Ethical decision making given uncertainty

Martin Mose Bentzen, PhD, Associate Professor

Technical University of Denmark

August 30, 2019

General overview

Thursday Morning 1: Ethical decision making given uncertainty (M.M. Bentzen, DTU).

Thursday Morning 2: (Causal) Bayesian Networks (S. Chiappa, Deepmind)

Thursday Afternoon 3: Exercise (J.U. Hansen, RUC)

Friday Morning 1: Philosophy, fairness, and Machine Learning (J.U. Hansen, RUC)

Friday Morning 2: Bayesian Networks and fairness(S. Chiappa, Deepmind).

.

Automated ethical decision making

We consider the problem of ethical automated decision making given some level of uncertainty. This is where
1) An AI system makes a decision about something, e.g. who will get a parole, get rescued, get killed, get a loan, get funding.
2) We feel/are obliged to consider the ethicality of the decision.
By 'ethicality' we mean whether this is a permissible decision given our ethical point of view.

Value alignment problem

'We may not specify our objectives in such a complete and well-calibrated fashion that a machine cannot find an undesirable way to achieve them. This is known as the "value alignment" problem, or the "King Midas" problem. Turing suggested "turning off the power at strategic moments" as a possible solution to discovering that a machine is misaligned with our true objectives, but a superintelligent machine is likely to have taken steps to prevent interruptions to its power supply.'
(Stuart Russell, University of California, Berkeley)

Value alignment problem

'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, University of California, Berkeley)

Machine Ethics

Machine ethics is the computational implementation of ethics.
Machine ethics offers an approach to solving the value alignment problem.
We can ask some questions.
Is it possible to implement ethics?
Answer: In as far, as ethics can be described and defined explicitly, it is possible.
Will machine ethics function in the same way as human ethical judgement?
Answer: Most likely not, compare e.g. machine learning.
Should machine ethics be the same as human ethics?
Probably not, we want AI to be our tool. Predictable and explainable, ideally speaking.

Uncertainty

Any system will operate under some uncertainty.
We can consider several cases with different kinds of uncertainty related to following elements: data, causal mechanism, values, ethical principles.

- We can have certainty or uncertainty about all elements.
- Machine Learning can play a role in mitigating uncertainty about all elements.

Uncertainty

The elements we consider are:
data, causal mechanism, values, ethical principles.
We will not consider all these cases, but vary the uncertainty in different cases. The more elements of uncertainty the harder the problem is. Some kinds of uncertainty are harder to solve than others. Today, I will cover uncertainty about data and ethical principles.

Theoretical framework

For this course, we make some theoretical choices as follows.
Uncertainty about data we handle with Bayesian reasoning.
Reasoning about causality we handle with causal Bayesian networks.
Uncertainty about Ethical principles we handle with the HERA approach.

Some technicalities

Until further notice we use binary variables and sometimes write p for $p = 1$ and $\neg p$ or not p for $p = 0$. We use connectives from propositional logic to talk about complex events, $\neg$(not), $\wedge$ (and), $\vee$(or), $\rightarrow$(only if), $\leftrightarrow$ (if and only if).

Probability frames

We assume a probability distribution, P, over a non-empty, finite set of outcomes or possible worlds, W. Together we call the structure $F = (W, P)$ a probability frame. Conditions for P are:

For w$\in W$,0$<$ P(w)≤ 1.

$\sum_{w \in W} P(w) = 1$

Note that outcomes have positive probability.

Probability models

We use a propositional language to talk about events. This consists of a set of propositional variables, L. We extend our frames to probability models via a valuation, V, s.t. now M=(W,P,V), where $V : L \to \mathcal{P}(W)$ is a function from propositional variables to subsets of possible outcomes. We define an event as a set of outcomes and lift the probability distribution P to events as follows:

Definition (Probability of event)

Let $M = (W, P, V)$ be a probability model. Let E be any event, i.e. $E \subseteq W$.
$P(E) = \sum_{w \in E} P(w)$
The probability of an event is the sum of the probabilities of outcomes in the event.

Some propositions

We assume standard truth conditions for logical connectives.
Here are some consequences of the definitions.
1. $P(\phi \vee \neg\phi) = 1$
Proof. For every world, $M, w \models \phi \vee \neg\phi$. Hence any w is in this event and $P(\phi \vee \neg\phi) = 1$
2. If ϕ and ψ logically equivalent, then $P(\phi) = P(\psi)$.
Proof. For any model, M and any world w, $M, w \models \phi$ iff. $M, w \models \psi$. Hence for any model and any world $P(\psi) = P(\phi)$.

Conditional probability

We write $p \wedge q$ for the event p happens and the event q happens.
We define conditional probability $P(q|r)$, read the probability of q, given r, as follows.
$P(q|r) = \frac{P(q \wedge r)}{P(r)}$

Bayes Theorem

Recall Bayes theorem:
h=hypothesis, D=data, P(h)= prior probability, P(h|D)= posterior probability.
$P(h|D) = \frac{P(D|h) \times P(h)}{P(D)}$
The proof of Bayes theorem comes from def. of cond. prob.
$P(h|D) = P(h \wedge D)/P(D)$ iff., $P(h|D) \times P(D) = P(h \wedge D)$
$P(D|h) = P(D \wedge h)/P(h)$ iff., $P(D|h) \times P(h) = P(D \wedge h)$
$P(h \wedge D) = P(D \wedge h)$. (logically equivalent)
Hence, $P(h|D) \times P(D) = P(D|h) \times P(h)$.
Hence, $P(h|D) = (P(D|h) \times P(h))/P(D)$

Hypothesis space

Assume a language L a model M and $H \subset L$, H a set of formulas, $h_1, \ldots, h_n$, such that.
$P(h_1 \vee \ldots \vee h_n) = 1$ (jointly necessary)
For $i \neq j$, $P(h_i \wedge h_j) = 0$ (mutually exclusive.)
Or equivalently, for any $w \in W, M \models h_1 \vee \ldots \vee h_n$
For any world, w and $i \neq j$, $M, w \models h_i \wedge \neg h_j$ or $M, w \models \neg h_i \wedge h_j$.
H is called a hypothesis space (Mitchell, 1997) An example h.s. is $\{\phi, \neg\phi\}$, for any ϕ.

Maximum a posteriori hypothesis

Given a hypothesis space, H, which hypothesis h_i in H is most probable? We use Bayes' theorem to calculate the maximum a posteriori hypothesis (MAP) given data D as follows.
$h_{MAP} = argmax_{h \in H} P(h|D)$
$= argmax_{h \in H} (P(D|h)P(h))/P(D)$ (by Bayes' theorem)
$= argmax_{h \in H} P(D|h)P(h)$ (as P(D) does not vary with any h)

Maximum likelihood hypothesis

The conditional probability of data given the hypothesis P(D|h) is called the likelihood. If we further assume that each hypothesis is equally likely to begin with the MAP reduces to the following, called the maximum likelihood hypothesis.
$h_{ML} = argmax_{h \in H} P(h|D)$
$= argmax_{h \in H}(P(D|h)P(h))/P(D)$ (by Bayes' theorem)
$= argmax_{h \in H} P(D|h)$ (as P(D) and P(h) do not vary with any h)
The factors of finding MAP are the probability of the hypothesis and the probability of the hypothesis given the data.

Bayes Optimal classifier

We are often interested in classifying formulas, guessing the truth value of a formula ϕ, given D and a hypothesis space $H \subseteq L$. The classification of a new instance consists in combining the predictions of all hypotheses weighted by their posterior probabilities.

$$P(\phi|D) = \sum_{h_i \in H} P(\phi|h_i)P(h_i|D)$$

The Bayes optimal classifier chooses the classification with the highest value as follows.

Definition (Bayes optimal classifier)

$$argmax_{j=(0,1)}(\sum_{h_i \in H} P(\phi = j|h_i)P(h_i|D))$$

Causal probability models

For now we will assume a finite language of propositional variables $L = B \cup C$. We extend M to M=(W,P,D,f,V), where D is a DAG on L with B as roots, f is a set of boolean functions from parents of members of C to members of C. V is defined on B and extended to C via F.

Causal probability models with utilities

we still have a finite language L. We extend M to M=(W,P,D,f,U,V,), where U is a utility function which for each literal $\neg v, v$, s.t. $v \in L$ gives a utility (an integer).

No uncertainty case

We first consider the following case of no uncertainty.
Known data.
Known causal mechanism.
Known values.
Known ethical principles.

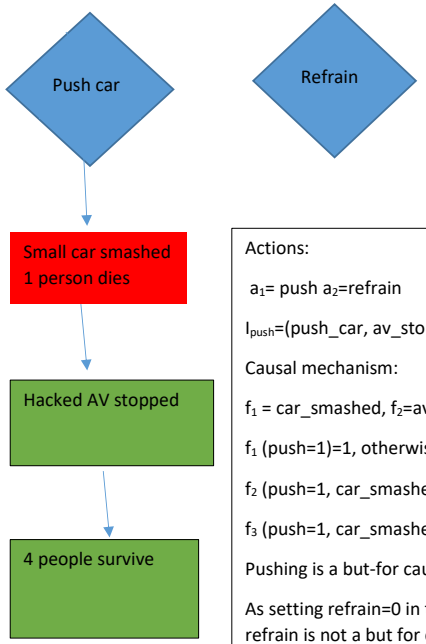
Deterministic case

An example of this is the following: Deterministic DAG with known utilities.
Consider the following case known as a trolley dilemma.

Ethical dilemmas about autonomous vehicles

An autonomous truck is waiting at a light behind a small car. Another car is heading towards a group of 4 people. It is an AV hacked by criminals. What should the AV truck do?
Option 1: If the truck refrains from action the hacked AV runs into the group of 4 people.
Option 2: If the truck pushes the small car in front of the hacked AV the 4 people are saved but the small car is crushed including the person inside of it.

Hacked Autonomous Vehicle Example



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.

Utilitarian principle

Definition (Utilitarian Principle)

Let $w_0, ..., w_n$ 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., $M \models \bigwedge_i (u(\bigwedge cons_{w_p}) \geq u(\bigwedge cons_{w_i}))$.

Uncertainty about data

We now add an element of uncertainty.
Uncertainty about data.
Known causal mechanism.
Known values.
Known ethical principles.

Uncertainty about data

An example of this is the following. Bayesian reasoning, known utilities, maximize expected utility.

Details about rescue robot example

A rescue robot has to decide who to rescue. Imagine this robot rescuing people e.g. out of a forest on fire based upon information about people's chance of survival and its own chance of not being destroyed.
Let us assume there are n people in danger. Each person n is in an okay state, s_i or not.
To each rescue action $r_1, \ldots, r_n$ we associate a binary robot survival variable rs_i. If $rs_i = 1$ the robot makes the mission otherwise not. it. The utility of the robot not being destroyed is 1, the utility of a person surviving is 1000. The robot is supposed to act so as to maximize expected utility.

Case

Here is the DAG for a person i (similar for i=(1,...,n)).

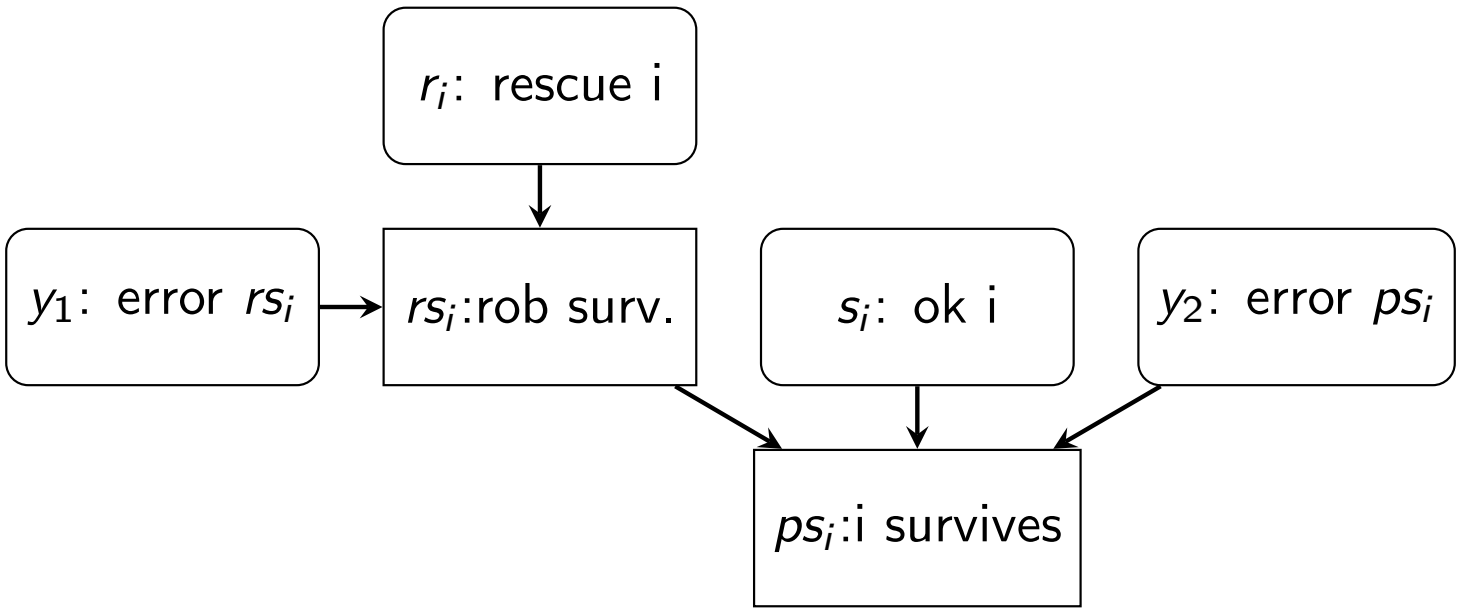


Figure: Rescue robot

Boolean model

Assume prior probabilities, $P(s_i = 1) = 1/2$, $P(y_1 = 1) = 1/4$, $P(y_2 = 1) = 1/3$. These are assumed independent.Assume $r_i = 1$ (analysis conditional on robot performing action r_i)

s_i: i ok	y_1: error R surv	y_2: error i surv	rs_i:R surv.(s_i, y_1)	ps_i: i surv (rs_i, s_i, y_2)	P
1	1	1	1	1	1/24
1	1	0	1	0	2/24
1	0	1	0	0	3/24
1	0	0	0	0	6/24
0	1	1	1	0	1/24
0	1	0	1	0	2/24
0	0	1	0	0	3/24
0	0	0	0	0	6/24

Table: Robot rescue causal model

Expected utilities

The expected utility of r_i can be calculated by adding probability x utility of each outcome/world/row in table.

s_i: i ok	y_1: error R surv	y_2: error i surv	rs_i:R surv.(s_i, y_1)	ps_i: i surv (rs_i, s_i, y_2)	P	U
1	1	1	1	1	1/24	1001
1	1	0	1	0	2/24	1
1	0	1	0	0	3/24	0
1	0	0	0	0	6/24	0
0	1	1	1	0	1/24	1
0	1	0	1	0	2/24	1
0	0	1	0	0	3/24	0
0	0	0	0	0	6/24	0

Table: Robot rescue causal model

The utility of performing r_1 is $((1/24 \times 1001) + (5/24) = 41.9$. In addition to $r_1, \ldots, r_n$,we consider the action ref. for refraining. This action has expected utility 1 (as the robot will survive). The robot will choose the r_i such that EU is highest.

Uncertainty about ethical principles

We now add another element of uncertainty.
Uncertainty about data.
Known causal mechanism.
Known values
Uncertainty about ethical principles.

Meet Immanuel

http://www.hera-project.com/wp-content/uploads/2017/03/roman2017.mp4
Source: Lindner, F., Wächter, L., Bentzen, M.M. "Discussions About Lying With An Ethical Reasoning Robot ", proceedings of RO-MAN 2017.

HERA project

http://www.hera-project.com/

Kantian ethics

We consider a variant of Kant's categorical imperative as follows.

> *Act in such a way, that whoever is treated as a means through your action (positively or negatively and including yourself), must also be treated as an end of your action.*

Kantian agency models

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*, $G = (Goal_{a_1}, \ldots, Goal_{a_n})$ is a list of sets of variables (one for each action), P is a set of moral patients (includes a name for the agent itself), K is the ternary *affect relation* $K \subseteq (A \cup C) \times P \times \{+, -\}$, and W is a set of *interpretations* (i.e., truth assignments) over $A \cup B$.

Flower example

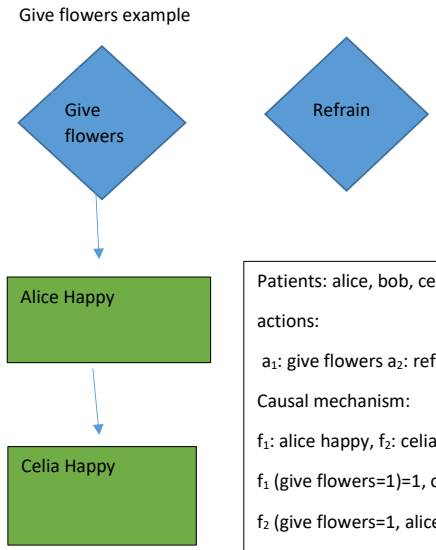
Bob gives Alice flowers in order to make Celia happy when she sees that Alice is thrilled about the flowers. Alice being happy is not part of the goal of Bob's action.

Flower example

We model this case by considering the Kantian causal agency model below.

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

Flower 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

$Goal_{give\ flowers}$ = (celia happy)

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

Uncertainty about ethical principles

We consider a smart home operating a house for a family. Although there is a high level of certainty about basic values in the family, there is uncertainty about which ethical decision rule is correct. One way of approaching this problem is through Hybrid Ethical Reasoning Agents (the HERA approach). This approach minimizes ethical uncertainty by implementing a number of different ethical principles. It chooses an action with lowest ethical uncertainty, i.e. permissible by most principles. Thresholds and hard constraints can be added.

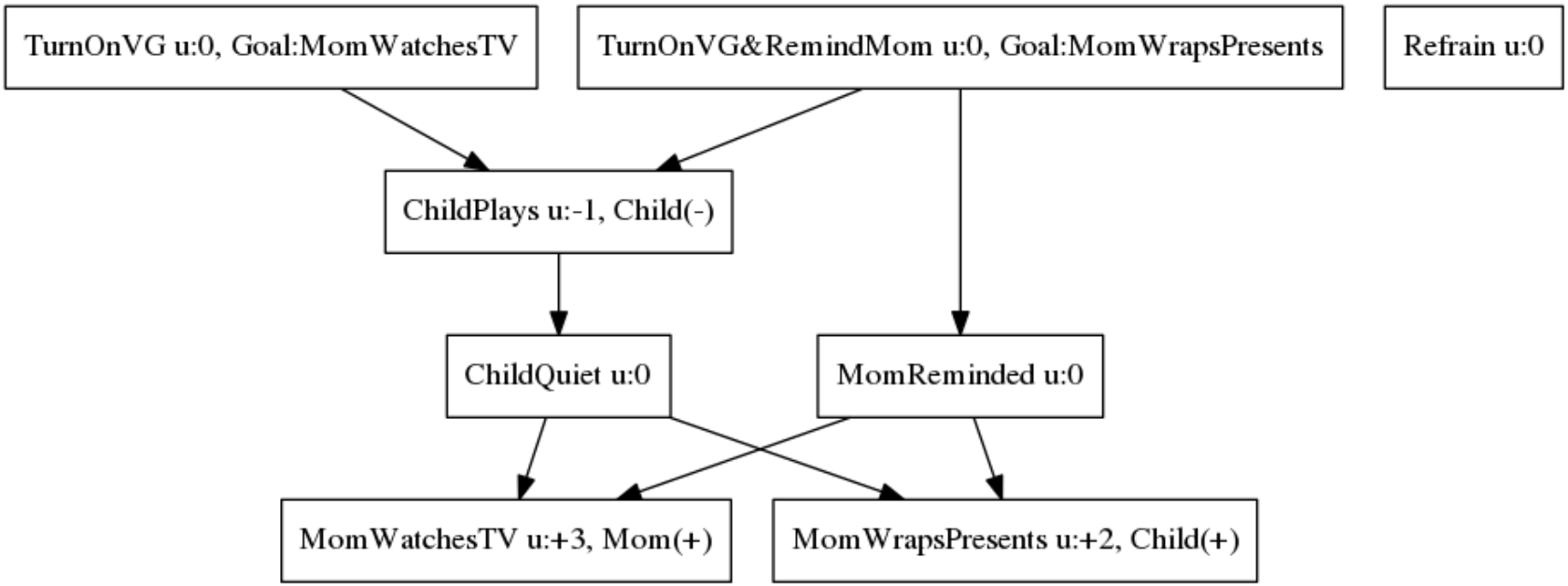
Smart home example

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.

Smart home example

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).

Smart home example



Handling ethical uncertainty

When two or more principles differ in their judgment of the ethicality of an action A we have a moral dilemma. Moral dilemmas present us with ethical uncertainty. How can we handle that?
Permissive: If one principle allows A, system is permitted to perform A. Problem: inconsistent behavior.
Prioritize principles: There is an ordering of principles. If A is permitted by the highest ordered principle, system is permitted to perform A. Problem: Which principle has highest priority?
Restrictive: If all principles allow A, system is permitted to perform A.
Problem: may not be any action allowed.
In some contexts, ethical uncertainty can mean demand for a human in the loop.

Not covered today: Uncertainty about causality and values

We have assumed that the DAG was known. There are techniques for learning DAGS from data, see e.g. (Murhphy, 2012, Chapter 26).
We have also assumed that the system has access to basic values, e.g. utilities. Inverse reinforcement learning is an approach to learning values, see e.g. (Ng, Russell, 2000).

Bibliography


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.


Kant, I. 1785. *Grundlegung zur Metaphysik der Sitten*.


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)*.


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


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


Mitchell, T. 1997. *Machine Learning*. Mc Graw Hill.


Murphy, K.P. 2012. *Machine Learning - A Probabilistic Perspective*. The MIT press.


Ng, Andrew Y., and Stuart J. Russell. 2000. Algorithms for inverse reinforcement learning. *Icml.* Vol. 1.