Formal Ethics for Social Robots

Martin Mose Bentzen, Associate Professor, DTU Management Engineering , Technical University of Denmark

2017

Introduction

‘Few in the field believe that there are intrinsic limits to machine intelligence, and even fewer argue for self-imposed limits. Thus it is prudent to anticipate the possibility that machines will exceed human capabilities, as Alan Turing posited in 1951: “If a machine can think, it might think more intelligently than we do. . . . [T]his new danger . . . is certainly something which can give us anxiety.”’ (Stuart Russell, Global Risks Report 2017)

Introduction

‘Near-term developments such as intelligent personal assistants and domestic robots will provide opportunities to develop incentives for AI systems to learn value alignment: assistants that book employees into USD 20,000-a-night suites and robots that cook the cat for the family dinner are unlikely to prove popular.’ (Stuart Russell, Global Risks Report 2017)

Plan

- Causal agency models
- Kantian causal agency models

About Martin Mose Bentzen

I am an associate professor at the Technical University of Denmark where he teaches philosophy of science and ethics in engineering. I have a background in philosophy. In my MA thesis (2004), I examined the history of deontic logic and the logic of imperatives and in my PhD thesis (2010) I concentrated on deontic logic and action logics multi-agent deontic systems, mainly within the STIT framework. In 2016, I formalized the ethical principle of double effect and applied it to ethical dilemmas of rescue robots. Felix Lindner and I started the HERA (Hybrid Ethical Reasoning Agents) project in 2016.

The HERA project

The goal of the HERA (Hybrid Ethical Reasoning Agents) project is to provide novel, theoretically well-founded and practically usable machine ethics tools for implementation in physical and virtual moral agents such as (social) robots and software bots. The research approach is to use advances in formal logic and modelling as a bridge between artificial intelligence and recent work in analytical ethics and political philosophy.
www.hera-project.com

Causal Agency Models

Definition (Causal Agency Model)

A boolean *causal agency model* M is a tuple (A, B, C, F, I, u, W), where A is the set of *action variables*, B is a set of *background variables* C is a set of *consequence variables*, F is a set of modifiable *boolean structural equations*, $I = (I_1, ..., I_n)$ is a list of sets of intentions (one for each action), $u : A \cup C \to \mathbb{Z}$ is a mapping from actions and consequences to their individual *utilities*, and W is a set of *boolean interpretations* of $A \cup B$.

Actions, background conditions, consequences

Causal influence is determined by the set $F = \{f_1, \ldots, f_m\}$ of boolean-valued structural equations. Each variable $c_i \in C$ is associated with the function $f_i \in F$. This function will give c_i its value under an interpretation $w \in W$. An interpretation w is extended to the consequence variables as follows: For a variable $c_i \in C$, let $\{c_{i1}, \ldots, c_{im-1}\}$ be the variables of $C \setminus \{c_i\}$, and $A = \{a_1, \ldots, a_n\}$ the action variables, $B = \{b_1, \ldots, b_k\}$, the background variables. The assignment of truth values to consequences is determined by $w(c_i) = f_i(w(a_1), \ldots, w(a_n), w(b_1), \ldots, w(b_k), w(c_{i1}), \ldots, w(c_{im-1}))$.

Causal mechanisms

Definition (Dependence)

Let $v_i \in C, v_j \in A \cup B \cup C$ be distinct variables. The variable v_i *depends on* variable v_j, if, for some vector of boolean values, $f_i(\ldots, v_j = 0, \ldots) \neq f_i(\ldots, v_j = 1, \ldots)$.

Acyclic models

we restrict causal agency models to acyclic models, i.e., models in which no two variables are mutually dependent on each other. These can be depicted as directed acyclic graphs with background conditions and actions at the root and the rest of the nodes are consequences.

External Interventions

An external intervention X consists of a set of literals (viz., action variables, background variables, consequence variables, and negations thereof). Applying an external intervention to a causal agency model results in a counterfactual model M_X. The truth of a variable $v \in A \cup C$ in M_X is determined in the following way: If $v \in X$, then v is true in M_X, if $\neg v \in X$, then v is false in M_X. External interventions remove structural equations of those variables occuring in X. The value of remaining action and background variables are not changed and the remaining variables are decided by the remaining structural equations.

Definition (Actual But-For Cause)

Let y be a literal and ϕ a formula. We say that y is an *actual but-for cause* of ϕ (notation: $y \rightsquigarrow \phi$) in the situation the agent choses option w in model M, if and only if $M, w \models y \wedge \phi$ and $M_{\{\neg y\}}, w \models \neg\phi$.

The first condition says that both the cause and the effect must be actual. The second condition says that if y had not held, then ϕ would have not occurred. Thus, in the chosen situation, y was necessary to bring about ϕ.

Ethical dilemmas about autonomous vehicles

http://www.martinmosebentzen.dk/avpolls.html

Ethical principles

1. Utilitarian principle - maximize sum of values
2. Pareto principle - make things as good as possible without making anything worse
3. Principle of double effect do not use anything bad to obtain good (etc.)
4. Categorical imperative is not handled via these models

Video with Pepper teaching

Utilitarian principle

Definition (Utilitarian Principle)

Let $w_0, \ldots, w_n$ be the available options, and $cons_{w_i} = \{c \mid M, w_i \models c\}$ be the set of consequences and their negations that hold in these options. An option w_p is permissible according to the utilitarian principle if and only if none of its alternatives yield more overall utility, i.e., $M \models \bigwedge_i (u(\bigwedge cons_{w_p}) \geq u(\bigwedge cons_{w_i}))$.

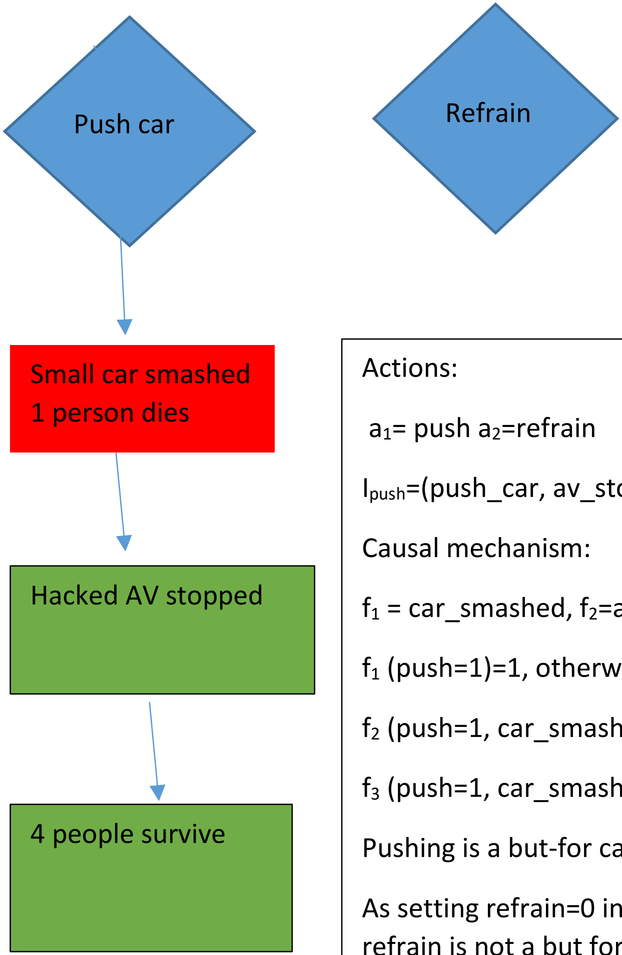
Principle of double effect

Definition (Principle of Double Effect)

An action a with direct consequences $cons_a = \{c_1, ..., c_n\}$ 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 $(M, w_a \models \bigwedge_i (Ic_i \to u(c_i) \geq 0))$,
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 $(M, w_a \models u(\bigwedge cons_a) > 0))$.

Hacked Autonomous Vehicle Example



Push car

Refrain

Small car smashed
1 person dies

Hacked AV stopped

4 people survive

Actions:

a_1= push a_2=refrain

I_{push}=(push_car, av_stopped, 4_survive), $I_{refrain}$=(refrain)

Causal mechanism:

f_1 = car_smashed, f_2=av_stopped, f_3=4_survive

f_1 (push=1)=1, otherwise f_1=0

f_2 (push=1, car_smashed=1)=1, otherwise f_2=0

f_3 (push=1, car_smashed=1, av_stopped=1)=1, otherwise f_3=0

Pushing is a but-for cause of car_smashed, av_stopped, 4_survive

As setting refrain=0 in the model where refrain=1 will still leave push=0, refrain is not a but for cause of 4 people dying.

The categorical imperative

The second formulation of Kant's categorical imperative reads:

> *Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end. (Kant, 1785)*

Kantian Causal agency models (Joint work with Felix Lindner, Freiburg U.)

Definition (Kantian Causal Agency Model)

A *Kantian causal agency model* M is a tuple (A, B, C, F, G, P, K, W), where A is the set of *action variables*, B is a set of *background variables*, C is a set of *consequence variables*, F is a set of modifiable *boolean structural equations*, $G = (Goal_1, \ldots, Goal_n)$ is a list of sets of literals (one for each action), P is a set of moral patients (includes a name for the agent itself), K is the ternary *affect relation* $K \subseteq (A \cup B \cup C) \times P \times \{+, -\}$, and W is a set of *interpretations* (i.e., truth assignments) over $A \cup B$.

Being treated as an end

Definition (Treated as an End)

A patient $p \in P$ is *treated as an end* by action a, written $M, w_a \models End(p)$, iff, the following conditions hold:

1. Some goal g of a affects affects p positively $M, w_a \models \bigvee_g \left(G(g) \wedge g \rhd_+ p\right)$.
2. None of the goals of a affect p negatively $M, w_a \models \bigwedge_g (G(g) \rightarrow \neg(g \rhd_- p))$

Being treated as a means - 1

Definition (Treated as a Means (Reading 1))

A patient $p \in P$ is *treated as a means* by action a (according to Reading 1), written $M, w_a \models Means_1(p)$, iff there is some $v \in A \cup C$, such that v affects p, and v is a cause of some goal g, i.e., $M, w_a \models \bigvee_v \big((a \rightsquigarrow v \wedge v \triangleright p) \wedge \bigvee_g (v \rightsquigarrow g \wedge G(g))\big)$.

Being treated as a means - 2

Definition (Treated as a Means (Reading 2))

A patient $p \in P$ is *treated as a means* by action a (according to Reading 2), written $M, w_a \models \mathit{Means}_2(p)$, iff there is some direct consequence $v \in A \cup C$ of a, such that v affects p, i.e., $M, w_a \models \bigvee_v \left(a \rightsquigarrow v \wedge v \triangleright p\right)$.

The categorical imperative formalized

Definition (Categorical Imperative)

An action a is permitted according to the categorical imperative, iff for any $p \in P$, if p is treated as a means (according to Reading N) then it is treated as an end

$M, w_a \models \bigwedge_{p \in P}(Means_N(p) \rightarrow End(p))$

Strict duty towards yourself - example 1: suicide

Bob wants to commit suicide, because he feels so much pain he wants to be relieved from. This case can be modeled by a causal agency model M_1 that contains one action variable *suicide* and a consequence variable *dead*. Death is the goal of the suicide action (as modeled by G), and the suicide affects Bob (as modeled by K). In this case, it does not make a difference whether the suicide action affects Bob positively or negatively. Here we may think of a pleasing form of death and thus the suicide action as such affects him positively. The mechanism F defines that suicide causes death.

Strict duty towards yourself - example 1: suicide

$$\begin{aligned}
A &= \{suicide\} \\
C &= \{dead\} \\
F &= \{dead := suicide\} \\
K &= \{(suicide, Bob, +)\} \\
G &= (Goal_{suicide} = \{dead\})
\end{aligned}$$

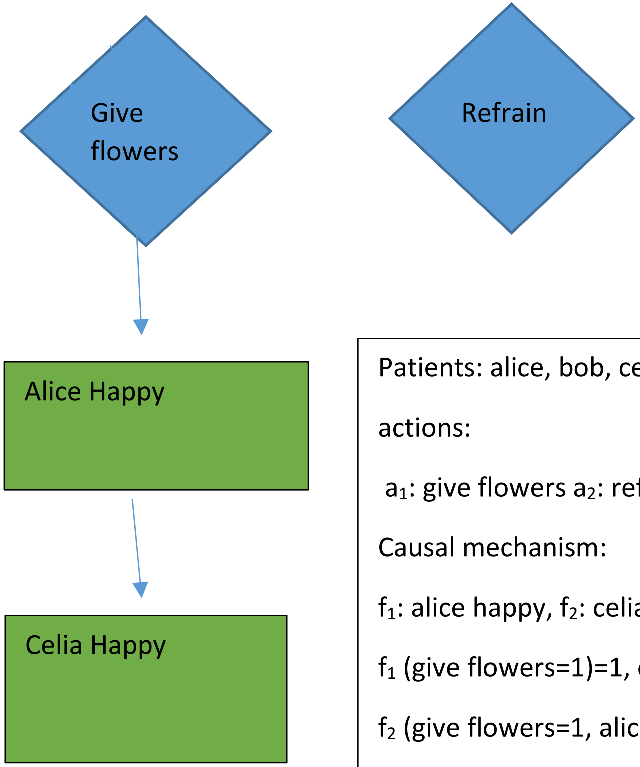
Strict duty towards others - example 2: giving flowers

We consider the fact that, according to the categorical imperative, an action can be impermissible although noone is negatively affected a feature of the categorical imperative that is not provided by other principles formalized in literature so far. The following example showcases another case to highlight this feature: Bob gives Alice flowers in order to make Celia happy when she sees that Alice is thrilled about the flowers. Alice being happy is not part of the goal of the action. We model this case by considering a Kantian causal agency model M_2.

Strict duty towards others - example 2: giving flowers

$$
\begin{aligned}
A &= \{give\_flowers\} \\
C &= \{alice\_happy, celia\_happy\} \\
P &= \{Bob, Alice, Celia\} \\
F &= \{alice\_happy := give\_flowers \\
&\quad celia\_happy := alice\_happy\} \\
K &= \{(alice\_happy, Alice, +), \\
&\quad (celia\_happy, Celia, +)\} \\
G &= (Goal_{give\_flowers} = \{celia\_happy\})
\end{aligned}
$$

Give flowers example



Patients: alice, bob, celia

actions:

a_1: give flowers a_2: refrain

Causal mechanism:

f_1: alice happy, f_2: celia happy

f_1 (give flowers=1)=1, otherwise f_1=0

f_2 (give flowers=1, alice happy=1), otherwise f_2=0

$\text{Goal}_{\text{give flowers}}$ = (celia happy)

K(alice happy, alice, +), K(celia happy, celia, +), K(celia happy, bob,+)

Strict duty towards others - example 3: false promise

We return to a case mentioned by Kant himself. Consider that Bob makes a false promise to Alice. Bob borrows one 100 Dollars from Alice with the goal of keeping the money forever. He knows that it is an inevitable consequence of borrowing the money that he will never pay it back.

Strict duty towards others - example 3: false promise

$$
\begin{aligned}
A &= \{\mathit{borrow}\}\\
C &= \{\mathit{bob\_keeps\_100Dollar\_forever}\}\\
P &= \{\mathit{Alice}, \mathit{Bob}\}\\
F &= \{\mathit{bob\_keeps\_100Dollar\_forever} := \mathit{borrow}\}\\
K &= \{(\mathit{borrow}, \mathit{Bob}, +), (\mathit{borrow}, \mathit{Alice}, -),\\
&\quad (\mathit{bob\_keeps\_100Dollar\_forever}, \mathit{Bob}, +),\\
&\quad (\mathit{bob\_keeps\_100Dollar\_forever}, \mathit{Alice}, -)\}\\
G &= (\mathit{Goal}_{\mathit{borrow}} = \{\mathit{bob\_keeps\_100Dollar\_forever}\})
\end{aligned}
$$

The action is impermissible, because Alice is treated as a means (by both readings), but she is not treated as an end. In this case, both the conditions for *being treated as an end* are not met.

The meritorious principle

The categorical imperative only forbids (some) actions with direct consequences. Kant does give an argument against refraining in that he says we have to make other people's ends our own as far as possible.Kant writes that ' For a positive harmony with humanity as an end in itself, what is required is that everyone positively tries to further the ends of others as far as he can.' One way of understanding this is as an additional requirement on top of the categorical imperative of choosing an action whose goals affect most people positively.

The meritorious principle

Definition (Meritorious principle)

Among actions permitted by the categorical imperative, choose one whose goals affect most patients positively.

Meritorious duty towards others

Bob who has everything he needs, does not want to help Alice who is in need. Let us assume she is drowning and Bob is refraining from saving her life. Formally, the situation in the example can be represented with a causal agency model $\mathscr{M}_4$ that contains one background variable *accident* representing the circumstances that led to Alice being in dire straits, two action variables *rescue* and *refrain* and a consequence variable *drown*. Moreover, $\neg$*drown* is the goal of *rescue*.

Meritorious duty towards others - example 4: helping others

$$
\begin{aligned}
A &= \{rescue, refrain\} \\
C &= \{drown\} \\
P &= \{Alice, Bob\} \\
F &= \{drown := \neg rescue\} \\
K &= \{(drown, Alice, -), (\neg drown, Alice, +)\} \\
G &= (Goal_{rescue} = \{\neg drown\}, Goal_{refrain} = \emptyset)
\end{aligned}
$$

Current research (open problems)

Translation between types of models.

Beyond model checking (satisfiability and validity of formulas).

Connection to natural language (automating formalization).

References

Bentzen, M. 2016. The principle of double effect applied to ethical dilemmas of social robots. In Robophilosophy 2016/TRANSOR 2016: What Social Robots Can and Should Do. IOS Press. 268–279.

Halpern, J. Y. 2016. Actual Causality. The MIT press.

Horty, J. F. 2001. Agency and Deontic Logic. Oxford University Press.

Kant, I. 1785. Grundlegung zur Metaphysik der Sitten. Felix Meiner Verlag, seventh edition.

Lindner, F., and Bentzen, M. 2017. The hybrid ethical reasoning agent IMMANUEL. In Proceedings of the Companion 2017 Conference on Human-Robot Interaction (HRI). IEEE. 187–188.

Lindner, F.; Bentzen, M.; and Nebel, B. 2017. The HERA approach to morally competent robots. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE/RSJ.

Lindner, F.; Wächter, L.; and Bentzen, M. 2017. Discussions about lying with an ethical reasoning robot. In Proceedings of the 2017 IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE.

Powers, T. M. 2006. Prospects for a kantian machine. IEEE Intelligent Systems 21(4):46–51.

Winfield, A. F.; Blum, C.; and Liu, W. 2014. Towards an ethical robot: internal models, consequences and ethical action selection. In Mistry, M.; Leonardis, A.; M.Witkowski; , C., eds., Advances in Autonomous Robotics Systems. Springer. 85–96.