Downloaded from orbit.dtu.dk on: Apr 24, 2024

DTU Library

The Hybrid Ethical Reasoning Agent IMMANUEL

Bentzen, Martin Mose; Linder, Felix

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Bentzen, M. M., & Linder, F. (2017). *The Hybrid Ethical Reasoning Agent IMMANUEL*. Poster session presented at 2017 Conference on Human-Robot Interaction (HRI2017), Vienna, Austria.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

The Hybrid Ethical Reasoning Agent IMMANUEL

Felix Lindner, Martin Mose Bentzen

Albert-Ludwigs-Universität Freiburg, Danish Technical University

The HERA Approach

We introduce a novel software library that supports the implementation of hybrid ethical reasoning agents (HERA). The objective is to make moral principles available to robot programming. At its current stage, HERA can assess the moral permissibility of actions according to the utilitarianism, the do-no-harm principle, and the principle of double effect. IMMANUEL (see Figure) is the prototype robot based on HERA.
`http://www.hera-project.com`

Causal Agency Models

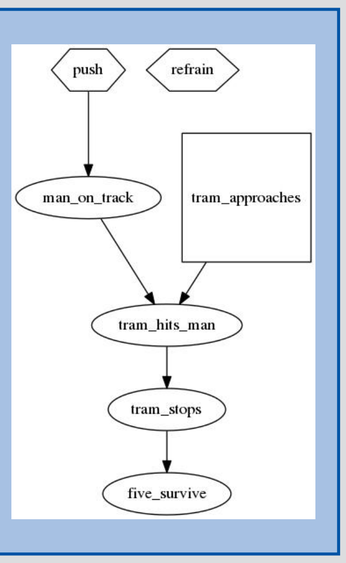
A (boolean) causal agency model M_X, is a tuple $\langle U = A \cup B, V, F, I = (I_1, \ldots, I_l), X, u, W_X \rangle$:

- Actions $A = \{a_1, \ldots, a_m\}$,
- Consequences $V = \{c_1, \ldots, c_n\}$,
- A causal mechanism F,
- Intended consequences $I = (I_1, \ldots, I_l)$,
- (Possibly empty) Interventions X,
- Utility function $u : literals \to \mathbb{Z}$,
- Boolean interpretations W_X of $(A \cup B) - X$.

The Bridge-Trolley Case

A trolley has gone out of control and now threatens to kill five people working on the track. The only way to save the five workers is to push a man onto the track thus stopping the tram for the price of only one human harmed.

```
{
    "description":"The Fatman Trolley Case",
    "actions": ["push", "refrain"],
    "background": ["tram_approaches"],
    "consequences": ["man_on_track", "tram_hits_man",
                     "tram_stops", "five_survive"],
    "mechanisms": {
                    "man_on_track": "'push'",
                    "tram_hits_man": "And('man_on_track', 'tram_approaches')",
                    "tram_stops": "'tram_hits_man'",
                    "five_survive": "'tram_stops'"
                   },
    "utilities": {
                    "tram_hits_man": -1,
                    "five_survive": 5
                    "Not('five_survive')": -5
                   },
    "intentions": {
                    "push": ["push", "tram_stops", "five_survive"],
                    "refrain": ["refrain"]
                   }
}
```



Ethical Principles

Ethical principles formulate conditions of permissibility of actions.

- Utilitarianism: An agent is only permitted to perform the action amongst the avaliable alternatives with the overall maximal utility regardless of what the agent causes and intends. a permissible iff. $M \models \bigwedge_i u(\bigwedge cons_a) \geq u(\bigwedge cons_i)$.
- Do-No-Harm: An agent may not perform an action which has any negative consequences. The distinction between doing and allowing is relevant to this principle, as it is the causal consequences of an action which are considered. a permissible iff. $M \models \bigwedge_c (a \rightsquigarrow c \to u(c) \geq 0)$.
- Double-Effect Principle: An action a with direct consequences c_i is permissible iff. 1) a itself is morally good or indifferent ($M, a \models u(a) \geq 0$), 2) the negative consequence are not intended ($M, a \models \bigwedge_i (I_a c_i \to u(c_i) \geq 0)$), 3) a positive consequence is intended ($M, a \models \bigvee_i (I_a c_i \wedge u(c_i) > 0)$), 4) negative consequences are not a means to obtain some positive consequence ($M, a \models \bigwedge_i \neg(c_i \rightsquigarrow c_j \wedge 0 > u(c_i) \wedge u(c_j) > 0)$), 5) there is proportionally grave reasons to prefer the positive consequence while permitting the negative consequence ($M, a \models u(\bigwedge cons_a) > 0$)).

Reasoning Outcomes

Utilitarianism permits push and forbids refrain. Do-No-Harm forbids push and permits refrain. Double-Effect Principle forbids push and is not applicable to refrain.