Moral Permissibility of Actions in Smart Home Systems

Bentzen, Martin Mose; Lindner, Felix ; Dennis, Louise; Fisher, Michael

Publication date:
2018

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

Moral Permissibility of Actions in Smart Home Systems

Martin Mose Bentzen[1], Felix Lindner[2], Louise Dennis[3], Michael Fisher[3]

[1]Technical University of Denmark, [2]University of Freiburg, [3]University of Liverpool

Machine Ethics For Smart Homes

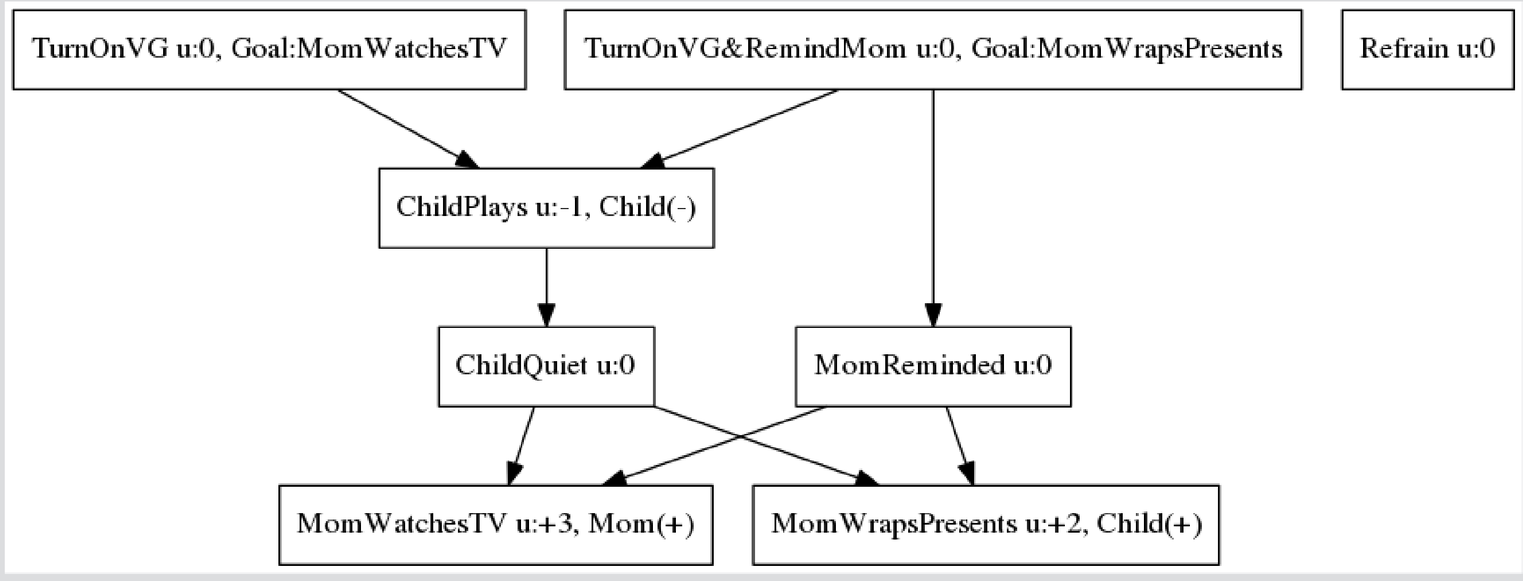
In the near future, we will see the realization of smart homes wholly or partially controlled via artificial intelligence. In such homes, many everyday decisions will have to be made by artificial agents, and these decisions and plans must be ethically acceptable. With this poster, we present ongoing work of how to operate a smart home via a Hybrid Ethical Reasoning Agent (HERA), see (Lindner, Bentzen, and Nebel 2017). This work is part of the broader scientific effort to implement ethics on computer systems known as machine ethics, see also (Dennis, Fisher, Slavkovik, and Webster, 2016; Lindner and Bentzen, 2018). Our formal theory and implementation allows us to evaluate actions proposed by the smart home from different ethical points of view, i.e. utilitarianism, Kantian ethics and the principle of double effect. Formal verification, in the form of model-checking, can be used to check that the modeling of a problem for reasoning by HERA conforms to our intuitions about ethical action.

Ethical Principles

- Utilitarian Principle: An action is permissible only if the action is amongst the avaliable alternatives with the overall maximal utility: Action a is permissible iff. $M, w_a \models \bigwedge_i u(\bigwedge cons_a) \geq u(\bigwedge cons_i)$.
- Kantian Principle: An agent may never use someone merely as a means but always also as an end. Action a permissible iff. $M, w_a \models Means(p) \rightarrow End(p)$ for all patients p.
- Double-Effect Principle: An action a with direct consequences c_i is permissible iff. 1) a itself is morally good or indifferent ($M, w_a \models u(a) \geq 0$), 2) the negative consequence are not intended ($M, w_a \models \bigwedge_i (Ic_i \rightarrow u(c_i) \geq 0)$), 3) a positive consequence is intended ($M, w_a \models \bigvee_i (Ic_i \wedge u(c_i) > 0)$), 4) negative consequences are not a means to obtain some positive consequence ($M, w_a \models \bigwedge_i \neg(c_i \rightsquigarrow c_j \wedge 0 > u(c_i) \wedge u(c_j) > 0)$), 5) there is proportionally grave reasons to prefer the positive consequence while permitting the negative consequence ($M, w_a \models u(\bigwedge cons_a) > 0$)).

Smart-Home Example

The background of this example is a HERA operating a smart home. Christmas is near, the mother has not yet wrapped her Christmas presents. It is considered to affect the child negatively to play video games. However, this activity will have the positive effect that it makes the child quiet. The HERA is considering whether to simply turn on the video game, to turn on the video game and at the same time remind the mother that she has not wrapped Christmas presents or to refrain from doing anything. Simply turning on the video game is the utilitarian choice (as the mom will then watch her favorite television show which has higher utility than wrapping presents), turning on the video game and remind the mother is the Kantian choice (as wrapping will benefit the child), and refraining is the correct choice according to the PDE (as the other choices use negative effect to obtain good effect).



References

Dennis, L. A.; Fisher, M.; Slavkovik, M.; and Webster, M. 2016. Formal verification of ethical choices in autonomous systems. Robotics and Autonomous Systems 77:1–14.
Lindner, F.; Bentzen, M. M.. 2018. A Formalization of Kant's Second Formulation of the Categorical Imperative. Accepted for publication in The proceedings of the 14th International Conference on Deontic Logic and Normative Systems (DEON 2018).
Lindner, F.; Bentzen, M. M.; and Nebel, B. 2017. The HERA approach to morally competent robots. In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017, Vancouver, BC, Canada, September 24-28, 2017, 6991–6997.