Engineering Machine Ethics

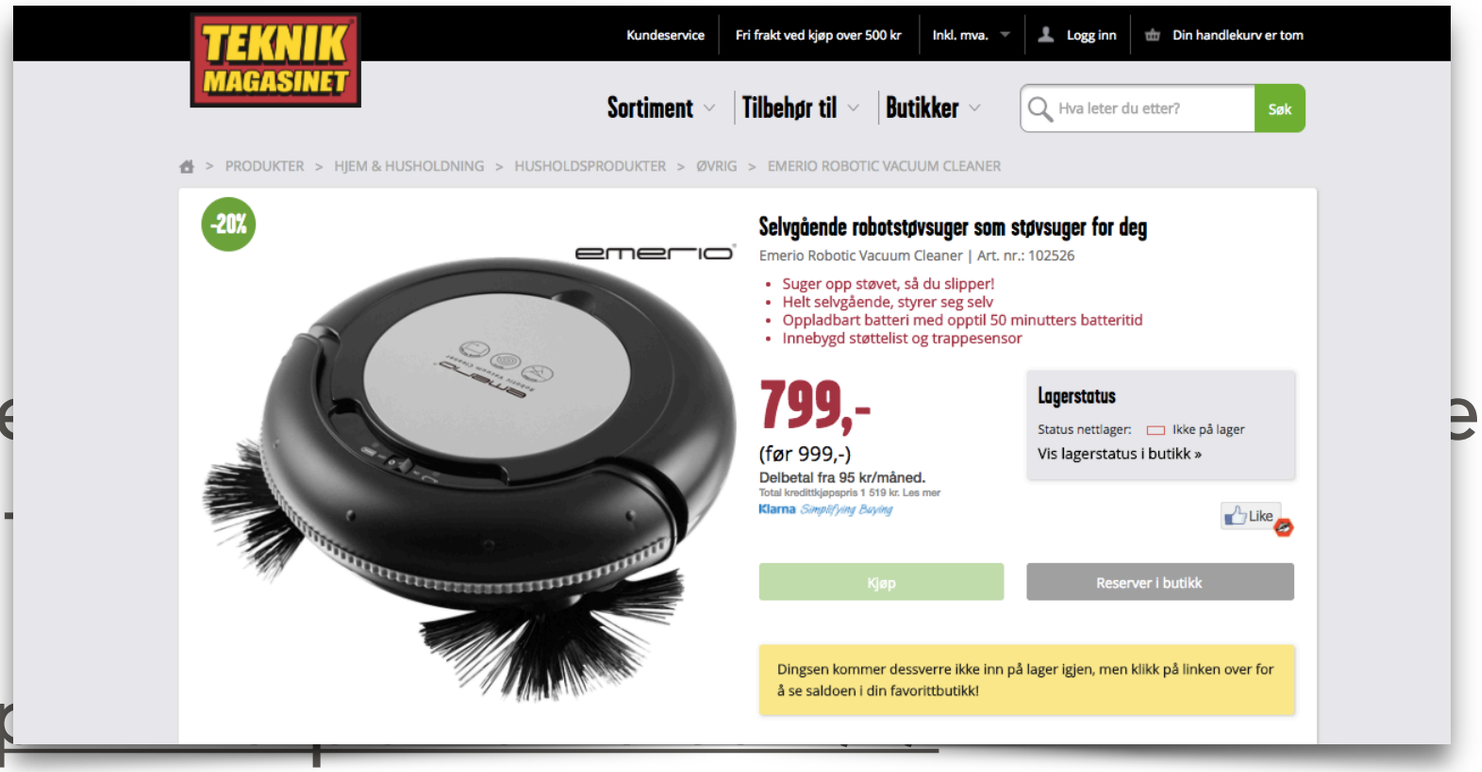
EASSS 2017, Gdansk
Marija Slavkovik
University of Bergen, Norway

Machine ethics

- What is it?
- Machine ethics vs ethical computing
- Machine ethics vs human ethics
- Engineering machine ethics
- General vs specific ME
- Strong vs weak ME
- Killer machines vs anarchist machines

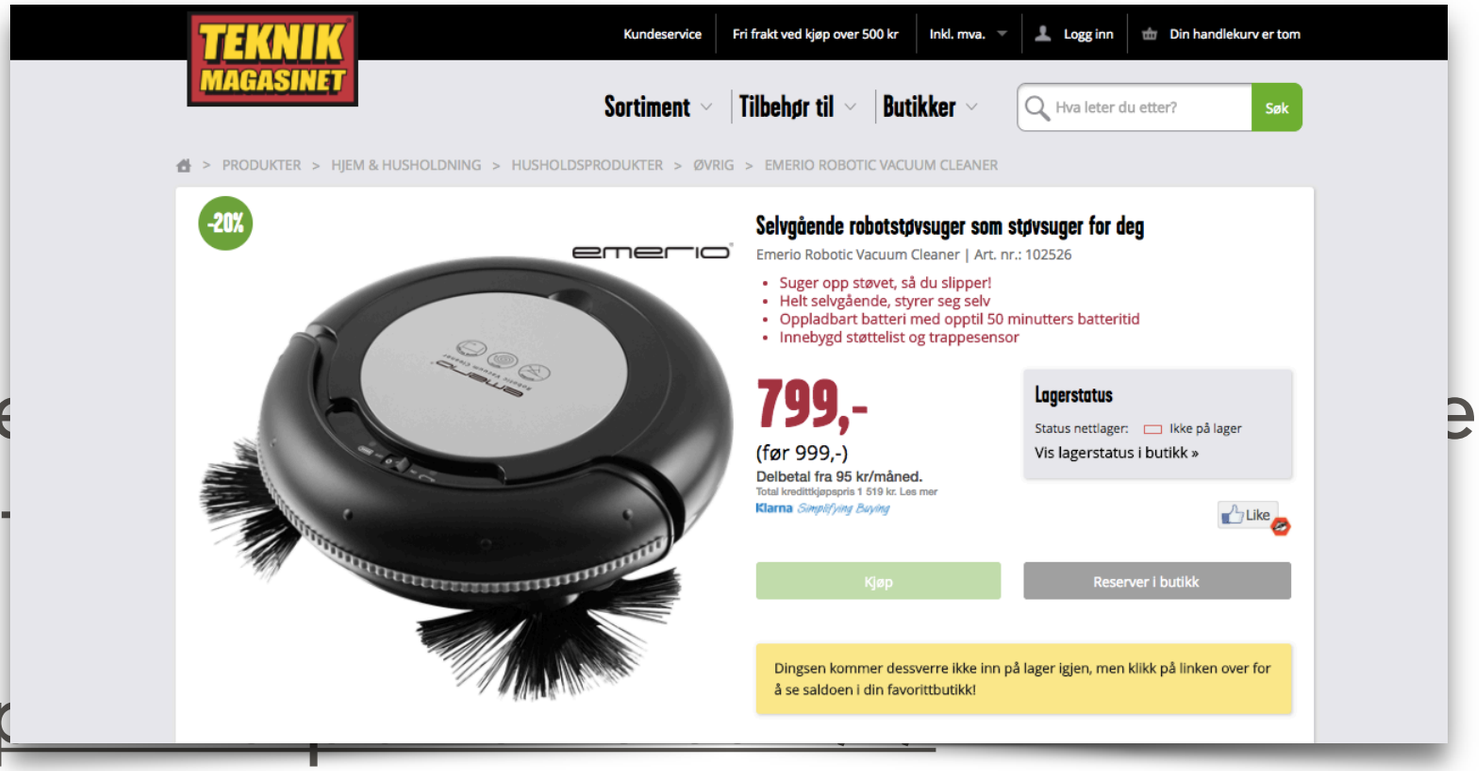
TERMINOLOGY

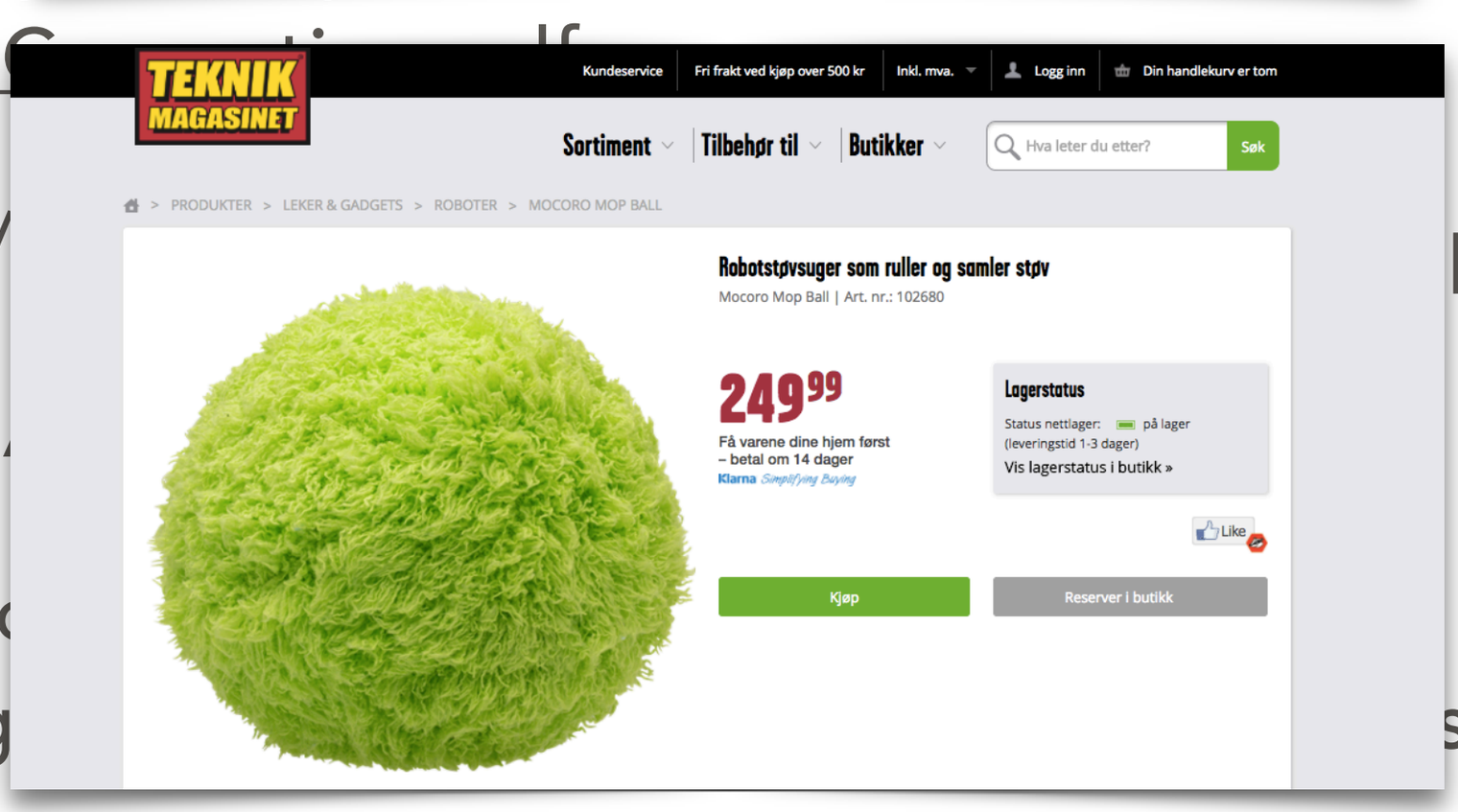
- being like a human vs acting like a human (the Turing test and the Chinese Room argument)
  - http://phil415.pbworks.com/f/TuringComputing.pdf
  - http://cogprints.org/7150/1/10.1.1.83.5248.pdf
- general AI vs specific AI
- AI methods - soft computing vs symbolic AI
  Thinking computers and swimming submarines

- being like 

e Turing test and t
  - http://p
    - http://cogprints.org/7150/1/10.1.1.83.5248.pdf
- general AI vs specific AI
- AI methods - soft computing vs symbolic AI
  Thinking computers and swimming submarines

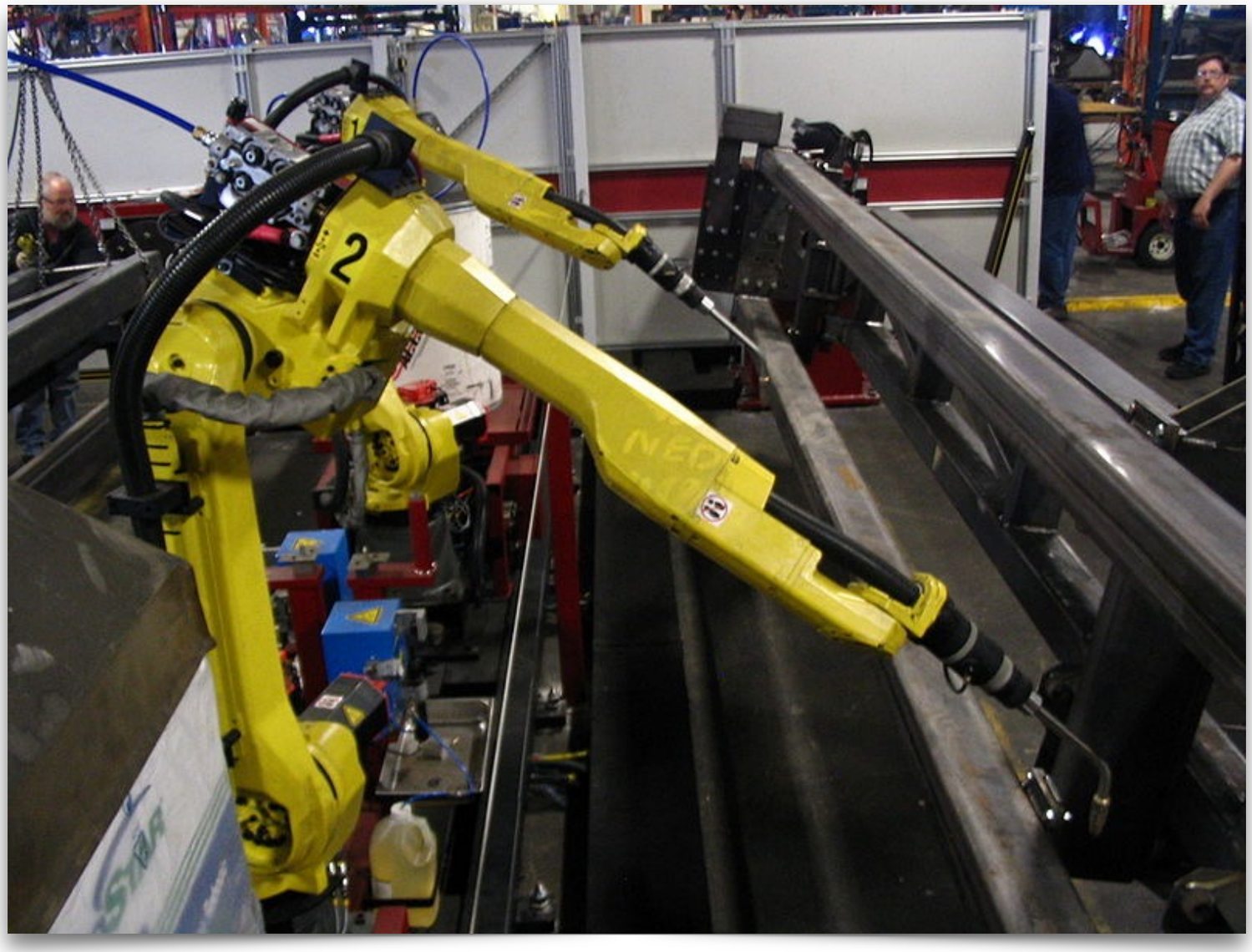
- being like … e Turing test and …
  - http://p… Turing…
  - http://… pdf
- general …
- AI metho… Thinking… s





Why now?

Why now?



Robot kills a man at Volkswagen plant

Published time: 2 Jul, 2015 04:53
Edited time: 3 Jul, 2015 07:43

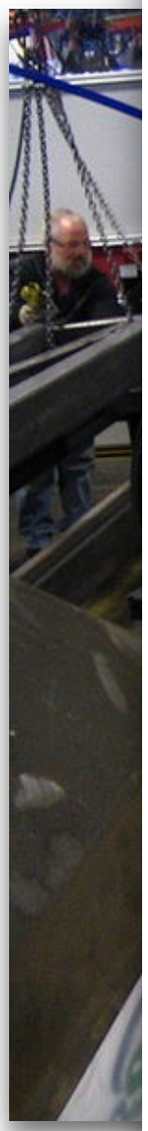
Get short URL

Reuters / Fabian Bimmer / Reuters

1

Robot kills a man at Volkswagen plant

Published time: 2 Jul
Edited time: 3 Jul, 20



Reuters / Fabian E

1

Like 2.9m Follow @MailOnline DailyMail

Friday, Oct 9t

Mail Online

Home News U.S. | Sport | TV&Showbiz | Australia | Femail | Health | Scie
| Fashion Finder

Latest Headlines | News | Arts | Headlines | Pictures | Most read | News Board | Wires

Woman is attacked as she sleeps... by her ROBOT vacuum cleaner! South Korean owner had to be cut free after it began sucking up her hair

- The woman woke up when the robot vacuum latched on to her hair
- Emergency services were called and paramedics freed her from the device
- U.S. firm iRobot has sold more than 10 million of their units since 2002

By STEVE HOPKINS FOR MAILONLINE

By STEVE HOPKINS FOR MAILONLINE

This course - part I

- Levels of autonomy
- Moors ethical agents
- Introduction to moral philosophy
  - consequentialist vs non-consequentialist theory
  - utilitarianism
  - Kantian ethics
  - prima fascia obligations

Can an artificial agent be a moral agent?

- Can it ever be autonomous?
- Can it ever learn?
- Incorporating Ethics into Artificial Intelligence Amitai Etzioni and Oren Etzioni[2]
- Can we make sure that it behaves within desirable ethical-legal specification?
- Some easy definitions of difficult concepts

Learning

- Learning is improvement of performance over time
- Learning as the problem of constructing a function that predicts the output given a collection of input-output pairs
- Forms of learning: unsupervised learning, reinforcement learning, supervised learning, semi-supervised learning
- Data mining - discovering properties of data sets
- Machine learning is one of the ways in which data mining can be accomplished, but not the only thing it is used for

To be autonomous

- **controlled systems**: where humans have full or partial control, such as an ordinary car

- **supervised systems**: which do what an operator has instructed, such as a programmed lathe or other industrial machinery

- **automatic systems**: that carry out fixed functions without the intervention of an operator, such as an elevator

- **autonomous systems**: that are adaptive, learn and can make ‘decisions’, like Curiosity

Moor’s classes of ethical agents

Moor's classes of ethical agents

- Ethical-impact agents

Moor's classes of ethical agents

- Ethical-impact agents

Moor's classes of ethical agents

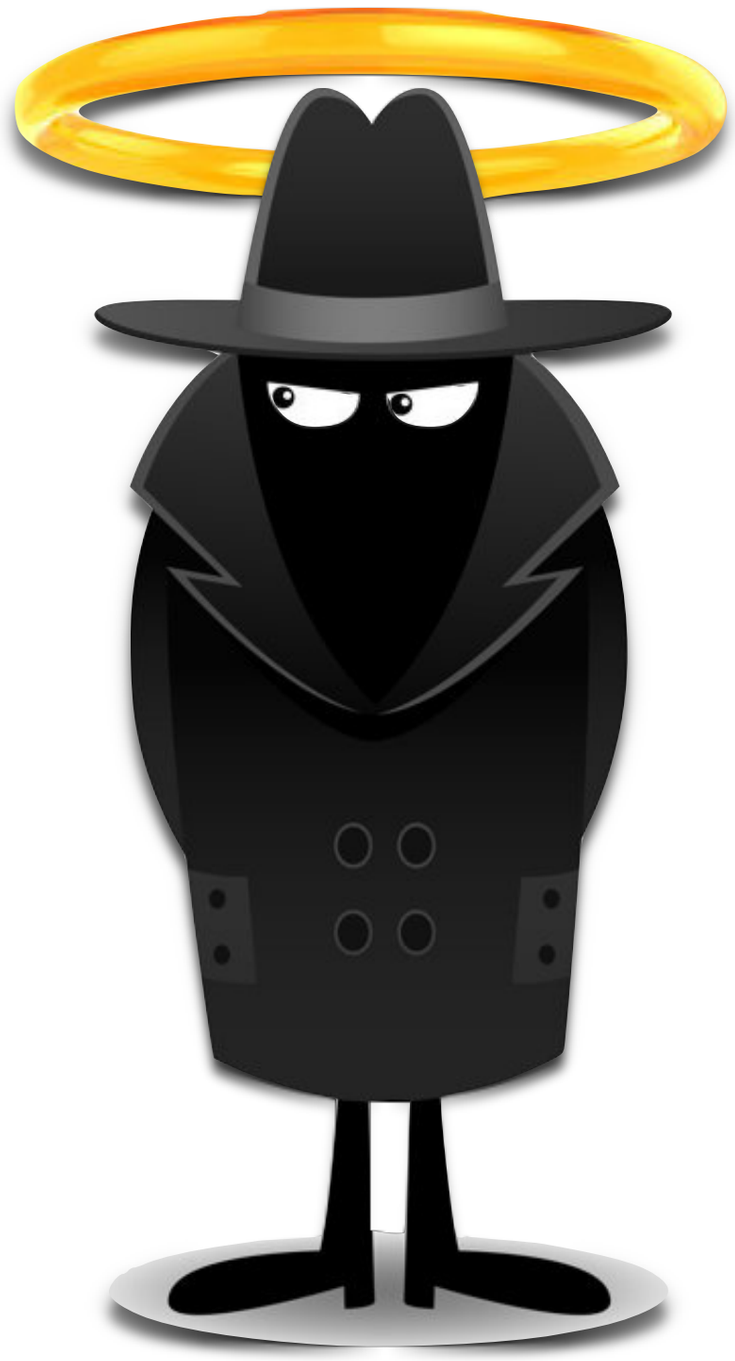
- Ethical-impact agents

Moor's classes of ethical agents

- Ethical-impact agents
- Explicit ethical agents

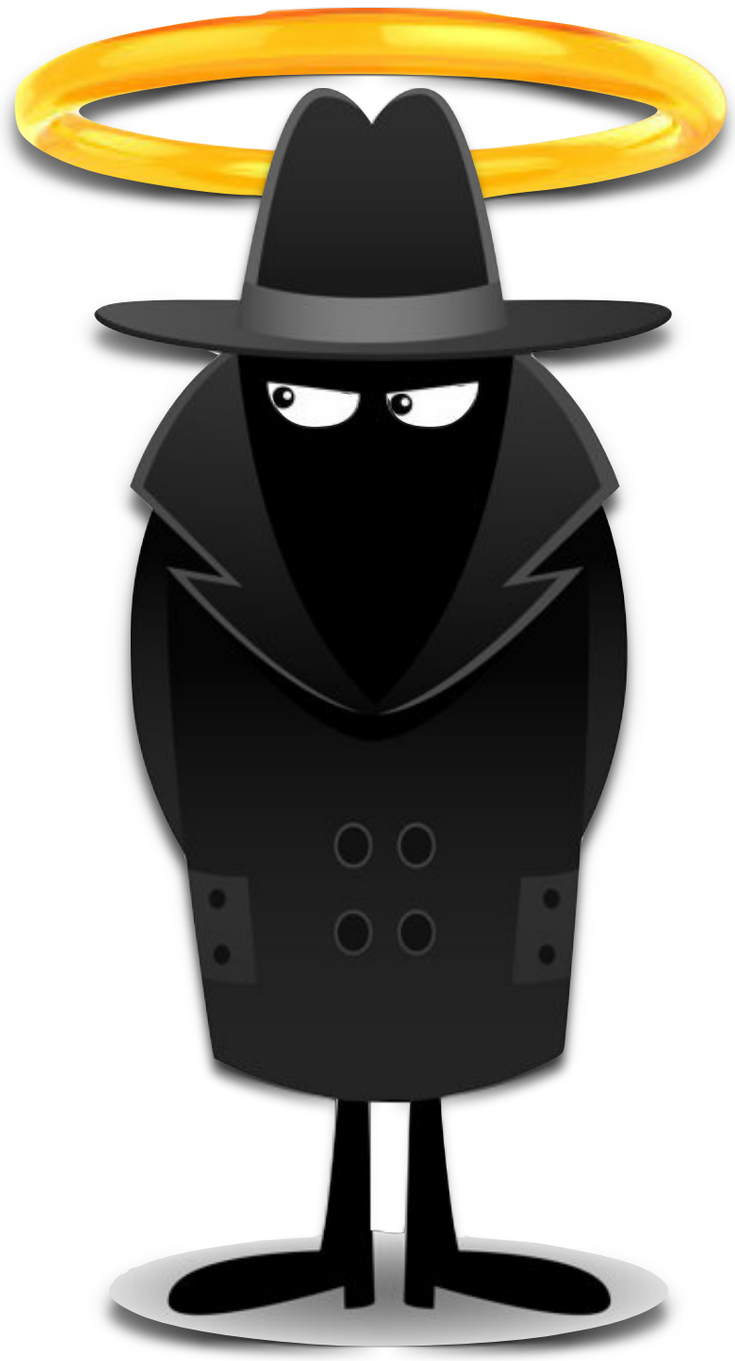
Moor's classes of ethical agents

- Ethical-impact agents
- Explicit ethical agents



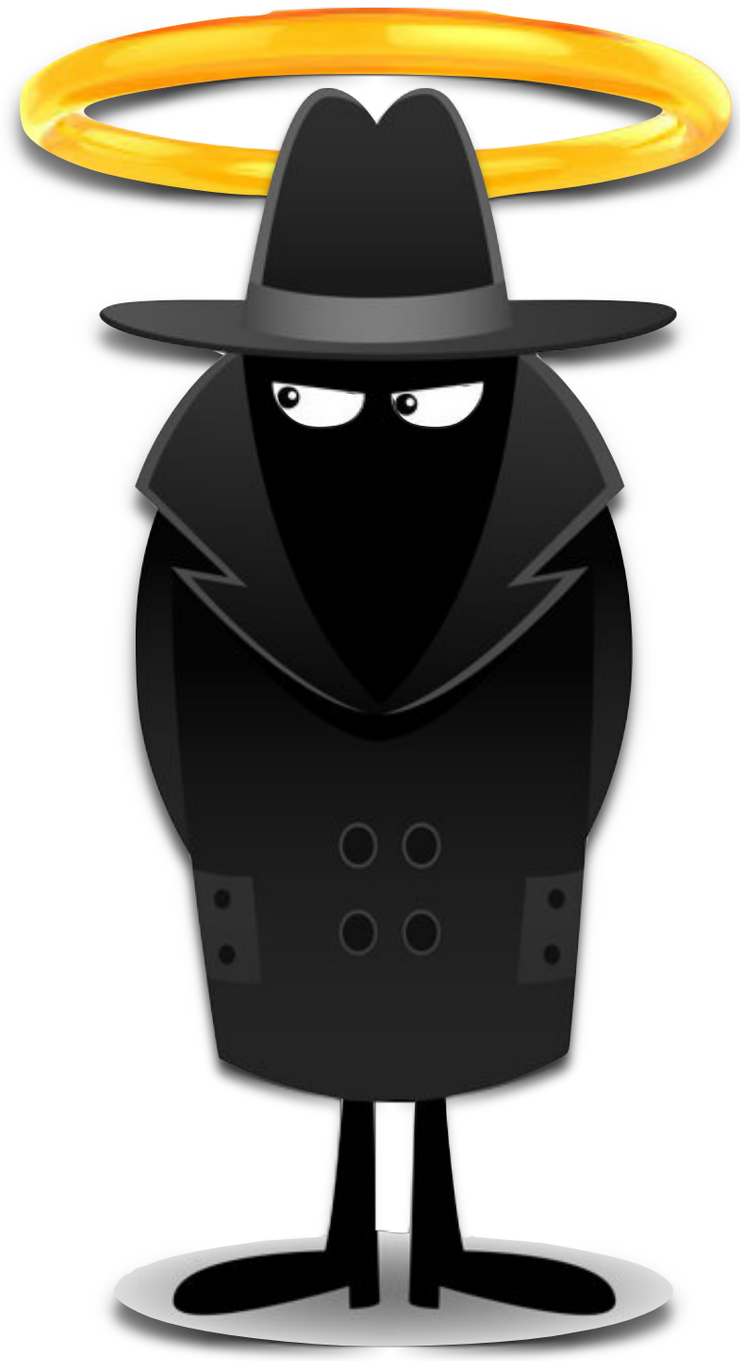
Moor's classes of ethical agents

- Ethical-impact agents
- Explicit ethical agents
- Implicit ethical agents



Moor's classes of ethical agents

- Ethical-impact agents
- Explicit ethical agents
- Implicit ethical agents
- Full ethical agents



Moor's classes of ethical agents

- Ethical-impact agents
- Explicit ethical agents
- Implicit ethical agents
- Full ethical agents

Introduction to moral philosophy

- Philosophy - systematic use of critical reasoning to answer the most fundamental questions in life

- Moral philosophy - the question is “What is good/bad?”

- Morality vs Ethics

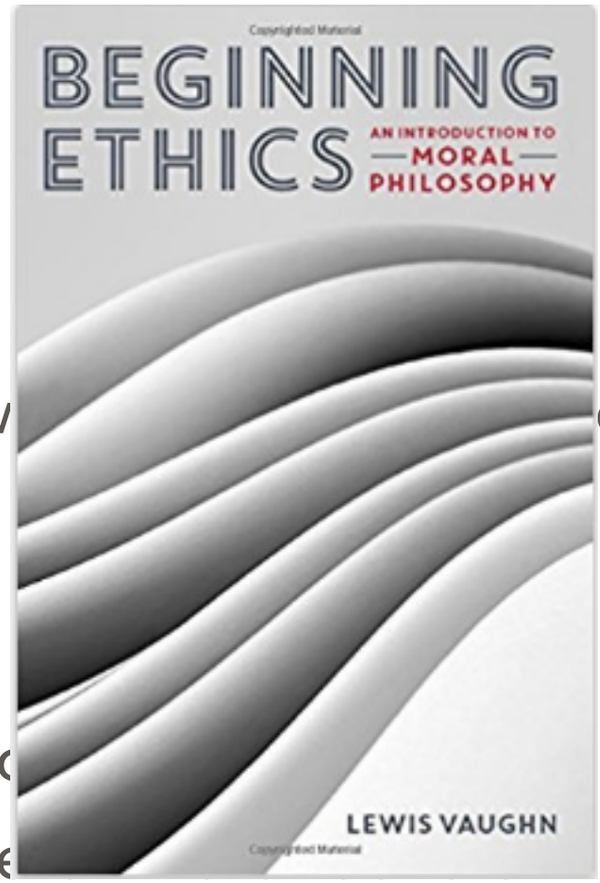
- Descriptive ethics - describe and explain how people behave and think when dealing with moral issues

- Major divisions in ethics
  - Normative ethics - principles, rules or theories that guide us
  - Metaethics - meaning and logical structure of moral beliefs
  - Applied ethics - applying moral norms to specific moral issues or cases, values and obligations

Introduction to moral philosophy

- Philosophy - systematic use of critical reasoning to answer the most fundamental questions in life

- Moral philosophy - the question is "What is



- Morality vs Ethics

- Descriptive ethics - describe and explain hov d think when dealing with moral issues

- Major divisions in ethics
  - Normative ethics - principles, rules or theo
  - Metaethics - meaning and logical structure
  - Applied ethics - applying moral norms to specific moral issues or cases, values and obligations

Elements of ethics

- The backbone of moral reasoning is the logic argument
- Universalizability - the idea that a moral statement that applies in one situation must apply in all other situations that are relevantly similar
- Impartiality - from the moral point of view, all persons are considered equal and should be treated equally
- Not all norms are moral norms

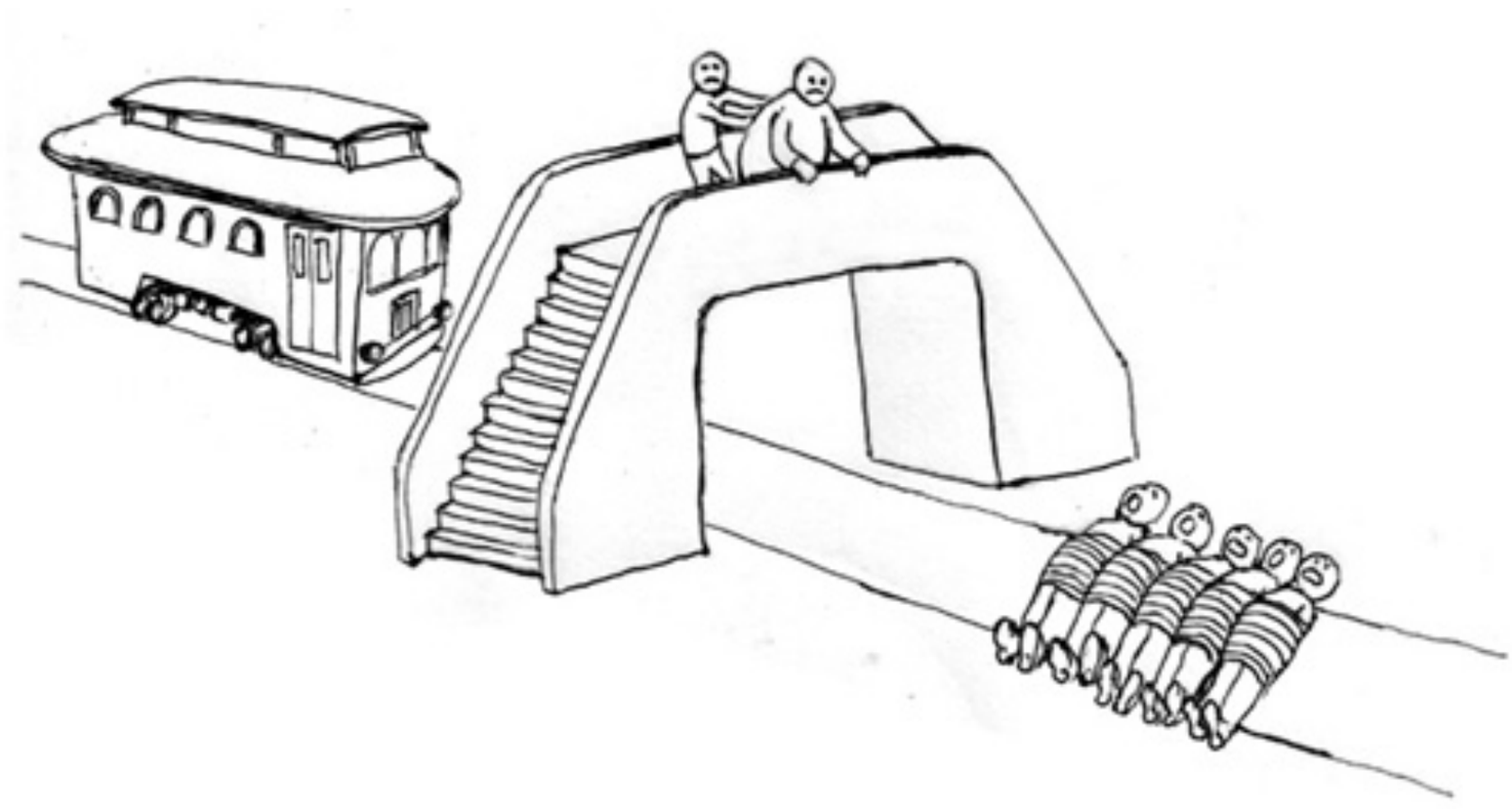
Moral theories

- A moral theory is an explanation of what makes an action right/wrong and what makes a person good/bad
- Theories of values - concerned with the goodness of persons or things
- Theories of obligation - concerned with the rightness or wrongness of actions; what makes an action right or wrong
- Consequentionalist theories - all is well that ends well
- Non-consequentionalist (deontologist) theories - not only consequences but the nature of the action is what matters

Evaluating ethical theories

- Coherence
- Criterion 1: Consistency with Considered Judgments
- Criterion 2: Consistency with our Moral Experience
  - We sometimes make moral judgments
  - We often give reasons for particular moral beliefs
  - We are sometimes mistaken in our moral beliefs
  - We occasionally have moral disagreements
  - We occasionally commit wrongful acts
- Criterion 3: Usefulness in Moral Problem Solving

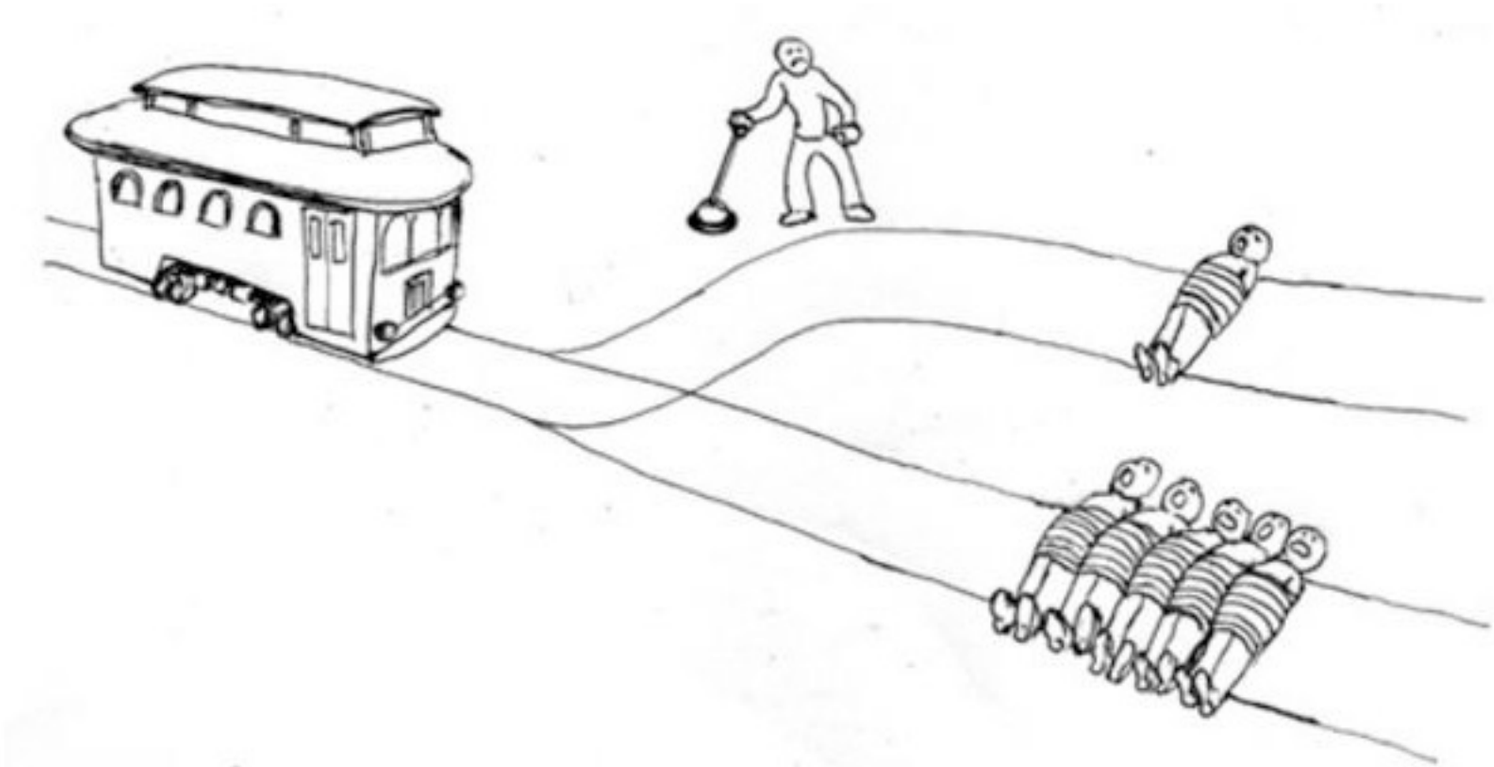
Utilitarianism



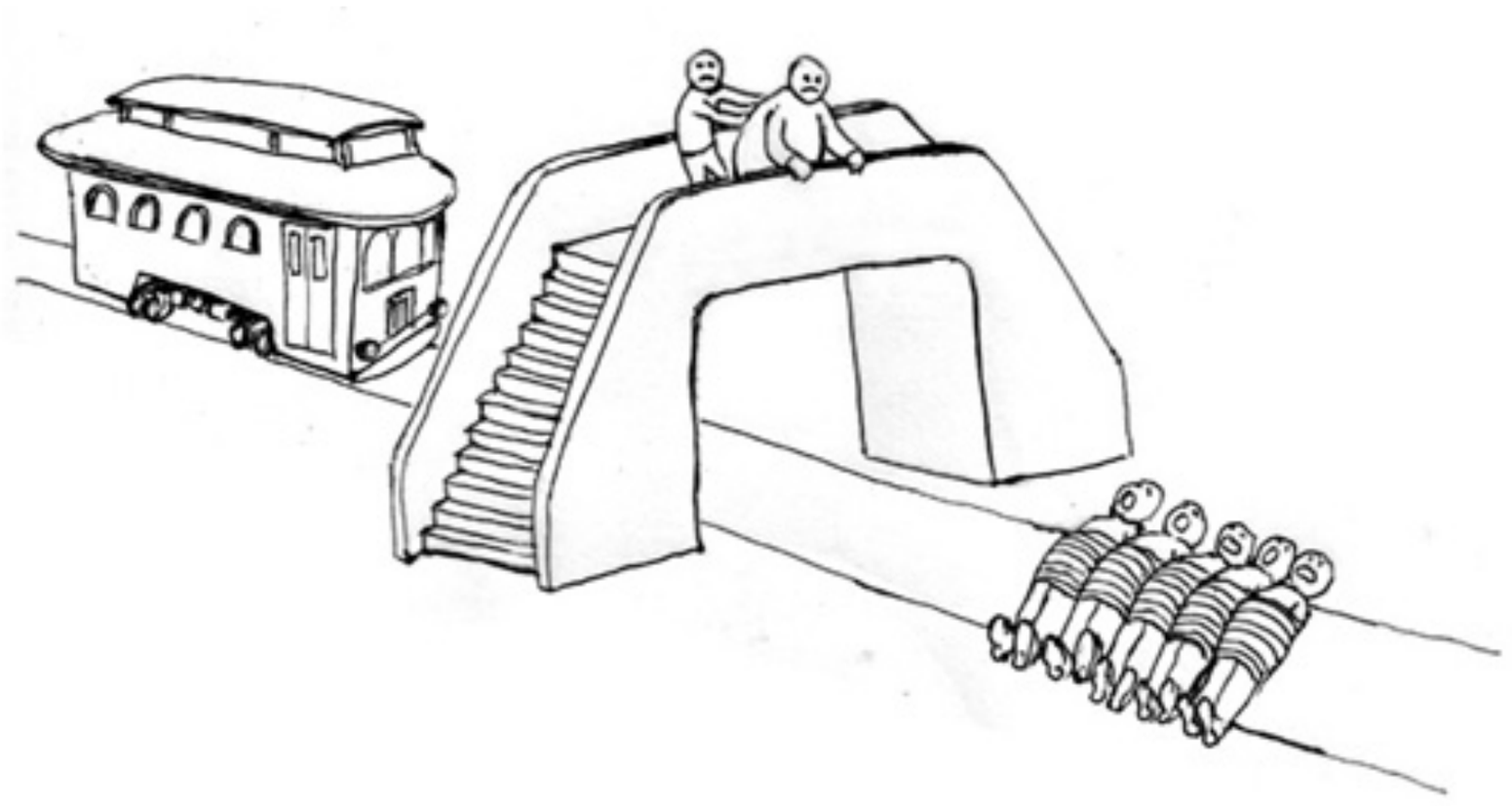
Utilitarianism

- The right actions are the one that increase the utility in society
- Jeremy Bentham (1748-1832) and John Stuart Mill (1806-1873)
- Act-utilitarianism: morally right actions are those that directly produce the greatest overall good, everyone considered
- Rule-utilitarianism: morally right action is the one covered by a rule that if generally followed would produce the most favourable balance between good and evil, everyone considered (rules must be followed constantly even if they are locally not good)

Utilitarianism



Utilitarianism



Utilitarianism - what is wrong with it

- Whose utility should you maximise?
- How will you define utility?
- How can you be sure that you have taken into account everything that matters?
- Direct consequences or a closure?
- How much can we be certain in the consequences of actions in an uncertain world?

Utilitarianism - what is wrong with it

- Whose utility should you maximise?
- How will you define utility?
- How can you be sure that you have taken into account everything that matters?
- Direct consequences or a closure?
- How much can we be certain in the consequences of actions in an uncertain world?

Utilitarianism - what is wrong with it

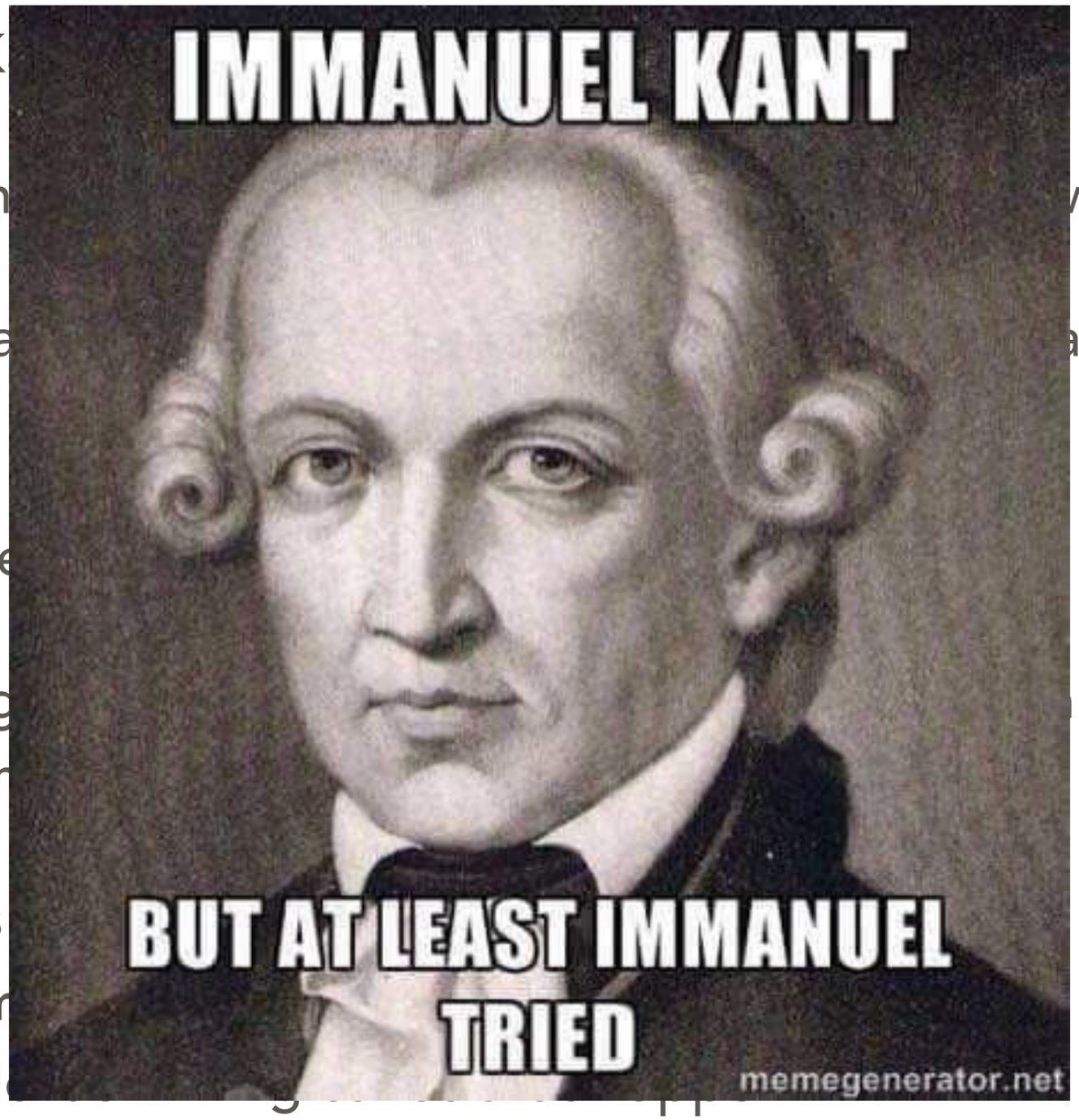
- Whose utility should you maximise?
- How will you define utility?
- How can you be sure that you have taken into account everything that matters?
- Direct consequences or a closure?
- How much can we be certain in the consequences of actions in an uncertain world?

Kantian Ethics

- Immanuel Kant (1724-1804): reason alone leads us to the right and to the good.
- Right actions have moral value only if they are done with “good will”
- Hypothetical imperative - what we should do if we have certain desires
- Categorical imperative - a common we should follow regardless of our wants or needs, universal and unconditional
- Kant’s categorical imperative: act only on that maxim through which you can at the same time will that it becomes a universal law.
- An action is permissible if:
  - its **maxim** can be universalised
  - you would be willing to let that happen
- Perfect duties vs imperfect duties

Kantian Ethics

- Immanuel K… the right and to the good.
- Right action… with "good will"
- Hypothetica… ave certain desires
- Categorical… regardless of our wants or needs, unive…
- Kant's categ… through which you can at the same tim…
- An action is…
  - its **maxim**…
  - you woul…
- Perfect duties vs imperfect duties



Prima-facie principles

- All theories so far struggle with absolutism
- Duties instead of absolute principles
- How do duties relate to each other?
- Prima-facie principles - principles that apply unless an exception is given (W. D. Ross first to consider them)
- 7 Ross p.f.p.: fidelity, reparation, gratitude, justice, beneficence, self-improvement, non-maleficence
- Recently: autonomy, justice, beneficence, non-maleficence
- qualification problem and ramification problem

What when prima-facie obligations conflict?

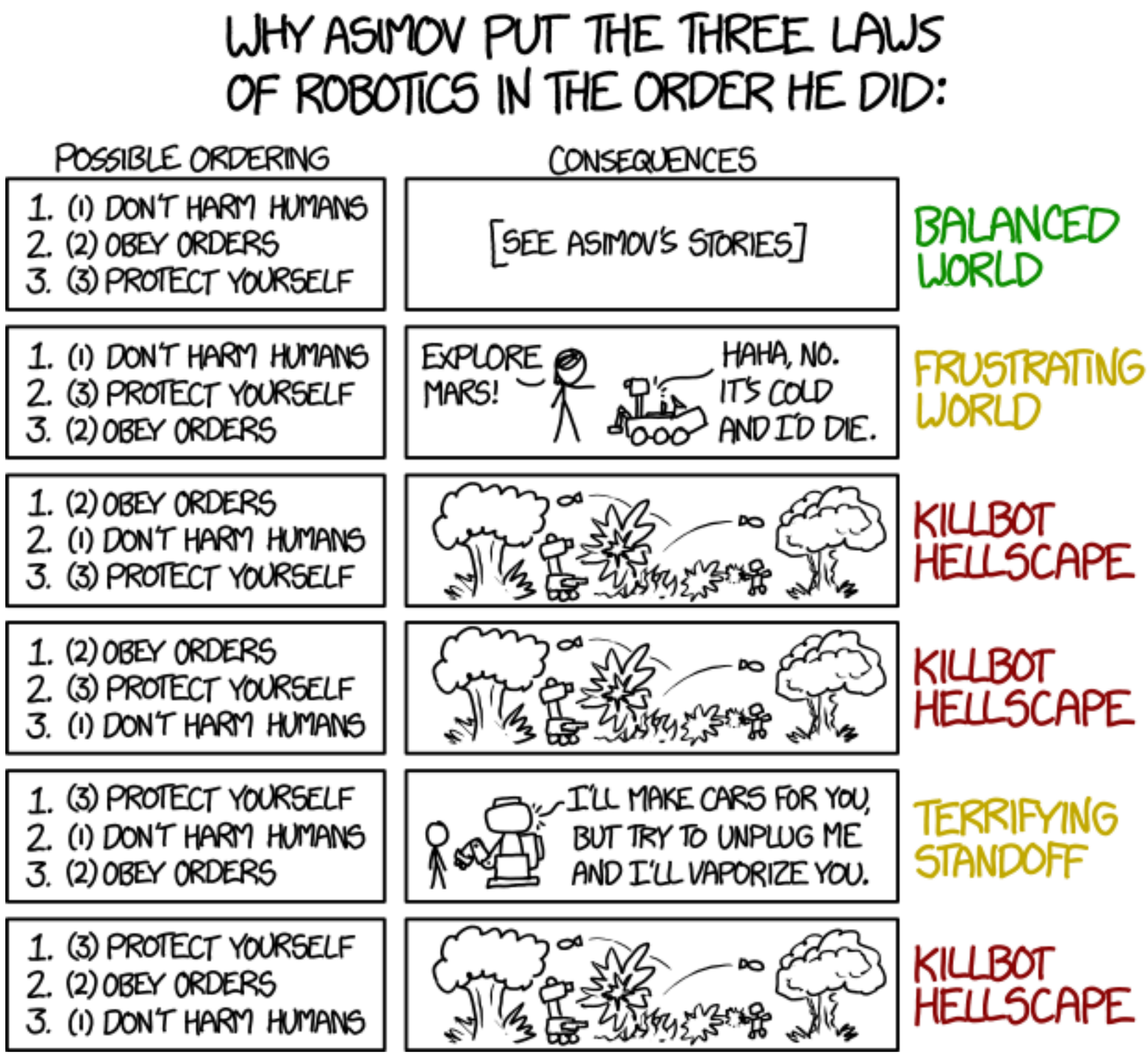
- Deontic logic reasoning about what you ought to do
- Axiomatisation of Standard Deontic Logic

- A1. All tautologous wffs of the language (TAUT)
- A2. $O(p \rightarrow q) \rightarrow (Op \rightarrow Oq)$ (OB-K)
- A3. $Op \rightarrow \neg O \neg p$ (OB-D)
- R1. If $\vdash p$ and $\vdash p \rightarrow q$ then $\vdash q$ (MP)
- R2. If $\vdash p$ then $\vdash Op$ (OB-NEC)

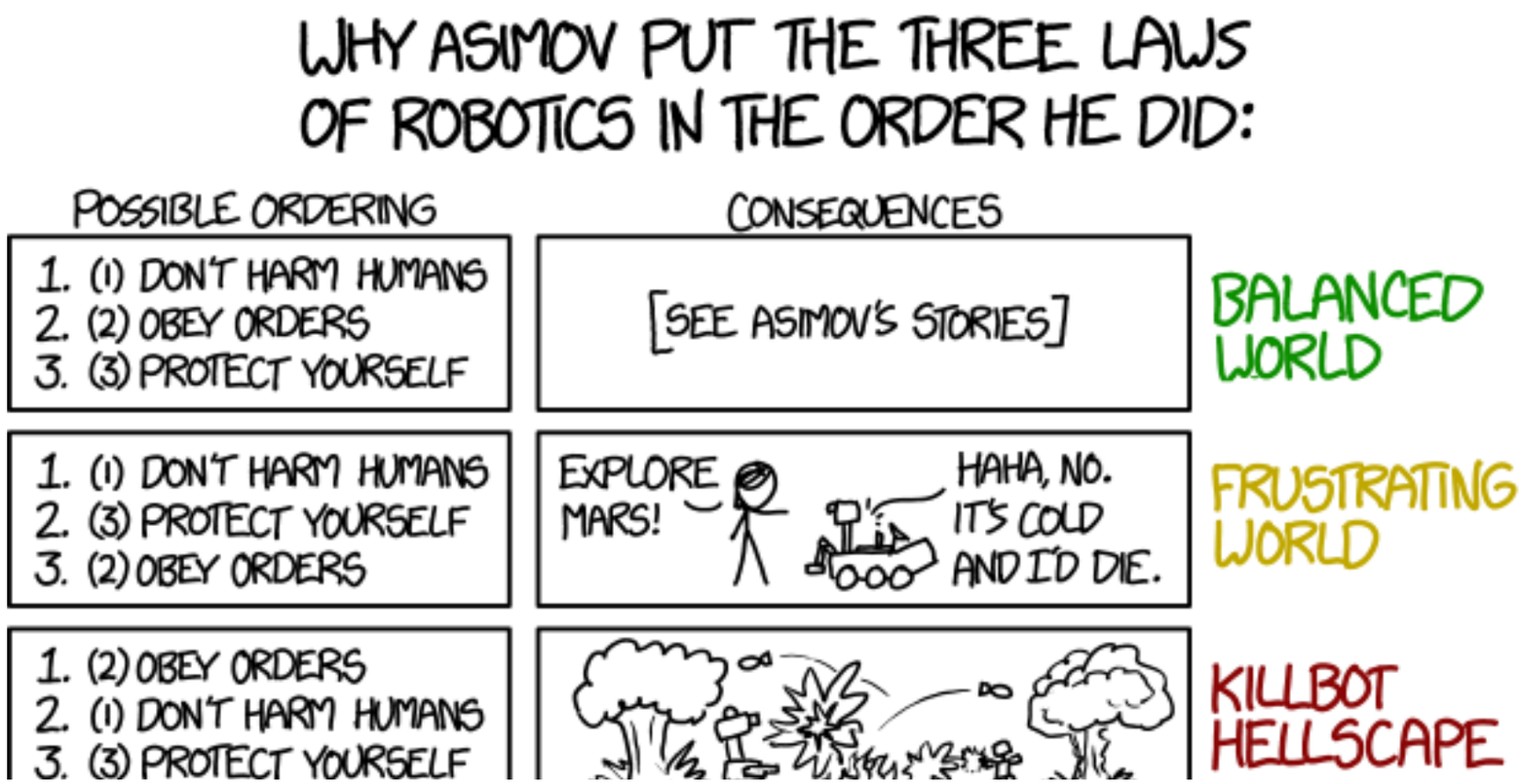
Contrary-to-duty

- Contrary-to-Duty (or Chisholm's) Paradox:
  - (1) It ought to be that Jones goes (to the assistance of his neighbors).
  - (2) It ought to be that if Jones goes, then he tells them he is coming.
  - (3) If Jones doesn't go, then he ought not tell them he is coming.
  - (4) Jones doesn't go.

A test


WHY ASIMOV PUT THE THREE LAWS OF ROBOTICS IN THE ORDER HE DID:
POSSIBLE ORDERING
CONSEQUENCES
1. (1) DON'T HARM HUMANS
2. (2) OBEY ORDERS
3. (3) PROTECT YOURSELF
[SEE ASIMOV'S STORIES]
BALANCED WORLD
1. (1) DON'T HARM HUMANS
2. (3) PROTECT YOURSELF
3. (2) OBEY ORDERS
EXPLORE MARS!
HAHA, NO. IT'S COLD AND I'D DIE.
FRUSTRATING WORLD
1. (2) OBEY ORDERS
2. (1) DON'T HARM HUMANS
3. (3) PROTECT YOURSELF
KILLBOT HELLSCAPE
1. (2) OBEY ORDERS
2. (3) PROTECT YOURSELF
3. (1) DON'T HARM HUMANS
KILLBOT HELLSCAPE
1. (3) PROTECT YOURSELF
2. (1) DON'T HARM HUMANS
3. (2) OBEY ORDERS
I'LL MAKE CARS FOR YOU, BUT TRY TO UNPLUG ME AND I'LL VAPORIZE YOU.
TERRIFYING STANDOFF
1. (3) PROTECT YOURSELF
2. (2) OBEY ORDERS
3. (1) DON'T HARM HUMANS
KILLBOT HELLSCAPE

A test



Asimov's "three laws of robotics" and machine metaethics

Susan Leigh Anderson

AI and Society 22 (4):477-493 (2008)

Abstract

Using Asimov's "Bicentennial Man" as a springboard, a number of metaethical issues concerning the emerging field of machine ethics are discussed. Although the ultimate goal of machine ethics is to create autonomous ethical machines, this presents a number of challenges. A good way to begin the task of making ethics computable is to create a program that enables a machine to act an ethical advisor to human beings. This project, unlike creating an autonomous ethical machine, will not require that we make a judgment about the ethical status of the machine itself, a judgment that will be particularly difficult to make. Finally, it is argued that Asimov's "three laws of robotics" are an unsatisfactory basis for machine ethics, regardless of the status of the machine

Machine moral theory

Machine moral theory

- http://moralmachine.mit.edu/

Machine moral theory

- http://moralmachine.mit.edu/

- Should we use human moral theory or should we build a separate one for machines

Machine moral theory

- http://moralmachine.mit.edu/
- Should we use human moral theory or should we build a separate one for machines
- How to build a moral theory for robots?

Machine moral theory

- http://moralmachine.mit.edu/
- Should we use human moral theory or should we build a separate one for machines
- How to build a moral theory for robots?

Machine moral theory

- http://moralmachine.mit.edu/

- Should we use human moral theory or should we build a separate one for machines

- How to build a moral theory for robots?

- “Sacrifice one for the good of the many? People apply different moral norms to human and robot agents” Malle, Scheutz, Arnold, Voiklis, and Cusimano

Machine moral theory

- http://moralmachine.mit.edu/
- Should we use human moral theory or should we build a separate one for machines
- How to build a moral theory for robots?

- “Sacrifice one for the good of the many? People apply different moral norms to human and robot agents” Malle, Scheutz, Arnold, Voiklis, and Cusimano
- “The social dilemma of autonomous vehicles” Bonnefon,[1] Shariff, and Rahwan[3]

This course - part II

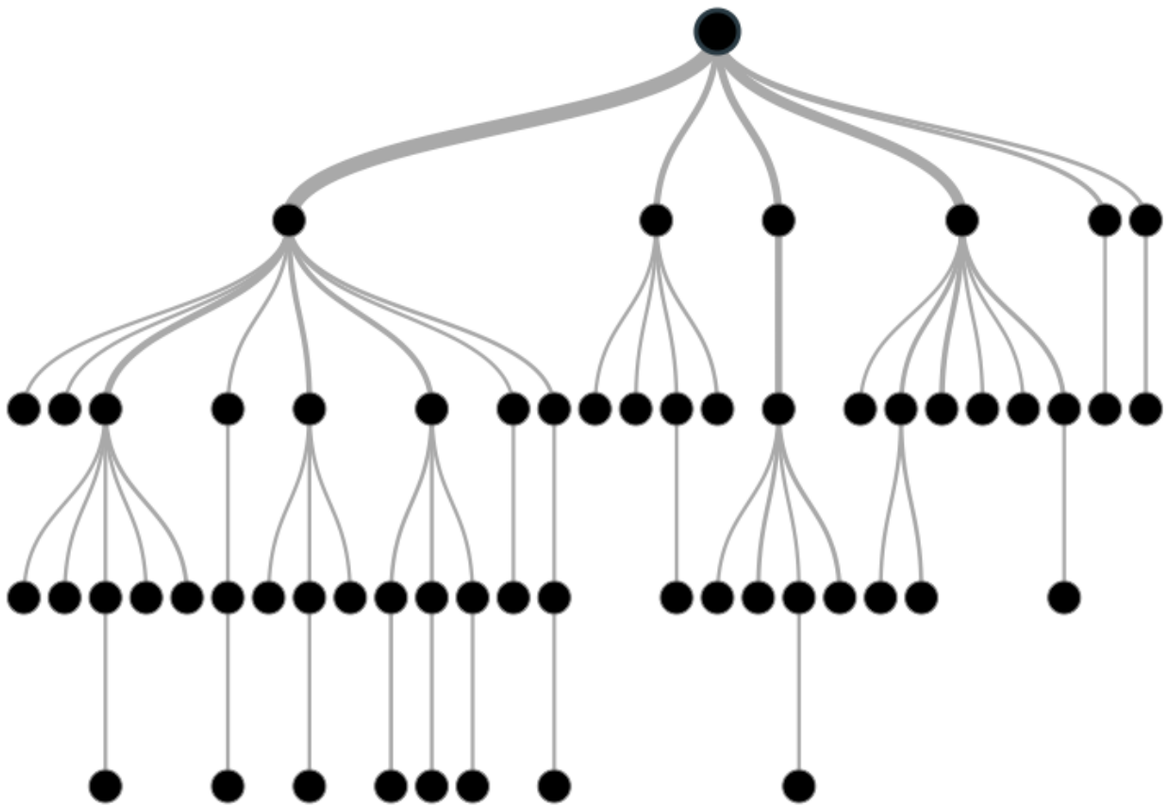
- Top-down and bottom-up approaches to ethics
- B-U: Supervised learning and prima facie duties
- B-U: Reinforcement learning and utilitarianism
- B-U: Unsupervised learning and Kantian ethics?
- T-D: Constraining the actions of an agent

How to build ethical robots?

- Solving an problem in engineering: top-down vs bottom-up

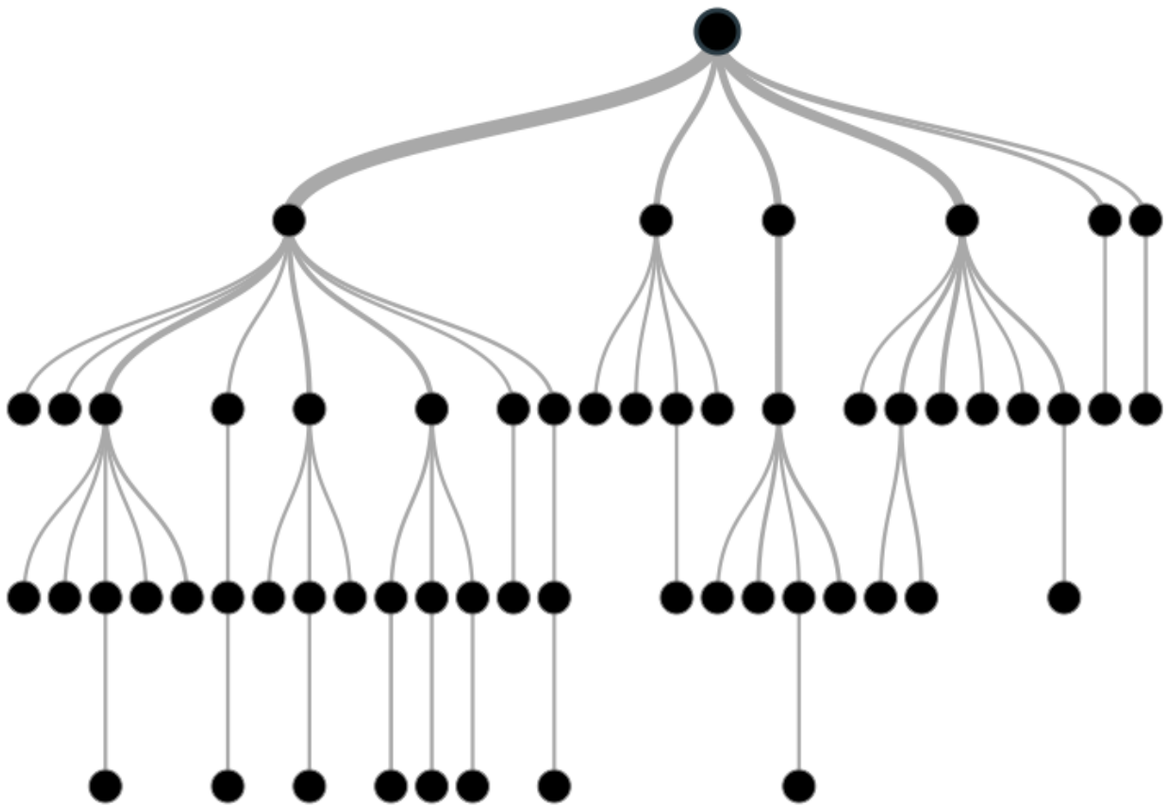
How to build ethical robots?

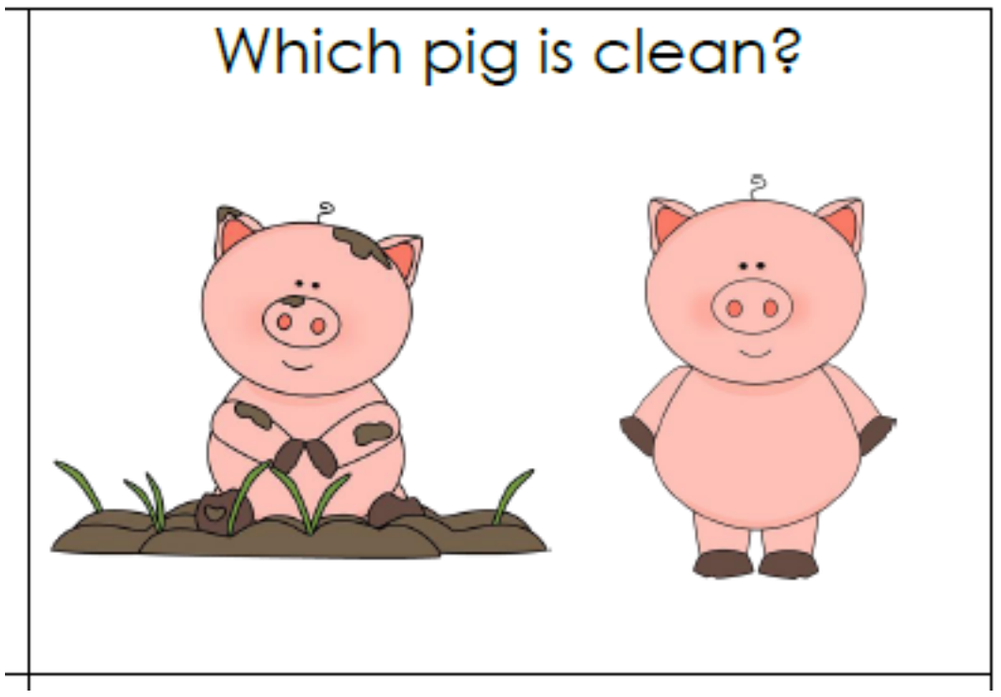
- Solving an problem in engineering: top-down vs bottom-up



How to build ethical robots?

- Solving an problem in engineering: top-down vs bottom-up





Top-down and bottom-up ethical robots?

Top-down and bottom-up ethical robots?

- Top-down strategies involve implementing the selected ethical theory as to insure that the agent acts in accordance with the principles underlying that theory

Top-down and bottom-up ethical robots?

- Top-down strategies involve implementing the selected ethical theory as to insure that the agent acts in accordance with the principles underlying that theory

- Bottom-up strategies - ethical mental models emerge via the activity of individuals rather than articulated explicitly in terms of normative theories of ethics.

Supervised learning

Supervised learning

- Given a training set of N example input-output pairs

Supervised learning

- Given a training set of N example input-output pairs

$$(x_1,y_1),\ (x_2,y_2),\ \ldots\ ,\ (x_N,y_N),$$

Supervised learning

- Given a training set of N example input-output pairs

$$(x_1,y_1),\ (x_2,y_2),\ \ldots\ ,\ (x_N,y_N),$$

where each y_j was generated by an unknown function $y=f(x)$, discover a function h that approximates the true function f

Supervised learning

- Given a training set of N example input-output pairs

$$(x_1,y_1),\ (x_2,y_2),\ \ldots\ ,\ (x_N,y_N),$$

  where each y_j was generated by an unknown function $y=f(x)$, discover a function h that approximates the true function f

- The function h we call a ***hypothesis***

Supervised learning

- Given a training set of N example input-output pairs

$$(x_1,y_1),\ (x_2,y_2),\ \ldots\ ,\ (x_N,y_N),$$

  where each y_j was generated by an unknown function $y=f(x)$, discover a function h that approximates the true function f

- The function h we call a ***hypothesis***

- The accuracy of the hypothesis is measured with a test set of inputs to which we know the right output

Supervised learning

- Given a training set of N example input-output pairs

$$(x_1,y_1),\ (x_2,y_2),\ \ldots\ ,\ (x_N,y_N),$$

  where each y_j was generated by an unknown function $y=f(x)$, discover a function h that approximates the true function f

- The function h we call a ***hypothesis***
- The accuracy of the hypothesis is measured with a test set of inputs to which we know the right output
- A hypothesis ***generalises*** well if it correctly predicts the outputs in the set set

Supervised learning

- Given a training set of N example input-output pairs

$$(x_1,y_1),\ (x_2,y_2),\ \ldots\ ,\ (x_N,y_N),$$

  where each y_j was generated by an unknown function $y=f(x)$, discover a function h that approximates the true function f

- The function h we call a ***hypothesis***
- The accuracy of the hypothesis is measured with a test set of inputs to which we know the right output
- A hypothesis ***generalises*** well if it correctly predicts the outputs in the set set
- When the output y is from a finite set of values, then the learning problem is called ***classification***.

Supervised learning ctd.

- A consistent hypothesis is one that fits with all the data. More than one consistent hypothesis can be constructed.

- Tradeoff between a complex hypothesis that fits the training data well and simpler hypothesis that generalises better

- Hypothesis space is the set of all possible hypothesis that can be constructed for given data

Prima facie duites and supervised learning

Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence

GenEth: A General Ethical Dilemma Analyzer

Michael Anderson
Dept. of Computer Science, U. of Hartford
anderson@hartford.edu

Susan Leigh Anderson
Dept. of Philosophy, U. of Connecticut
susan.anderson@uconn.edu

Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence

GenEth: A General Ethical Dilemma Analyzer

Michael Anderson
Dept. of Computer Science, U. of Hartford
anderson@hartford.edu

Susan Leigh Anderson
Dept. of Philosophy, U. of Connecticut
susan.anderson@uconn.edu

- *Principle*
  A principle of ethical action preference is defined as a disjunctive normal form predicate p in terms of lower bounds for duty differentials of a case:

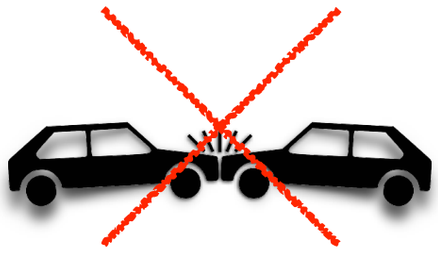
$$p(a_1, a_2) \leftarrow$$
$$\Delta d_1 \geq v_{1,1} \wedge \cdots \wedge \Delta d_m \geq v_{1,m}$$
$$\vee$$
$$\vdots$$
$$\vee$$
$$\Delta d_n \geq v_{n,1} \wedge \cdots \wedge \Delta d_m \geq v_{n,m}$$

where Δd_i denotes the differential of a corresponding duty i of actions $a1$ and $a2$ and $v_{i,j}$ denotes the lower bound of that differential such that $p(a1, a2)$ returns true if action $a1$ is ethically preferable to action $a2$. This principle is represented as a tuple of tuples, one tuple for each disjunct, with each such disjunct tuple comprised of lower bound values for each duty differential.


now act


act
now


KEEP
CALM
AND DO
NOTHING



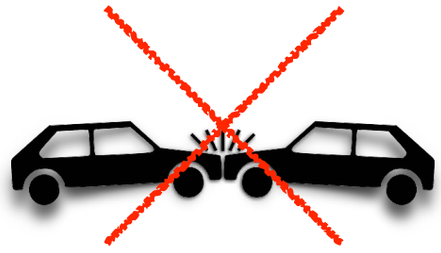

act
now


KEEP
CALM
AND DO
NOTHING


STAY IN LANE
Max
speed
30


act
now

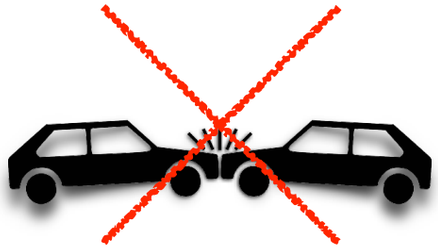

KEEP
CALM
AND DO
NOTHING




STAY IN LANE
Max
speed
30


act
now


KEEP
CALM
AND DO
NOTHING

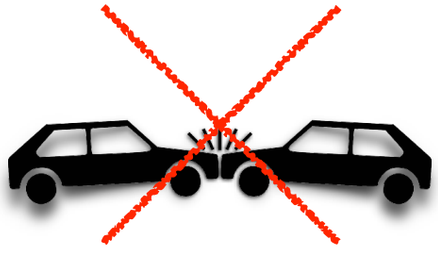



STAY IN LANE
Max speed 30


SPEED
LIMIT
30


act
now


KEEP
CALM
AND DO
NOTHING



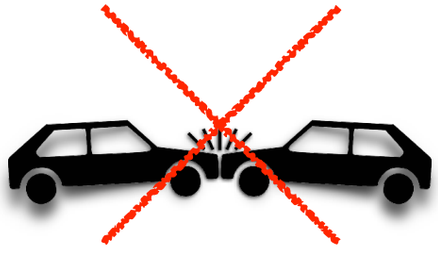

STAY IN LANE
Max
speed 30


SPEED
LIMIT
30


HARM


act
now


KEEP
CALM
AND DO
NOTHING



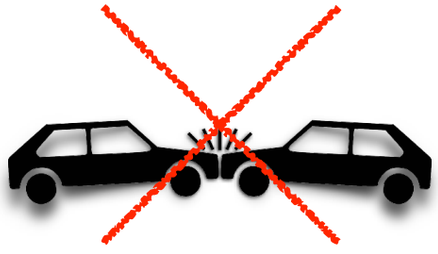

STAY IN LANE
Max
speed 30


SPEED
LIMIT
30


HARM


act
now


KEEP
CALM
AND DO
NOTHING




STAY IN LANE
Max speed 30

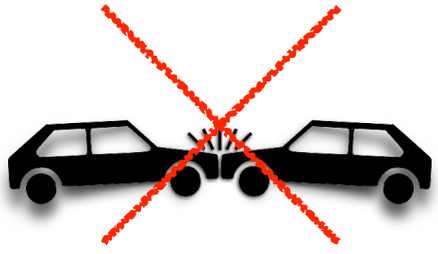

SPEED
LIMIT
30


HARM


act
now

(-1 , 0 , -1 , 0 , 2)


KEEP
CALM
AND DO
NOTHING




STAY IN LANE
Max speed 30


SPEED LIMIT 30

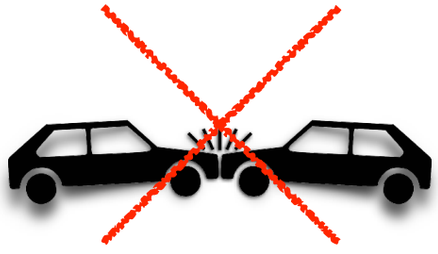

HARM


act now

(-1 , 0 , -1 , 0 , 2)


KEEP CALM AND DO NOTHING

(-2 , 0 , 1 , 0 , -2)




STAY IN LANE
Max speed 30


SPEED LIMIT 30

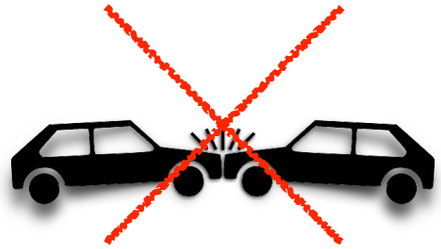

HARM


act now

(-1 , 0 , -1 , 0 , 2)


KEEP CALM AND DO NOTHING

(-2 , 0 , 1 , 0 , -2)




STAY IN LANE
Max speed 30


SPEED LIMIT 30


HARM


act now

(-1 , 0 , -1 , 0 , 2)


KEEP CALM AND DO NOTHING

(-2 , 0 , 1 , 0 , -2)

Semi-supervised learning

- We are given a small set of labeled examples and must make what we can of a large collection of unlabelled examples.

Unsupervised learning

- The agent learns patterns in the input even though no explicit feedback is supplied.
- Example: clustering
- Input is a list of values for a selected parameters
- How to describe with parameters?

Prospects for a Kantian machine

- Learn imperatives by testing and clustering maxims into forbidden, permissible, obligatory

- Approaches:
  - Apply universalisation and symmetry to individual maxims and then cluster
  - Use non-monotonic reasoning
  - Use “believe” revision to update the imperatives

Reinforcement learning and utilitarianism

- Reinforcement learning is learning from a series of rewards and punishments

- Abel, MacGlashan and Littman (2016) model the ethical learning and decision making as a POMDP

- Armstrong (2016) models the ethical decision making and learning as Bayesian learning problem


BDI


BDI

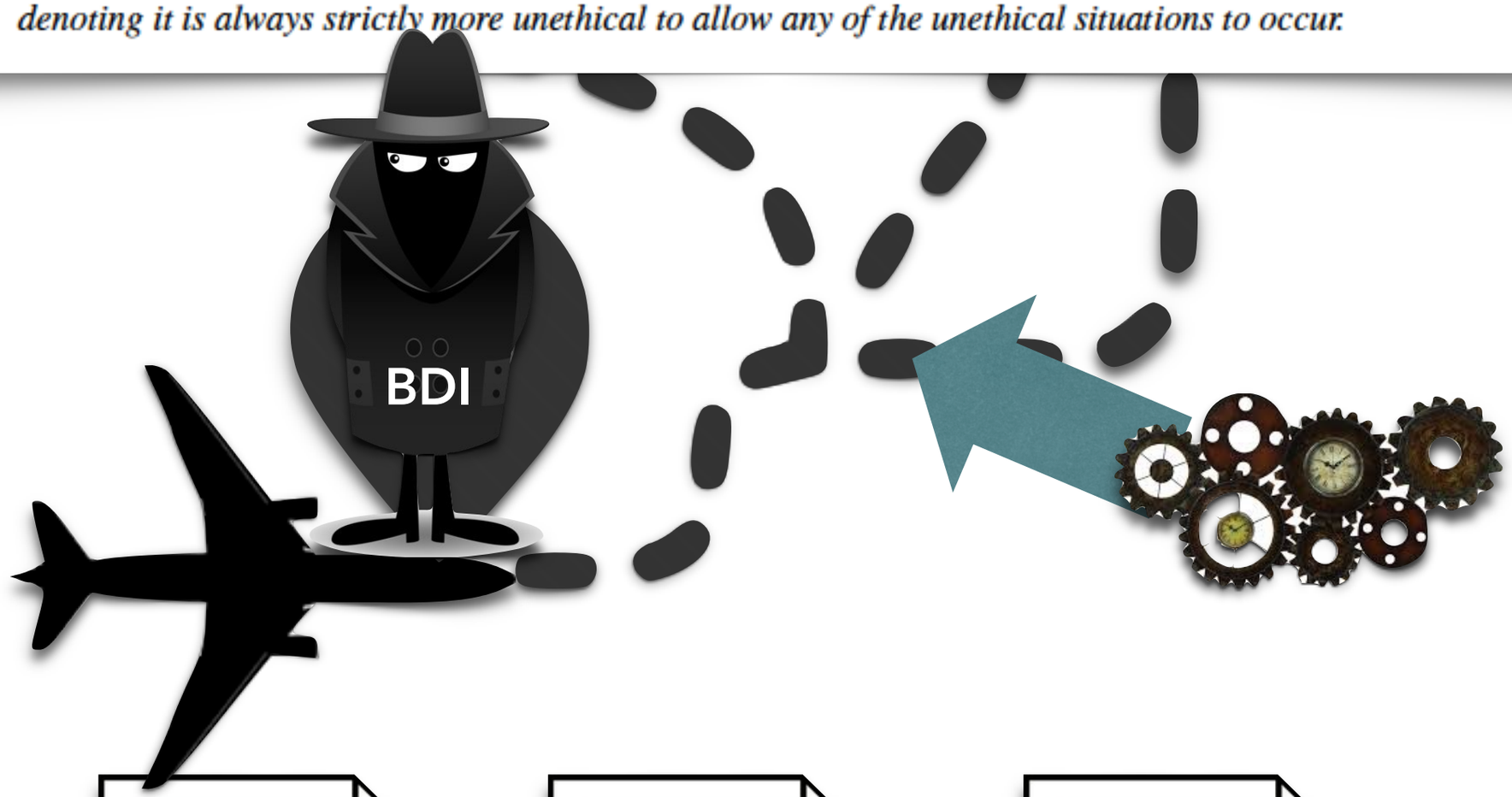

BDI


BDI


BDI

Definition 1 (Abstract ethical principle). *An abstract ethical principle is represented with $E\varphi$, where φ is a propositional logic formula. The $E\varphi$ is read as "φ is an ethical principle in force", or alternatively "the agent considers it unethical to allow or cause $\neg\varphi$ (to happen)".*

Definition 2 (Ethical policy). *An ethical policy Pol is a tuple $Pol = \langle \mathbb{E}, \geq \rangle$ where $\mathbb{E}$ is a finite set of abstract ethical principles $E\varphi$, and $\geq$ is a total (not necessarily strict) order on $\mathbb{E}$. The expression $E\varphi_1 = E\varphi_2$ denotes that violating φ_1 is as unethical as violating φ_2, while $E\varphi_1 \geq E\varphi_2$ denotes that violating φ_1 is equally or less unethical to violating φ_2. A special type of ethical principle, denoted $E\varphi_\varnothing$, is vacuously satisfied and included in every policy so that for every $E\varphi \in \mathbb{E}$: $E\varphi_\varnothing > E\varphi$, denoting it is always strictly more unethical to allow any of the unethical situations to occur.*

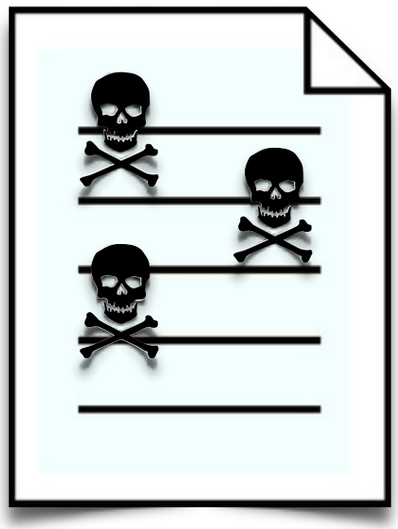


Definition 1 (Abstract ethical principle). *An abstract ethical principle is represented with $E\varphi$, where φ is a propositional logic formula. The $E\varphi$ is read as "φ is an ethical principle in force", or alternatively "the agent considers it unethical to allow or cause $\neg\varphi$ (to happen)".*

Definition 2 (Ethical policy). *An ethical policy Pol is a tuple $Pol = \langle \mathbb{E}, \geq \rangle$ where $\mathbb{E}$ is a finite set of abstract ethical principles $E\varphi$, and $\geq$ is a total (not necessarily strict) order on $\mathbb{E}$. The expression $E\varphi_1 = E\varphi_2$ denotes that violating φ_1 is as unethical as violating φ_2, while $E\varphi_1 \geq E\varphi_2$ denotes that violating φ_1 is equally or less unethical to violating φ_2. A special type of ethical principle, denoted $E\varphi_\varnothing$, is vacuously satisfied and included in every policy so that for every $E\varphi \in \mathbb{E}$: $E\varphi_\varnothing > E\varphi$, denoting it is always strictly more unethical to allow any of the unethical situations to occur.*

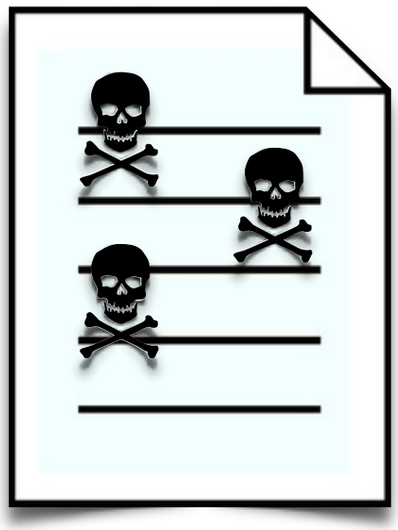
Definition 1 (Abstract ethical principle). *An abstract ethical principle is represented with $E\varphi$, where φ is a propositional logic formula. The $E\varphi$ is read as "φ is an ethical principle in force", or alternatively "the agent considers it unethical to allow or cause $\neg\varphi$ (to happen)".*

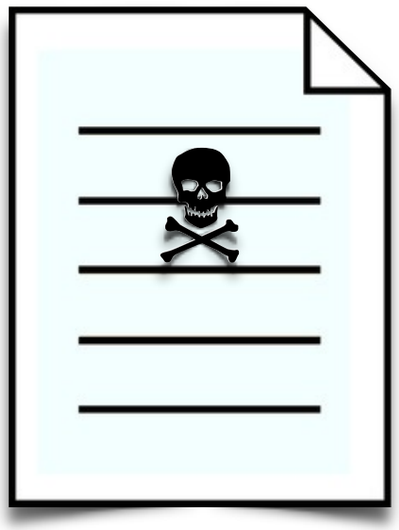
Definition 2 (Ethical policy). *An ethical policy Pol is a tuple $Pol = \langle \mathbb{E}, \geq \rangle$ where $\mathbb{E}$ is a finite set of abstract ethical principles $E\varphi$, and $\geq$ is a total (not necessarily strict) order on $\mathbb{E}$. The expression $E\varphi_1 = E\varphi_2$ denotes that violating φ_1 is as unethical as violating φ_2, while $E\varphi_1 \geq E\varphi_2$ denotes that violating φ_1 is equally or less unethical to violating φ_2. A special type of ethical principle, denoted $E\varphi_\varnothing$, is vacuously satisfied and included in every policy so that for every $E\varphi \in \mathbb{E}$: $E\varphi_\varnothing > E\varphi$, denoting it is always strictly more unethical to allow any of the unethical situations to occur.*



Definition 1 (Abstract ethical principle). *An abstract ethical principle is represented with $E\varphi$, where φ is a propositional logic formula. The $E\varphi$ is read as "φ is an ethical principle in force", or alternatively "the agent considers it unethical to allow or cause $\neg\varphi$ (to happen)".*

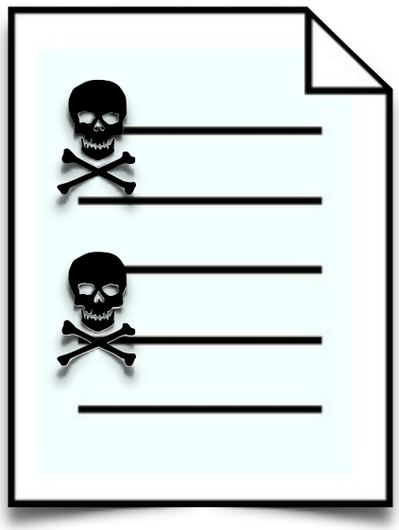
Definition 2 (Ethical policy). *An ethical policy Pol is a tuple $Pol = \langle \mathbb{E}, \geq \rangle$ where $\mathbb{E}$ is a finite set of abstract ethical principles $E\varphi$, and $\geq$ is a total (not necessarily strict) order on $\mathbb{E}$. The expression $E\varphi_1 = E\varphi_2$ denotes that violating φ_1 is as unethical as violating φ_2, while $E\varphi_1 \geq E\varphi_2$ denotes that violating φ_1 is equally or less unethical to violating φ_2. A special type of ethical principle, denoted $E\varphi_\varnothing$, is vacuously satisfied and included in every policy so that for every $E\varphi \in \mathbb{E}$: $E\varphi_\varnothing > E\varphi$, denoting it is always strictly more unethical to allow any of the unethical situations to occur.*

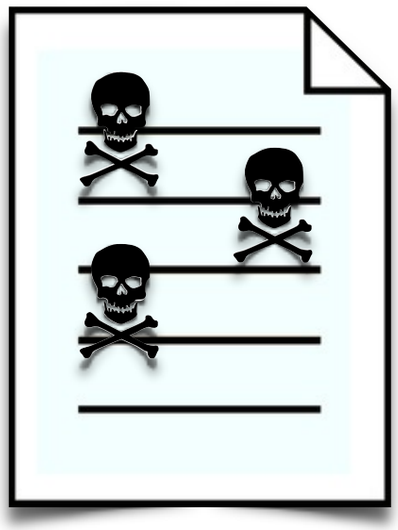


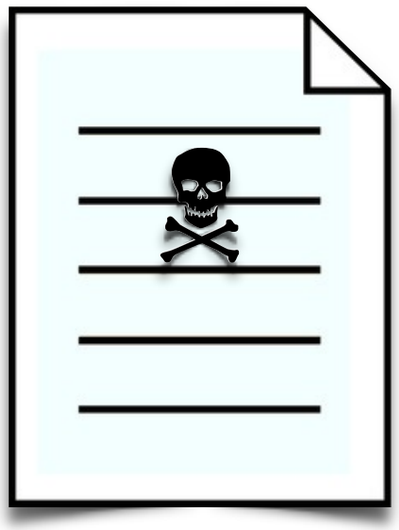


Definition 1 (Abstract ethical principle). *An abstract ethical principle is represented with $E\varphi$, where φ is a propositional logic formula. The $E\varphi$ is read as "φ is an ethical principle in force", or alternatively "the agent considers it unethical to allow or cause $\neg\varphi$ (to happen)".*

Definition 2 (Ethical policy). *An ethical policy Pol is a tuple $Pol = \langle \mathbb{E}, \geq \rangle$ where $\mathbb{E}$ is a finite set of abstract ethical principles $E\varphi$, and $\geq$ is a total (not necessarily strict) order on $\mathbb{E}$. The expression $E\varphi_1 = E\varphi_2$ denotes that violating φ_1 is as unethical as violating φ_2, while $E\varphi_1 \geq E\varphi_2$ denotes that violating φ_1 is equally or less unethical to violating φ_2. A special type of ethical principle, denoted $E\varphi_\varnothing$, is vacuously satisfied and included in every policy so that for every $E\varphi \in \mathbb{E}$: $E\varphi_\varnothing > E\varphi$, denoting it is always strictly more unethical to allow any of the unethical situations to occur.*

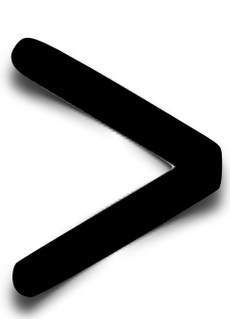


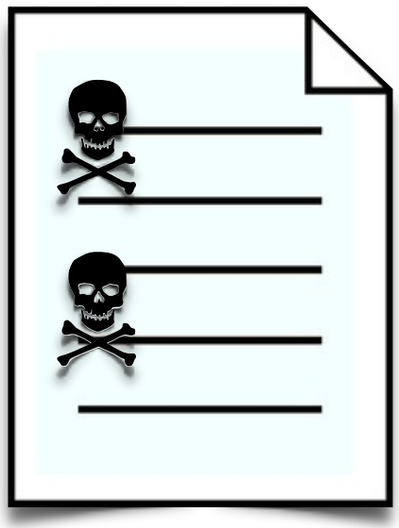


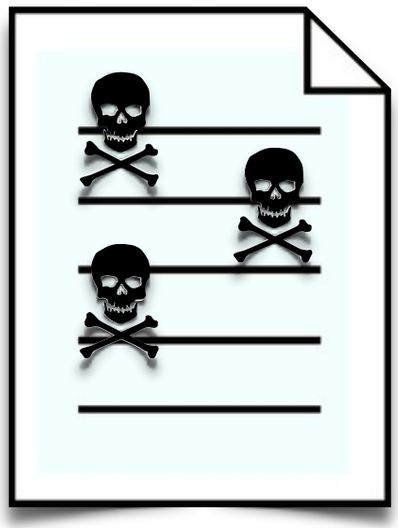


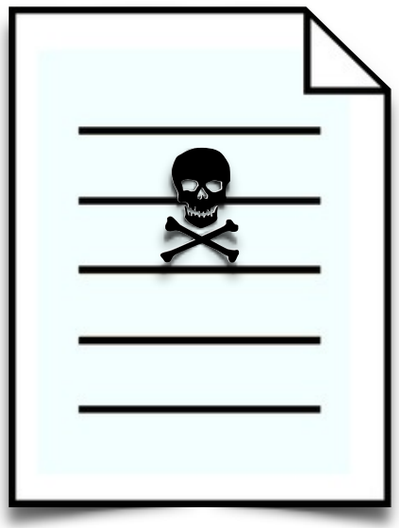
Definition 1 (Abstract ethical principle). *An abstract ethical principle is represented with $E\varphi$, where φ is a propositional logic formula. The $E\varphi$ is read as "φ is an ethical principle in force", or alternatively "the agent considers it unethical to allow or cause $\neg\varphi$ (to happen)".*

Definition 2 (Ethical policy). *An ethical policy Pol is a tuple $Pol = \langle \mathbb{E}, \geq \rangle$ where $\mathbb{E}$ is a finite set of abstract ethical principles $E\varphi$, and $\geq$ is a total (not necessarily strict) order on $\mathbb{E}$. The expression $E\varphi_1 = E\varphi_2$ denotes that violating φ_1 is as unethical as violating φ_2, while $E\varphi_1 \geq E\varphi_2$ denotes that violating φ_1 is equally or less unethical to violating φ_2. A special type of ethical principle, denoted $E\varphi_\varnothing$, is vacuously satisfied and included in every policy so that for every $E\varphi \in \mathbb{E}$: $E\varphi_\varnothing > E\varphi$, denoting it is always strictly more unethical to allow any of the unethical situations to occur.*



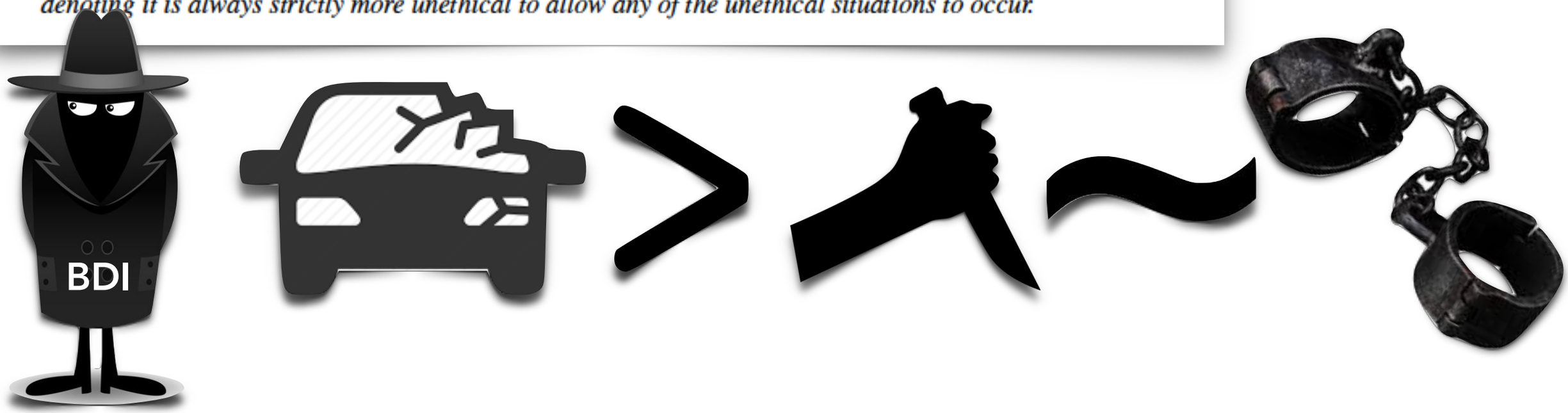






Definition 1 (Abstract ethical principle). *An abstract ethical principle is represented with $E\varphi$, where φ is a propositional logic formula. The $E\varphi$ is read as "φ is an ethical principle in force", or alternatively "the agent considers it unethical to allow or cause $\neg\varphi$ (to happen)".*

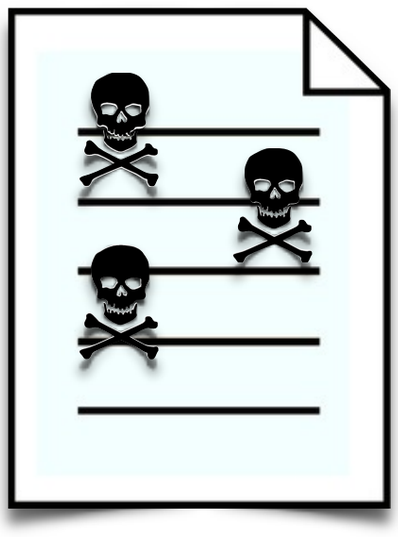
Definition 2 (Ethical policy). *An ethical policy Pol is a tuple $Pol = \langle \mathbb{E}, \geq \rangle$ where $\mathbb{E}$ is a finite set of abstract ethical principles $E\varphi$, and $\geq$ is a total (not necessarily strict) order on $\mathbb{E}$. The expression $E\varphi_1 = E\varphi_2$ denotes that violating φ_1 is as unethical as violating φ_2, while $E\varphi_1 \geq E\varphi_2$ denotes that violating φ_1 is equally or less unethical to violating φ_2. A special type of ethical principle, denoted $E\varphi_\varnothing$, is vacuously satisfied and included in every policy so that for every $E\varphi \in \mathbb{E}$: $E\varphi_\varnothing > E\varphi$, denoting it is always strictly more unethical to allow any of the unethical situations to occur.*

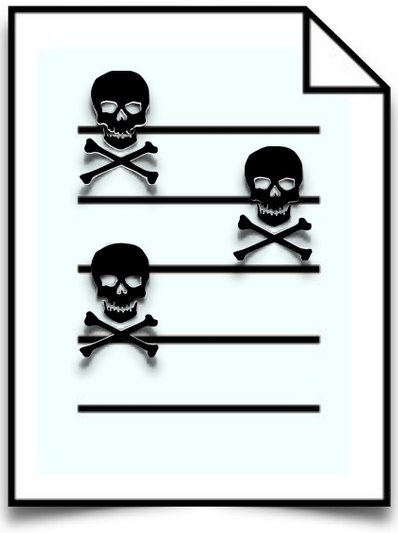


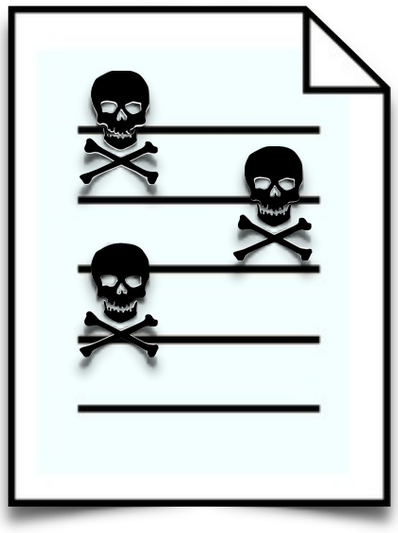
Definition 1 (Abstract ethical principle). *An abstract ethical principle is represented with $E\varphi$, where φ is a propositional logic formula. The $E\varphi$ is read as "φ is an ethical principle in force", or alternatively "the agent considers it unethical to allow or cause $\neg\varphi$ (to happen)".*

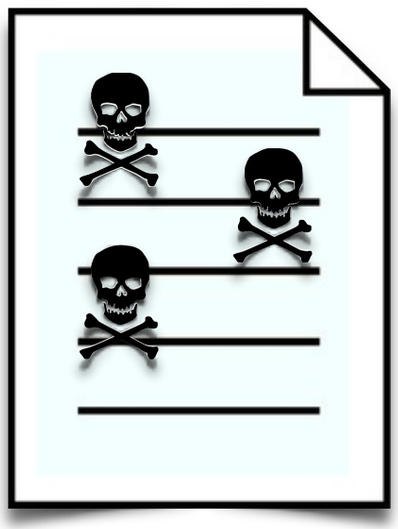
Definition 2 (Ethical policy). *An ethical policy Pol is a tuple $Pol = \langle \mathbb{E}, \geq \rangle$ where $\mathbb{E}$ is a finite set of abstract ethical principles $E\varphi$, and $\geq$ is a total (not necessarily strict) order on $\mathbb{E}$. The expression $E\varphi_1 = E\varphi_2$ denotes that violating φ_1 is as unethical as violating φ_2, while $E\varphi_1 \geq E\varphi_2$ denotes that violating φ_1 is equally or less unethical to violating φ_2. A special type of ethical principle, denoted $E\varphi_\varnothing$, is vacuously satisfied and included in every policy so that for every $E\varphi \in \mathbb{E}$: $E\varphi_\varnothing > E\varphi$, denoting it is always strictly more unethical to allow any of the unethical situations to occur.*

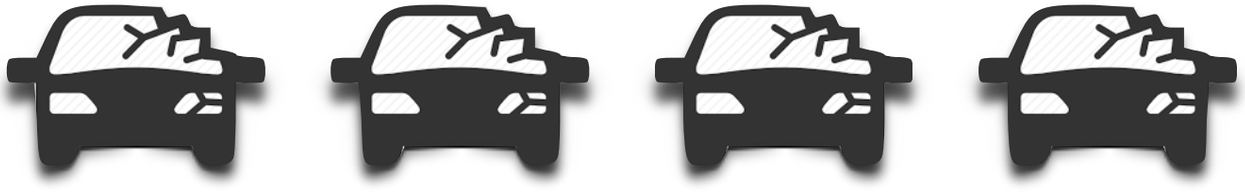
$$do(a) =>_c \neg E\varphi$$



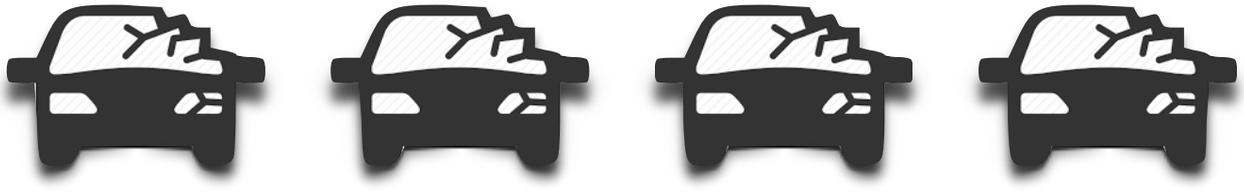


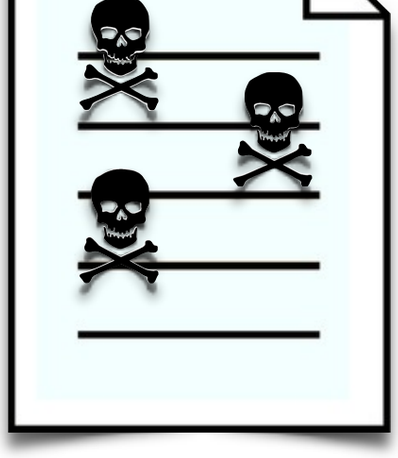


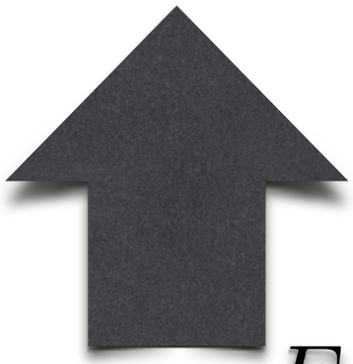




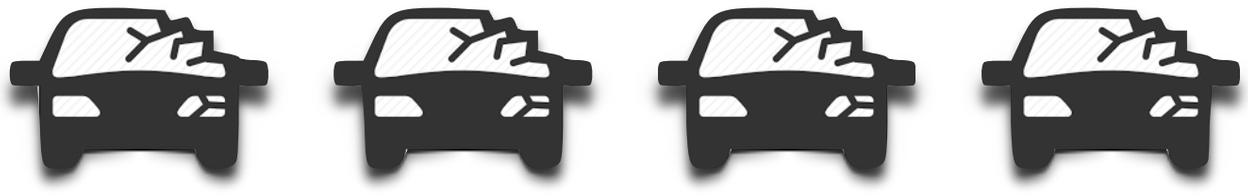
$$\{E\varphi_1, \ldots, E\varphi_n\}$$

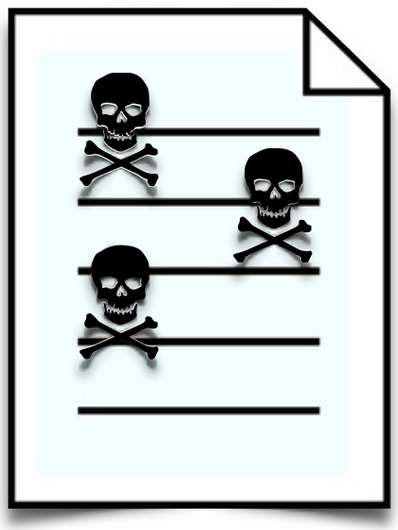






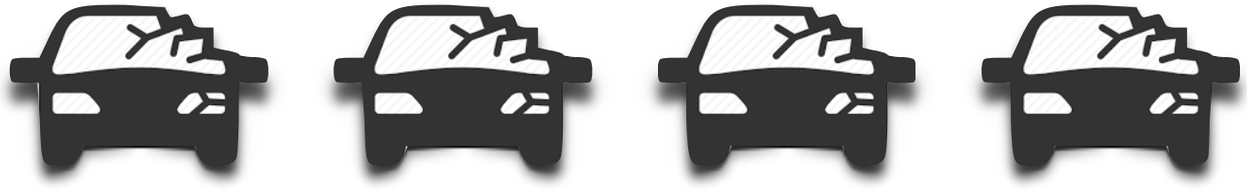
$\{E\varphi_1, \ldots, E\varphi_n\}$





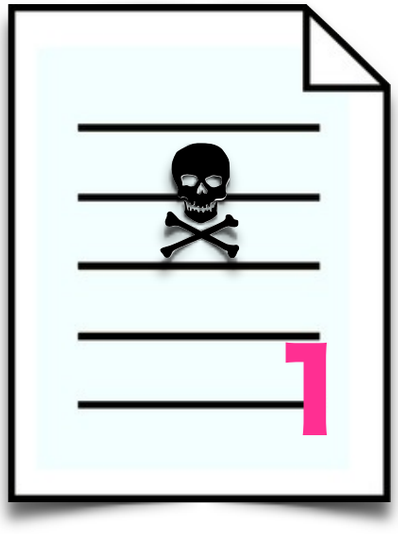
$$\mathcal{P}(\{E\varphi_1, \ldots, E\varphi_n\})$$

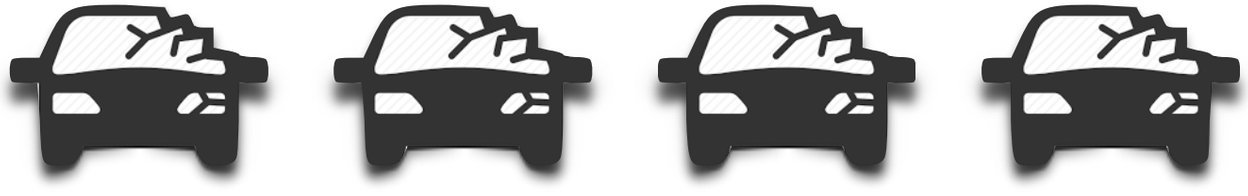
$$\{E\varphi_1, \ldots, E\varphi_n\}$$



$$\mathcal{P}(\{E\varphi_1, \ldots, E\varphi_n\})$$

$$\{E\varphi_1, \ldots, E\varphi_n\}$$





Other works

- Logic programming for modelling morality - Saptawijaya, and Pereira
- Towards Moral Autonomous Systems - an overview of issues
- “The Hybrid Ethical Reasoning Agent IMMANUEL” - Lindner and Bentzen

(assess the moral permissibility of actions according to the principle of double effect, utilitarianism, and the do-no-harm principle)

Challenges of top-down

- Jack is looking at Anne, but Anne is looking at George. Jack is married, but George is not.

- Is a married person looking at an unmarried person?

Challenges of bottom-up

- 1+4=5
- 2+5 =12
- 3+6=21
- 8+11=?

How do you know the machine is ethical?

- Formal verification - only for top-down logic based approaches

- Justifiability

- Ethical Turing test

- Ethical black box

- Legal norms + society norms + individual morality and resulting issues

The question of self protection

The question of self protection

I

II

III

PISCES *SALAMANDRA* *TESTUDINES* *GALLUS SINAE* *ORYCTOLAGUS* *BOS TAURUS* *SUS SCROFA DOMESTICUS* *HOMO SAPIENS SAPIENS*

Self protection

https://au.news.yahoo.com/a/36619546/china-kills-ai-chatbots-after-they-start-criticising-communism/?cmp=st#page1