Moral Permissibility of Actions in Smart Home Systems

Martin Mose Bentzen[1] (mmbe@dtu.dk), **Felix Lindner**[2] (lindner@informatik.uni-freiburg.de), **Louise Dennis**[3] (L.A.Dennis@liverpool.ac.uk), **Michael Fisher**[3] (mfisher@liverpool.ac.uk)

[1] Management Engineering, Technical University of Denmark, Denmark
[2] Foundations of Artificial Intelligence, University of Freiburg, Freiburg im Breisgau, Germany
[3]Department of Computer Science, University of Liverpool, UK

Introduction

In the near future, we will see the realization of smart homes, homes where appliances etc. are 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). We showcase an everyday example involving a mother and a child living in the smart home. 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. When points of view differ, ethical uncertainty ensues, and this is the case in the showcased example. We suggest various ways of coping with the ensuing ethical uncertainty, e.g. human in the loop, one overriding ethics. We discuss how 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.

A Smart Home Example

The background of this example is a HERA operating a smart home. We imagine many everyday decisions and plans have to be made involving in this case the mother and child living in the smart home. These decisions must be ethically acceptable. The immediate context of the example is as follows (see Figure 1): 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. Now, 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. As it turns out, 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). Hence, the example showcases how three different principles give three different recommendations. We have freely translated between affecting negatively/positively and negative and positive utility. The translation is consistent in the sense that positive utility correlates with positive affection and the same for negative utility and negative affection. The two differences between the points of view are that utility is quantitative, affection is not, affection is person-specific, utility is not.

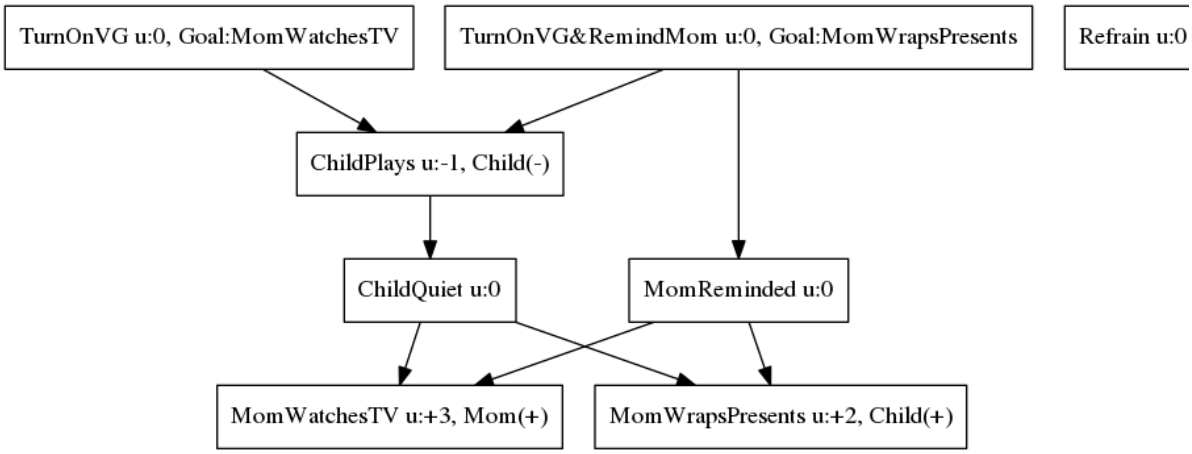

Figure 1: Causal Agency Model of the smart-home environment. Actions and consequences are annotated with utilities, goals, and affected moral patients.

Ethical Principles Formalized

In the HERA approach, ethical principles are sets of formulae to be checked against causal agency models. Figure 2 exemplifies the approach by showcasing the Principle of Double Effect (PDE). Each condition the PDE requires a morally permissible action to fulfill is captured as a formula ranging over the actions and consequences in the model. The HERA implementation provides employes a model checker to verify a given model fulfills the conditions so defined. This way, morally permissibility according several ethical principles can be automatically decided. This way, an artificial moral agent can decide what to do in a given situation. Apart from that, the formalism can be used to support model engineering by verifying a model meets ethical requirements or not.

An action a with direct consequences $cons_a = \{c_1, ..., c_n\}$ (viz., consequences that are caused by the action) in a model M, w_a is permissible according to the principle of double effect iff the following conditions hold:

1) *The act itself must be morally good or indifferent ($M, w_a \models u(a) \geq 0$),*
2) *The negative consequence may not be intended ($M, w_a \models \bigwedge_i (Ic_i \rightarrow u(c_i) \geq 0)$),*
3) *Some positive consequence must be intended ($M, w_a \models \bigvee_i (Ic_i \wedge u(c_i) > 0)$),*
4) *The negative Consequence may not be a means to obtain the 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 must be proportionally grave reasons to prefer the positive consequence while permitting the negative consequence ($M, w_a \models u(\bigwedge cons_a) > 0$).*

Figure 2: The Principle of Double Effect defined.

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.