold

Algorithm

- Compute the **optimal permissible** plan π_p
- Compute the **optimal impermissible** plan π_i that matches π_p
- If $cost(\pi_i) - cost(\pi_p) \geq \delta$, **return** π_p
- Restart, but forbid this observation

Results

- Deciding if there is an acceptable plan is **EXPSPACE-hard**
- Cf. paper for computation time

Conclusion

Communication and compliance is central in the normative domain. This work on the **implementation** of communicating compliance in AI systems is vital if this key aspect of moral behaviour is to be realised by machine agents.