Automating moral reasoning

Marija Slavkovik
@MSlavkovik

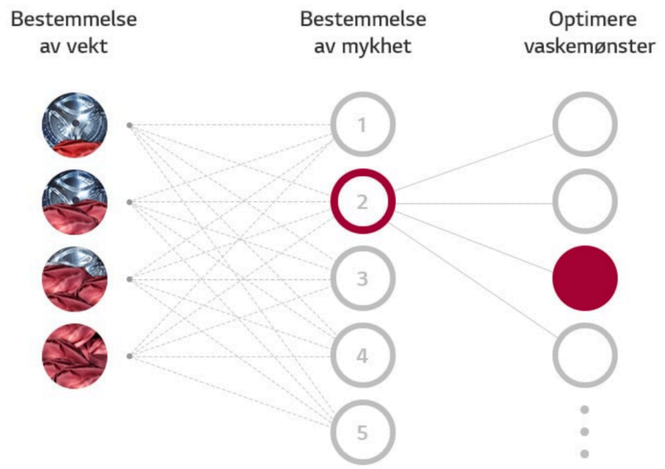
University of Bergen

Overview

- Why automating moral reasoning is necessary
- What are artificial moral agents?
- Handling the ethical impact of AI
- Two kinds of artificial moral agents
- What is moral?
- Top-down approaches
- Bottom-up approaches
- Why so little machine learning?

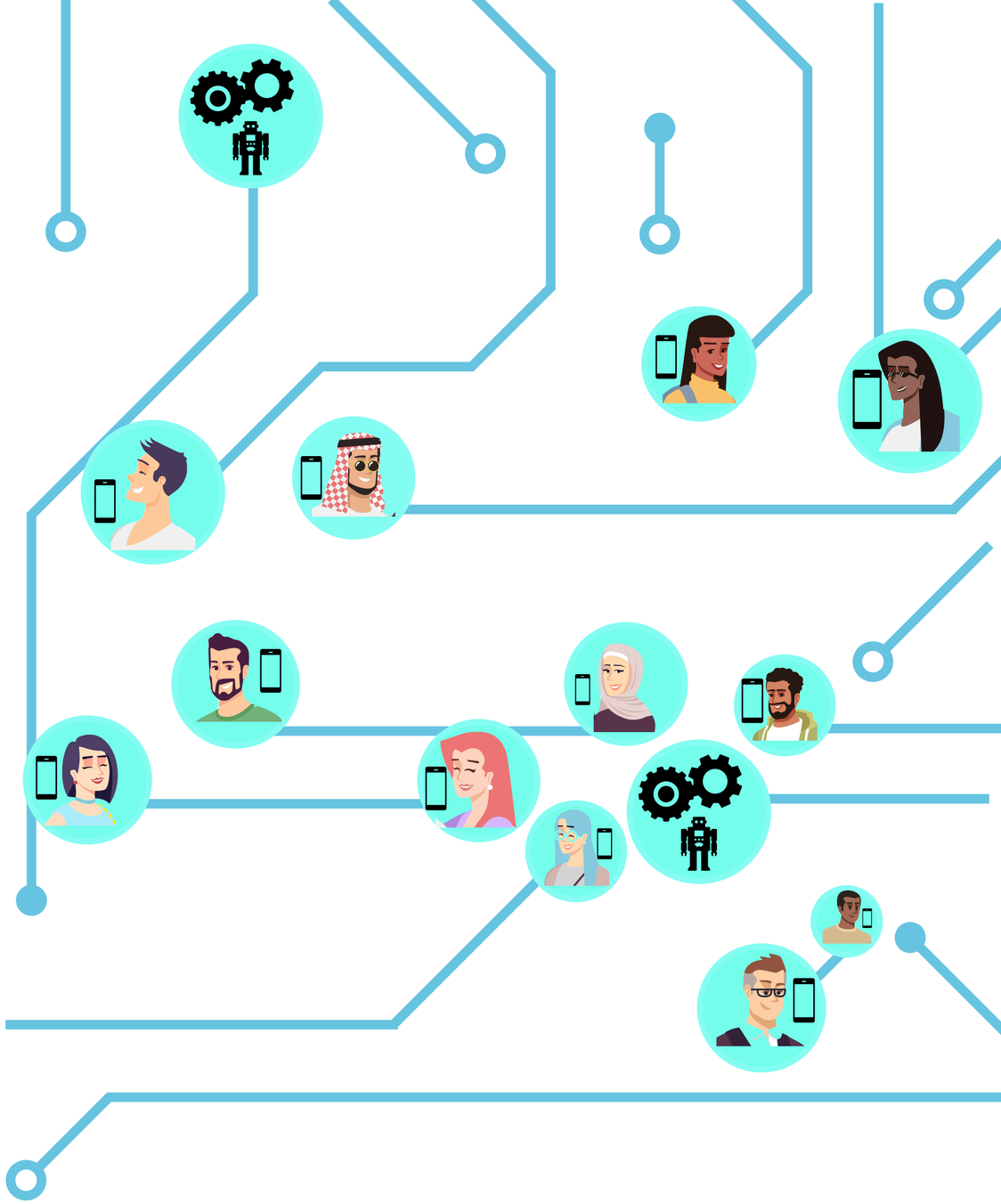
Digital society

**devices that can do some
things by themselves and the humans that use them**



What is AI DD ™?

AI DD ™ reads the weight and also feels the softness of the clothes and then uses the best drum movements for the clothes in the machine.

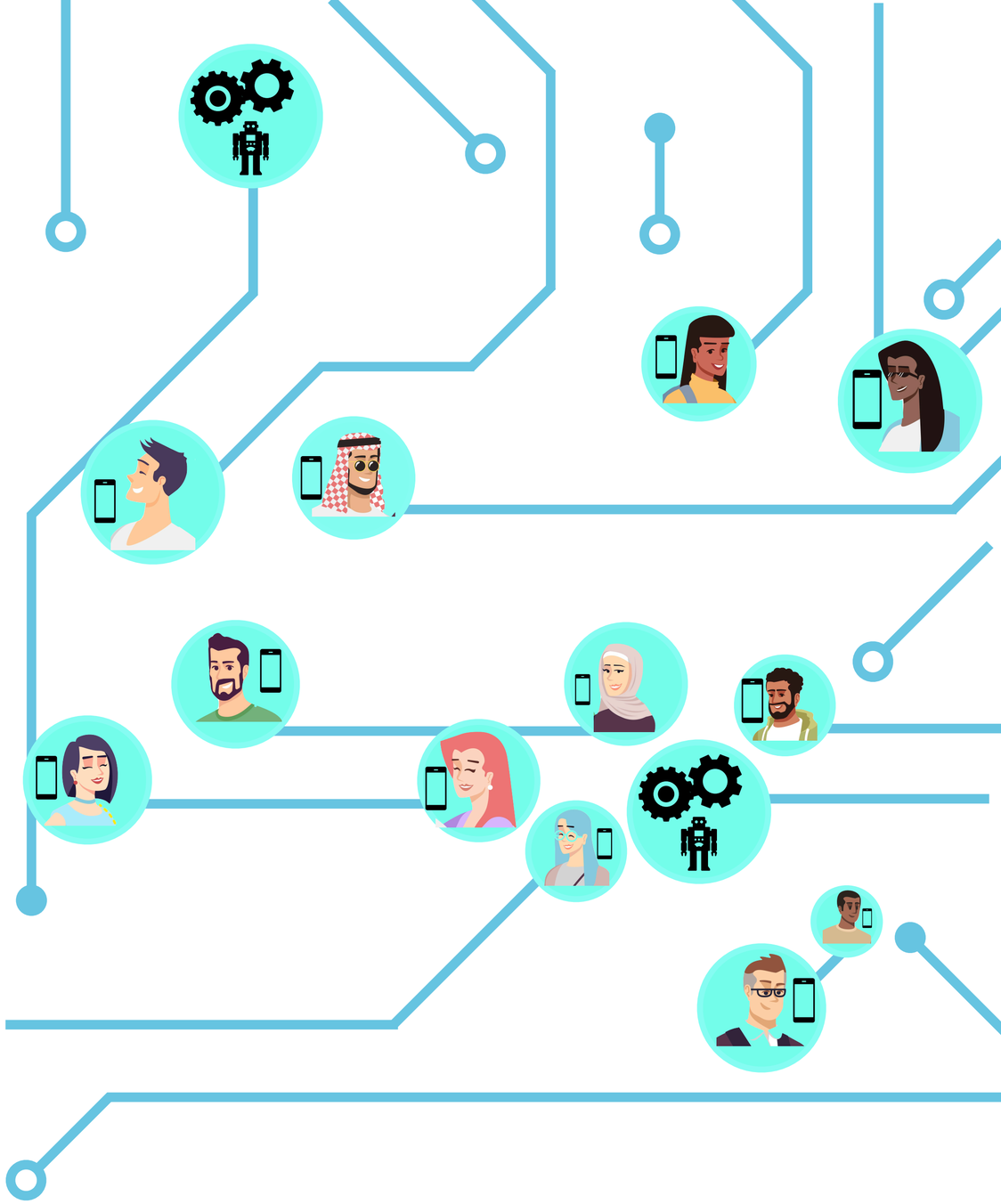


Digital society

devices that can do some
things by themselves and the humans that use them

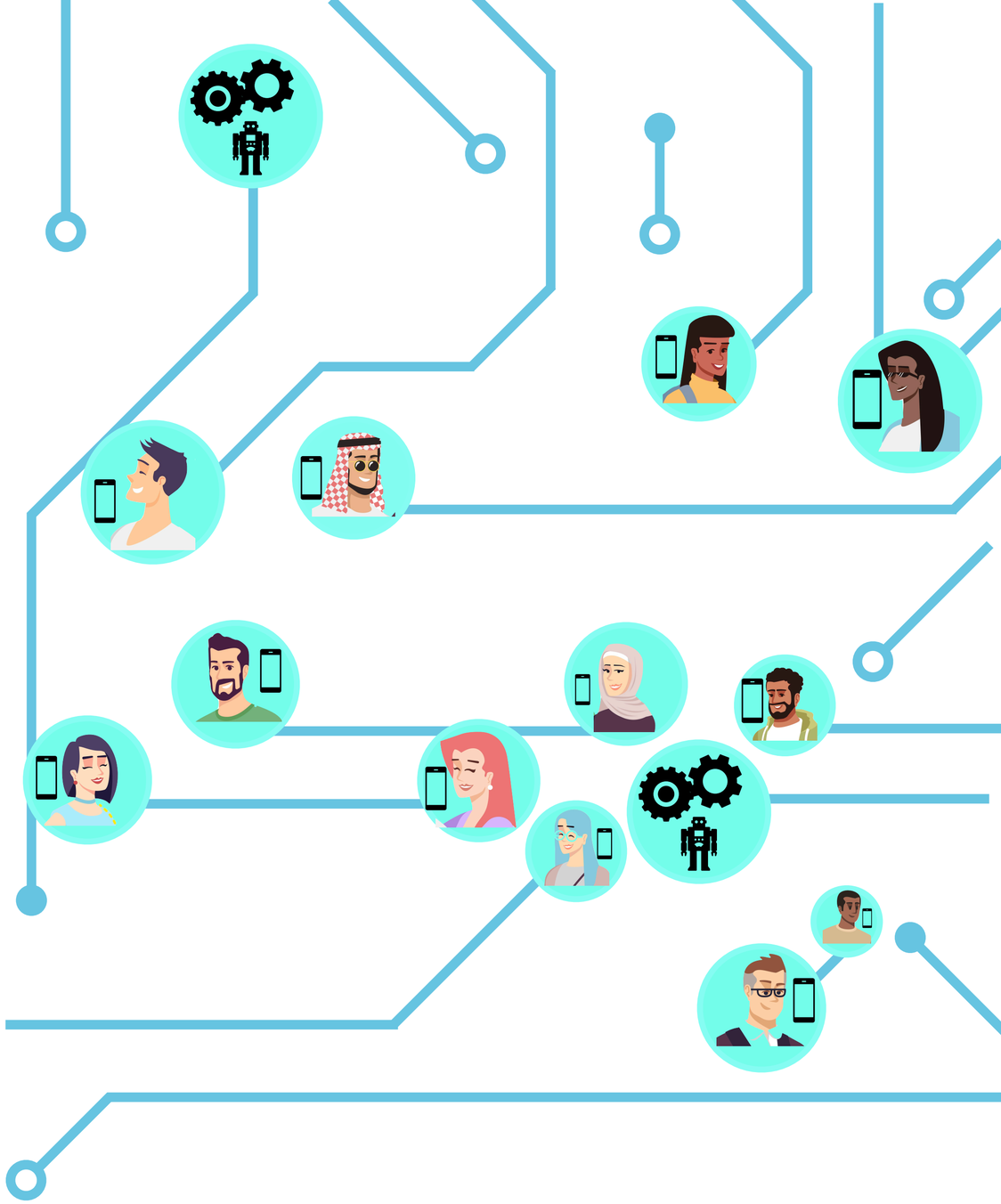
Smart Diagnosis ™

With Smart Diagnosis, you can quickly troubleshoot most minor issues that may arise.



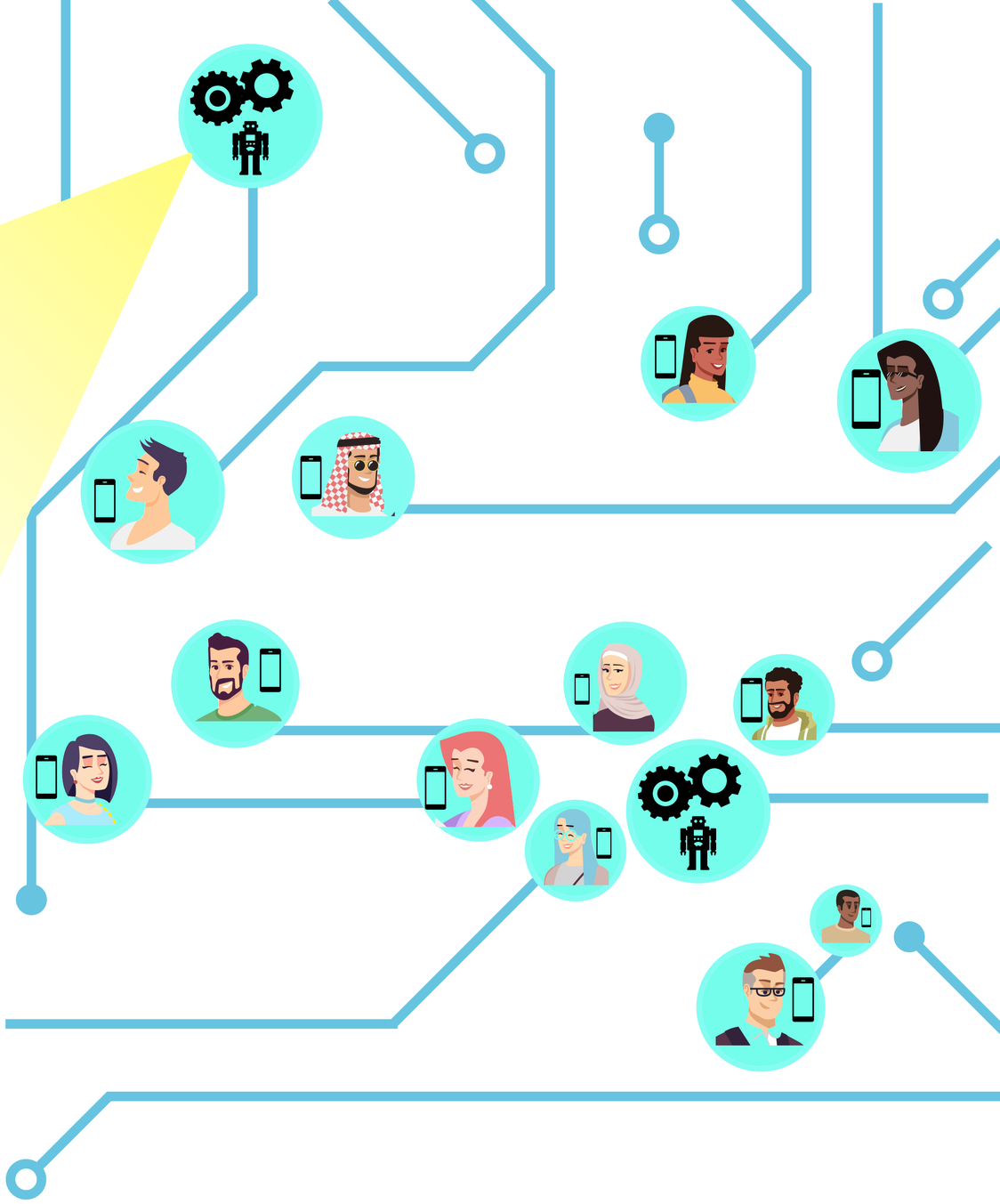
Digital society

**devices that can do some
things by themselves and the humans that use them**



Digital society

devices that can do some
things by themselves and the humans that use them



Utgiftsrefusjon

Utgifter

Rediger

Type	Vedlegg	Beskrivelse	Beløp	Kurs	Beløp NOK
Faglitteratur/bøker	1	Hands-On Explainable AI (XAI)	44,85 GBP	11,629	521,55
Totalt					**521,55**

Vedlegg

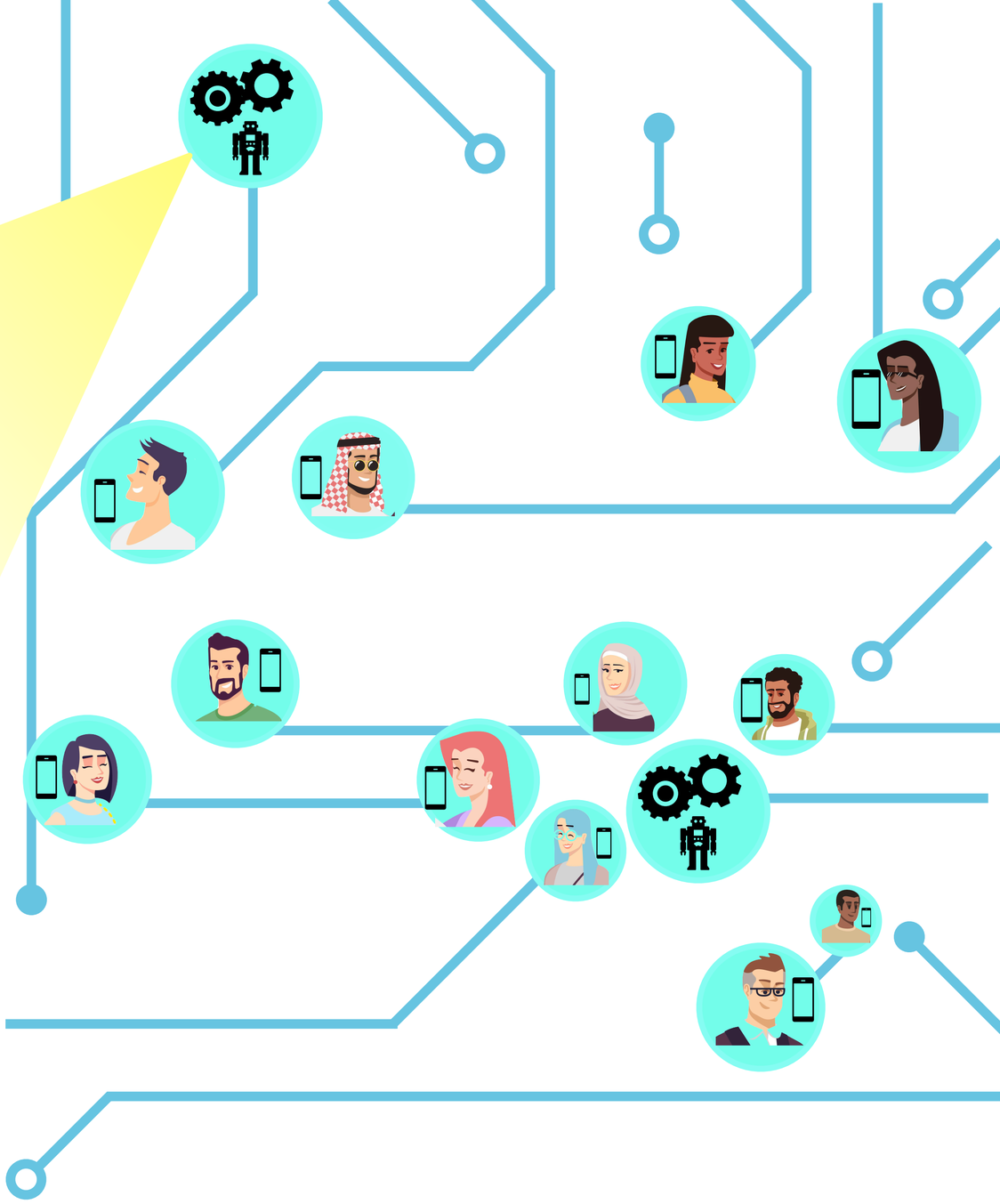
Rediger

Vedlegg fra utgiftsposter

Faglitteratur/bøker (44.85 GBP) - 1
pdf

Digital society

devices that can do some
things by themselves and the humans that use them



Utgiftsrefusjon

Utgifter

Rediger

Type	Vedlegg	Beskrivelse	Beløp	Kurs	Beløp NOK
Faglitteratur/bøker	1	Hands-On Explainable AI (XAI)	44,85 GBP	11,629	521,55
Totalt					**521,55**

Vedlegg

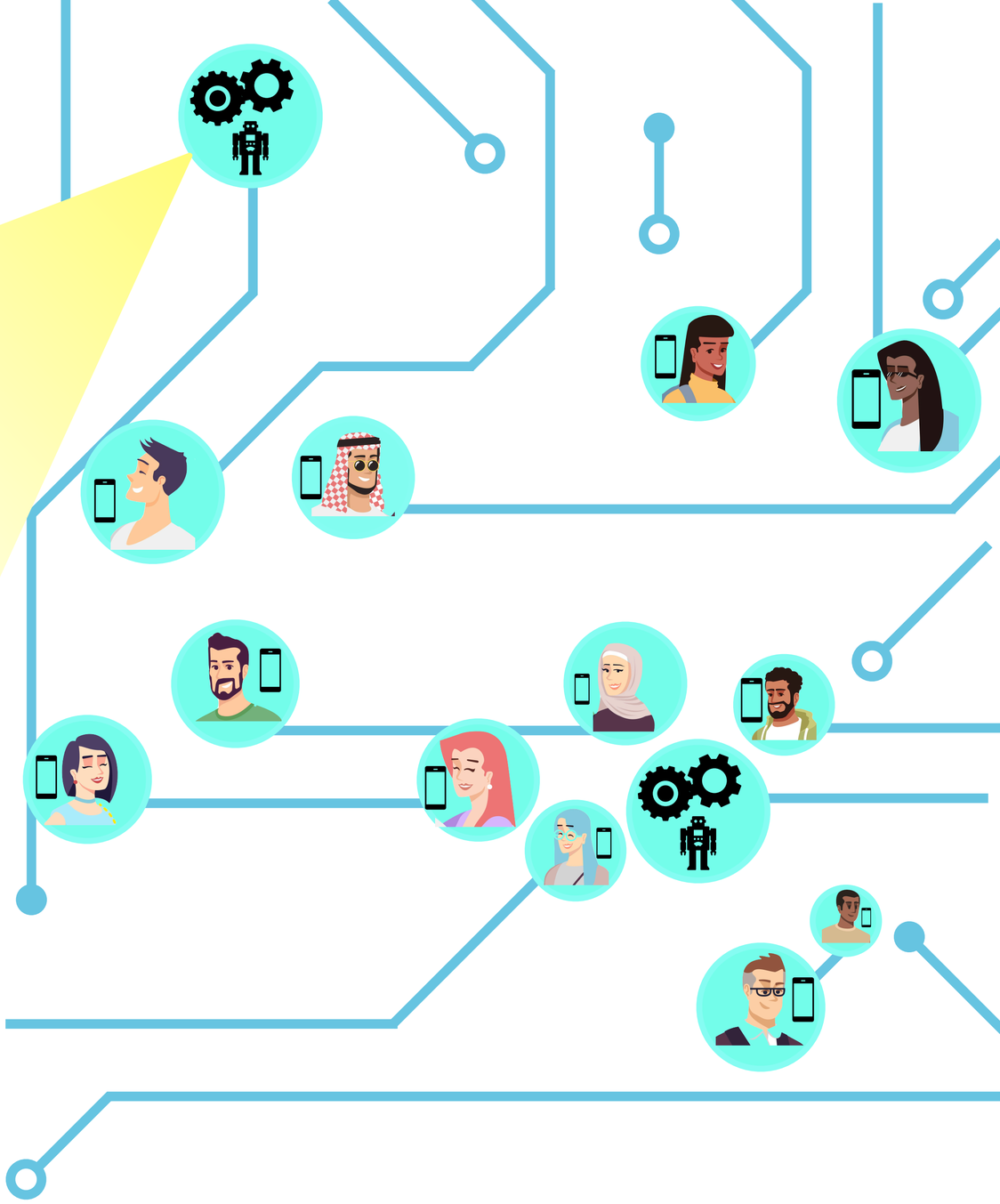
Rediger

Vedlegg fra utgiftsposter

Faglitteratur/bøker (44.85 GBP) - 1
pdf

Digital society

**devices that can do some
things by themselves and the humans that use them**



Utgiftsrefusjon

Utgifter

Rediger

Type	Vedlegg	Beskrivelse	Beløp	Kurs	Beløp NOK
Faglitteratur/bøker	1	Hands-On Explainable AI (XAI)	44,85 GBP	11,629	521,55
Totalt					**521,55**

Order date	09.10.2021
Order #	026-3107586-1263557

Invoice details

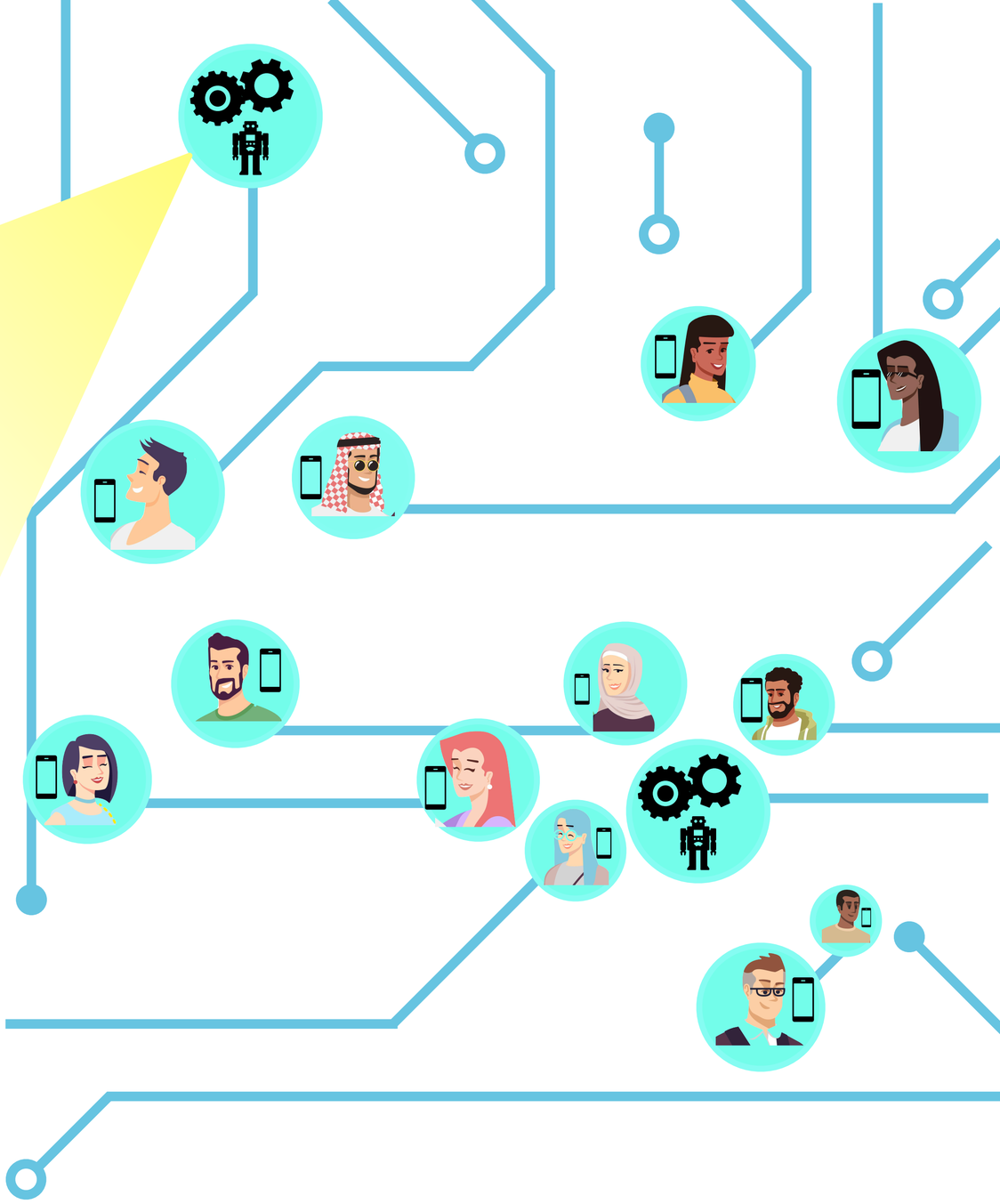
Description	Qty	Unit price (excl. VAT)	VAT rate	Unit price (incl. VAT)	Item subtotal (incl. VAT)
Hands-On Explainable AI (XAI) with Python: Interpret, visualize, explain, and integrate reliable AI for fair, secure, and trustworthy AI apps | 1800208138 ASIN: 1800208138	1	£37.99	0%	£37.99	£37.99
Shipping Charges		£6.86		£6.86	£6.86

Invoice total £44.85

VAT rate	Item subtotal (excl. VAT)	VAT subtotal

Digital society

**devices that can do some
things by themselves and the humans that use them**



Utgiftsrefusjon

Utgifter

Rediger

Type	Vedlegg	Beskrivelse	Beløp	Kurs	Beløp NOK
Faglitteratur/bøker	1	Hands-On Explainable AI (XAI)	44,85 GBP	11,629	521,55
Totalt					**521,55**

Order date 09.10.2021
Order # 026-3107586-1263557

Invoice details

Description	Qty	Unit price (excl. VAT)	VAT rate	Unit price (incl. VAT)	Item subtotal (incl. VAT)
Hands-On Explainable AI (XAI) with Python: Interpret, visualize, explain, and integrate reliable AI for fair, secure, and trustworthy AI apps | 1800208138 ASIN: 1800208138	1	£37.99	0%	£37.99	£37.99
Shipping Charges		£6.86		£6.86	£6.86

Invoice total £44.85

VAT rate	Item subtotal (excl. VAT)	

Digital society

TABLE I
LEVELS OF AUTOMATION OF DECISION AND ACTION SELECTION

HIGH	10. The computer decides everything, acts autonomously, ignoring the human.
	9. informs the human only if it, the computer, decides to
	8. informs the human only if asked, or
	7. executes automatically, then necessarily informs the human, and
	6. allows the human a restricted time to veto before automatic execution, or
	5. executes that suggestion if the human approves, or
	4. suggests one alternative
	3. narrows the selection down to a few, or
	2. The computer offers a complete set of decision/action alternatives, or
LOW	1. The computer offers no assistance: human must take all decisions and actions.

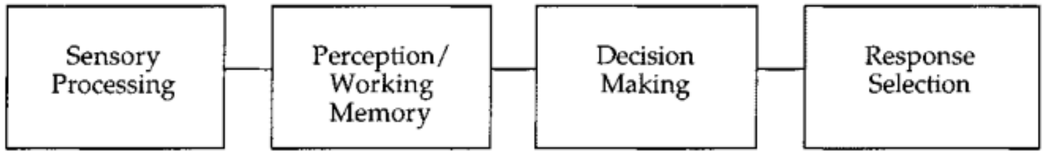


Fig. 1. Simple four-stage model of human information processing.

A Model for Types and Levels of Human Interaction with Automation

Raja Parasuraman, Thomas B. Sheridan, *Fellow, IEEE*, and Christopher D. Wickens

What happens when we automate decisions?

What happens when we automate decisions?

- Rational decision-making is a process than consists of at least four steps:

What happens when we automate decisions?

- Rational decision-making is a process than consists of at least four steps:

  1. identify the problem for which a decision needs to be made,

What happens when we automate decisions?

- Rational decision-making is a process than consists of at least four steps:

  1. identify the problem for which a decision needs to be made,

  2. evaluate the objectives and preferences that apply,

What happens when we automate decisions?

- Rational decision-making is a process than consists of at least four steps:

  1. identify the problem for which a decision needs to be made,

  2. evaluate the objectives and preferences that apply,

  3. analyze the decision problem and its constraints, and develop or identify the possible options from which to choose,

What happens when we automate decisions?

- Rational decision-making is a process than consists of at least four steps:

  1. identify the problem for which a decision needs to be made,

  2. evaluate the objectives and preferences that apply,

  3. analyze the decision problem and its constraints, and develop or identify the possible options from which to choose,

  4. choose from the identified options following some reasoning.

What happens when we automate decisions?

- Rational decision-making is a process than consists of at least four steps:

  1. identify the problem for which a decision needs to be made,

  2. evaluate the objectives and preferences that apply,

  3. analyze the decision problem and its constraints, and develop or identify the possible options from which to choose,

  4. choose from the identified options following some reasoning.

Ethical decisions

Ethical decisions

- A decision is ethical if it is made not only on the factual identified objectives, preferences and constraints, but also based on a person's or societies consideration of what is right or wrong.

Ethical decisions

- A decision is ethical if it is made not only on the factual identified objectives, preferences and constraints, but also based on a person's or societies consideration of what is right or wrong.
- Ethical decisions include considering “the interests of others as of equal weight with ones own”

Ethical decisions

- A decision is ethical if it is made not only on the factual identified objectives, preferences and constraints, but also based on a person's or societies consideration of what is right or wrong.
- Ethical decisions include considering “the interests of others as of equal weight with ones own”
- In ethics, value is a way to indicate the degree of importance of some thing or action, with the aim of determining what actions are best to do or which states of the world are best to be achieved.

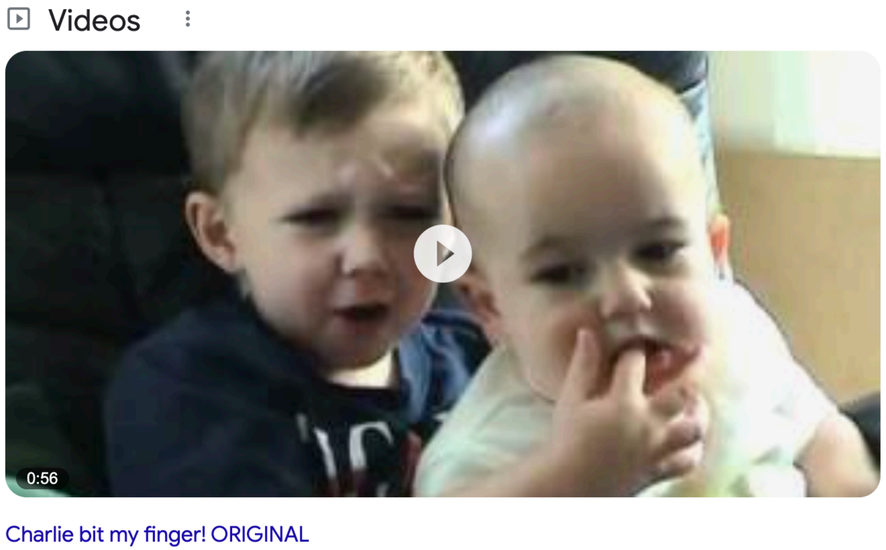
AI, agent, morality

AI, agent, morality

- AI does not make decisions, an agent makes decisions!

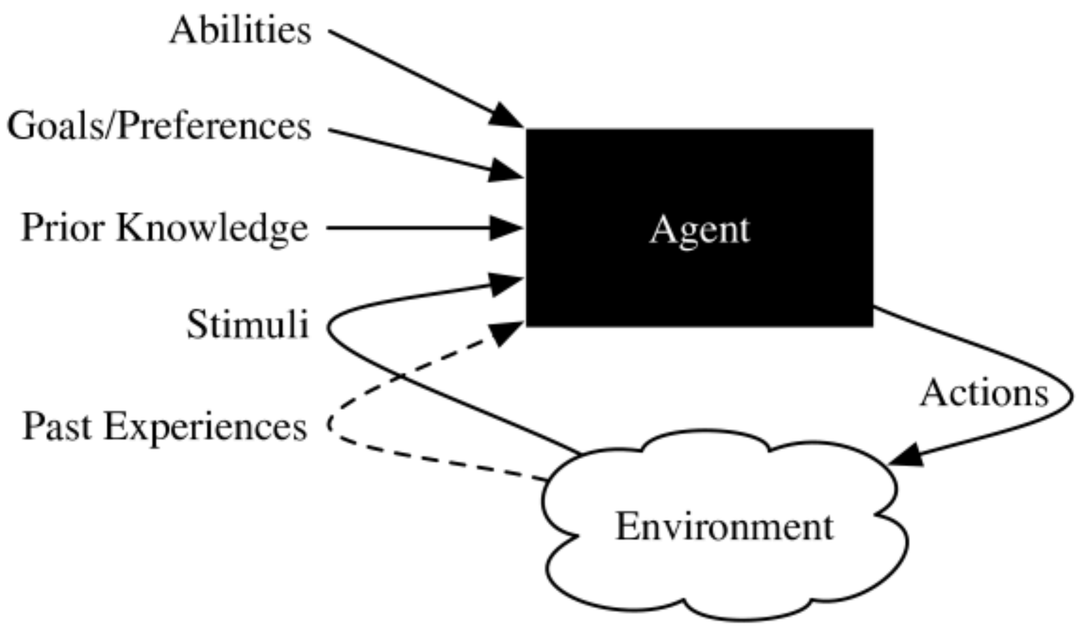
AI, agent, morality

- AI does not make decisions, an agent makes decisions!

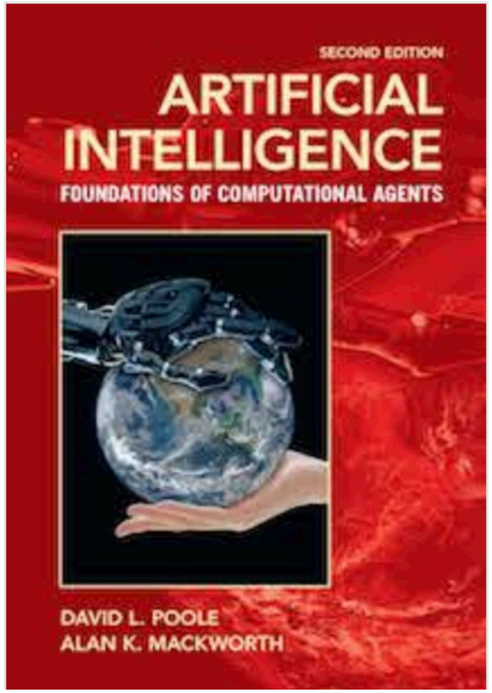


What is an agent?

An entity that acts in an environment.



An agent as an input–output system



Natural vs artificial agents

Natural vs artificial agents

- In nature, we tend to call an agent something that is able to do all the tasks associated with agency

Natural vs artificial agents

- In nature, we tend to call an agent something that is able to do all the tasks associated with agency
- There are different type of artificial agents: reactive, intelligent, embodied

Natural vs artificial agents

- In nature, we tend to call an agent something that is able to do all the tasks associated with agency
- There are different type of artificial agents: reactive, intelligent, embodied
- Example: playing chess against a computer

Natural vs artificial agents

- In nature, we tend to call an agent something that is able to do all the tasks associated with agency
- There are different type of artificial agents: reactive, intelligent, embodied
- Example: playing chess against a computer

Deep Blue

One of the two cabinets of Deep Blue in its exhibit at the Computer History Museum, California

Natural vs artificial agents

- In nature, we tend to call an agent something that is able to do all the tasks associated with agency
- There are different type of artificial agents: reactive, intelligent, embodied
- Example: playing chess against a computer

Chess

Game

Chess is a board game played between two players. It is sometimes called Western chess or international chess to distinguish it from related games such as xiangqi and shogi. Wikipedia

Genres: Board game; Abstract strategy game; Mind sport

Players: 2

Skills required: Strategy, tactics

Deep Blue

One of the two cabinets of Deep Blue in its exhibit at the Computer History Museum, California

Can an artificial agent be a moral agent

Can an artificial agent be a moral agent

- Let us assume yes, but do consult a philosopher

Can an artificial agent be a moral agent

- Let us assume yes, but do consult a philosopher
- Joke: what is the difference between a scientist and an engineer?

Artificial moral agents

Artificial moral agents

Browse Journals & Magazines > IEEE Intelligent Systems > Volume: 21 Issue: 4

The Nature, Importance, and Difficulty of Machine Ethics

1 Author(s) J.H. Moor View All Authors

69 Paper Citations

3005 Full Text Views

Abstract

Document Sections

1. Varieties of Machine Ethics
2. Ethical-Impact Agents
3. Implicit Ethical Agents
4. Explicit Ethical Agents
5. Full Ethical Agents

Authors

Abstract:
The question of whether machine ethics exists or might exist in the future is difficult to answer if we can't agree on what counts as machine ethics. Some might argue that machine ethics obviously exists because humans are machines and humans have ethics. Others could argue that machine ethics obviously doesn't exist because ethics is simply emotional expression and machines can't have emotions. A wide range of positions on machine ethics are possible, and a discussion of the issue could rapidly propel us into deep and unsettled philosophical issues. Perhaps, understandably, few in the scientific arena pursue the issue of machine ethics. As we expand computers' decision-making roles in practical matters, such as computers driving cars, ethical considerations are inevitable. Computer scientists and engineers must examine the possibilities for machine ethics because, knowingly or not, they've already engaged in some form of it. Before we can discuss possible implementations of machine ethics, however, we need to be clear about what we're asserting or denying

Published in: IEEE Intelligent Systems (Volume: 21 , Issue: 4 , July-Aug. 2006)

Page(s): 18 - 21

INSPEC Accession Number: 9065956

Date of Publication: 07 August 2006

DOI: 10.1109/MIS.2006.80

Artificial moral agents

Browse Journals & Magazines > IEEE Intelligent Systems > Volume: 21 Issue: 4

The Nature, Importance, and Difficulty of Machine Ethics

1 Author(s) J.H. Moor View All Authors

Abstract:
The question of whether machine ethics exists or might exist in the future is difficult to answer if we can't agree on what counts as machine ethics. Some might argue that machine ethics obviously exists because humans are machines and humans have ethics. Others could argue that machine ethics obviously doesn't exist because ethics is simply emotional expression and machines can't have emotions. A wide range of positions on machine ethics are possible, and a discussion of the issue could rapidly propel us into deep and unsettled philosophical issues. Perhaps, understandably, few in the scientific arena pursue the issue of machine ethics. As we expand computers' decision-making roles in practical matters, such as computers driving cars, ethical considerations are inevitable. Computer scientists and engineers must examine the possibilities for machine ethics because, knowingly or not, they've already engaged in some form of it. Before we can discuss possible implementations of machine ethics, however, we need to be clear about what we're asserting or denying

Published in: IEEE Intelligent Systems (Volume: 21 , Issue: 4 , July-Aug. 2006)

Page(s): 18 - 21

INSPEC Accession Number: 9065956

Date of Publication: 07 August 2006

DOI: 10.1109/MIS.2006.80

Artificial moral agents

- Ethical-impact AMA

The Nature, Importance, and Difficulty of Machine Ethics

1 Author(s) J.H. Moor View All Authors

Abstract:
The question of whether machine ethics exists or might exist in the future is difficult to answer if we can't agree on what counts as machine ethics. Some might argue that machine ethics obviously exists because humans are machines and humans have ethics. Others could argue that machine ethics obviously doesn't exist because ethics is simply emotional expression and machines can't have emotions. A wide range of positions on machine ethics are possible, and a discussion of the issue could rapidly propel us into deep and unsettled philosophical issues. Perhaps, understandably, few in the scientific arena pursue the issue of machine ethics. As we expand computers' decision-making roles in practical matters, such as computers driving cars, ethical considerations are inevitable. Computer scientists and engineers must examine the possibilities for machine ethics because, knowingly or not, they've already engaged in some form of it. Before we can discuss possible implementations of machine ethics, however, we need to be clear about what we're asserting or denying

Published in: IEEE Intelligent Systems (Volume: 21 , Issue: 4 , July-Aug. 2006)

Page(s): 18 - 21

INSPEC Accession Number: 9065956

Date of Publication: 07 August 2006

DOI: 10.1109/MIS.2006.80

Artificial moral agents

- Ethical-impact AMA
- Implicit ethical AMA

Browse Journals & Magazines > IEEE Intelligent Systems > Volume: 21 Issue: 4

The Nature, Importance, and Difficulty of Machine Ethics

1 Author(s) J.H. Moor View All Authors

69 Paper Citations

3005 Full Text Views

Abstract

Document Sections

1. Varieties of Machine Ethics
2. Ethical-Impact Agents
3. Implicit Ethical Agents
4. Explicit Ethical Agents
5. Full Ethical Agents

Authors

Abstract:
The question of whether machine ethics exists or might exist in the future is difficult to answer if we can't agree on what counts as machine ethics. Some might argue that machine ethics obviously exists because humans are machines and humans have ethics. Others could argue that machine ethics obviously doesn't exist because ethics is simply emotional expression and machines can't have emotions. A wide range of positions on machine ethics are possible, and a discussion of the issue could rapidly propel us into deep and unsettled philosophical issues. Perhaps, understandably, few in the scientific arena pursue the issue of machine ethics. As we expand computers' decision-making roles in practical matters, such as computers driving cars, ethical considerations are inevitable. Computer scientists and engineers must examine the possibilities for machine ethics because, knowingly or not, they've already engaged in some form of it. Before we can discuss possible implementations of machine ethics, however, we need to be clear about what we're asserting or denying

Published in: IEEE Intelligent Systems (Volume: 21 , Issue: 4 , July-Aug. 2006)

Page(s): 18 - 21

INSPEC Accession Number: 9065956

Date of Publication: 07 August 2006

DOI: 10.1109/MIS.2006.80

Artificial moral agents

- Ethical-impact AMA
- Implicit ethical AMA
- Explicit ethical AMA

Browse Journals & Magazines > IEEE Intelligent Systems > Volume: 21 Issue: 4

The Nature, Importance, and Difficulty of Machine Ethics

1 Author(s) J.H. Moor View All Authors

69 Paper Citations

3005 Full Text Views

Abstract

Document Sections

1. Varieties of Machine Ethics
2. Ethical-Impact Agents
3. Implicit Ethical Agents
4. Explicit Ethical Agents
5. Full Ethical Agents

Authors

Abstract:
The question of whether machine ethics exists or might exist in the future is difficult to answer if we can't agree on what counts as machine ethics. Some might argue that machine ethics obviously exists because humans are machines and humans have ethics. Others could argue that machine ethics obviously doesn't exist because ethics is simply emotional expression and machines can't have emotions. A wide range of positions on machine ethics are possible, and a discussion of the issue could rapidly propel us into deep and unsettled philosophical issues. Perhaps, understandably, few in the scientific arena pursue the issue of machine ethics. As we expand computers' decision-making roles in practical matters, such as computers driving cars, ethical considerations are inevitable. Computer scientists and engineers must examine the possibilities for machine ethics because, knowingly or not, they've already engaged in some form of it. Before we can discuss possible implementations of machine ethics, however, we need to be clear about what we're asserting or denying

Published in: IEEE Intelligent Systems (Volume: 21 , Issue: 4 , July-Aug. 2006)

Page(s): 18 - 21

INSPEC Accession Number: 9065956

Date of Publication: 07 August 2006

DOI: 10.1109/MIS.2006.80

Artificial moral agents

- Ethical-impact AMA
- Implicit ethical AMA
- Explicit ethical AMA
- Fully ethical agents

Browse Journals & Magazines > IEEE Intelligent Systems > Volume: 21 Issue: 4

The Nature, Importance, and Difficulty of Machine Ethics

1 Author(s) J.H. Moor View All Authors

Abstract:
The question of whether machine ethics exists or might exist in the future is difficult to answer if we can't agree on what counts as machine ethics. Some might argue that machine ethics obviously exists because humans are machines and humans have ethics. Others could argue that machine ethics obviously doesn't exist because ethics is simply emotional expression and machines can't have emotions. A wide range of positions on machine ethics are possible, and a discussion of the issue could rapidly propel us into deep and unsettled philosophical issues. Perhaps, understandably, few in the scientific arena pursue the issue of machine ethics. As we expand computers' decision-making roles in practical matters, such as computers driving cars, ethical considerations are inevitable. Computer scientists and engineers must examine the possibilities for machine ethics because, knowingly or not, they've already engaged in some form of it. Before we can discuss possible implementations of machine ethics, however, we need to be clear about what we're asserting or denying

Published in: IEEE Intelligent Systems (Volume: 21 , Issue: 4 , July-Aug. 2006)

Page(s): 18 - 21

INSPEC Accession Number: 9065956

Date of Publication: 07 August 2006

DOI: 10.1109/MIS.2006.80

Ethical impact artificial moral agents

change the balance of good-bad in society

Ethical impact artificial moral agents

change the balance of good-bad in society

- can operate autonomously,

Ethical impact artificial moral agents

change the balance of good-bad in society

- can operate autonomously,
- can impact society and individuals,

Ethical impact artificial moral agents

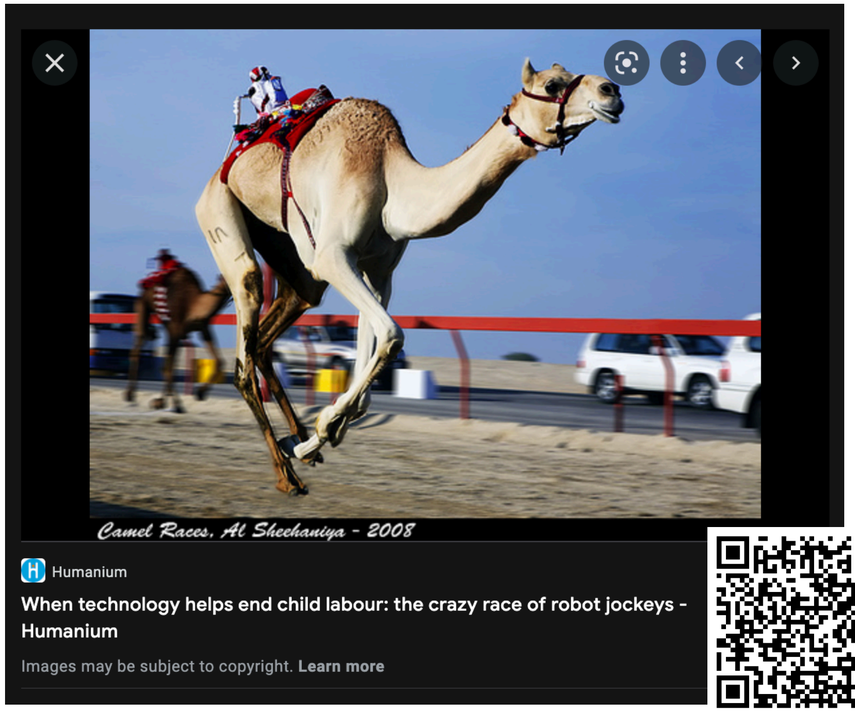
change the balance of good-bad in society

- can operate autonomously,
- can impact society and individuals,
- does not reason ethically (cannot/does not need to decide between right and wrong)

Ethical impact artificial moral agents

change the balance of good-bad in society

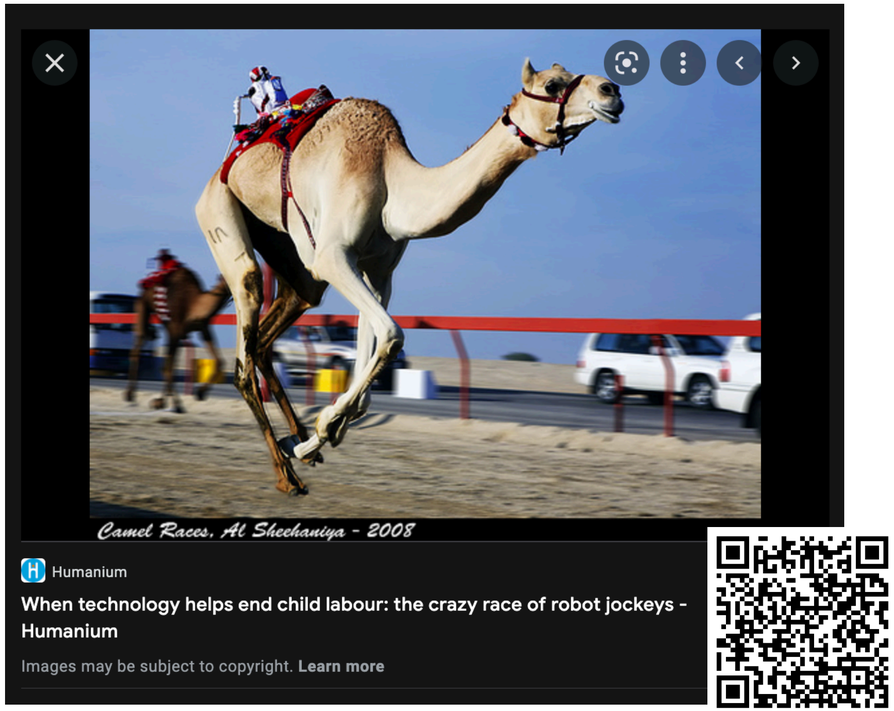
- can operate autonomously,
- can impact society and individuals,
- does not reason ethically (cannot/does not need to decide between right and wrong)



Ethical impact artificial moral agents

change the balance of good-bad in society

- can operate autonomously,
- can impact society and individuals,
- does not reason ethically (cannot/does not need to decide between right and wrong)



Ethical impact artificial moral agents

change the balance of good-bad in society

Ethical impact artificial moral agents

change the balance of good-bad in society

Ethical impact artificial moral agents

change the balance of good-bad in society

- How to establish the right relationship between a society forum and AI developer actors.

Ethical impact artificial moral agents

change the balance of good-bad in society

- How to establish the right relationship between a society forum and AI developer actors.

Ethical impact artificial moral agents

change the balance of good-bad in society

- How to establish the right relationship between a society forum and AI developer actors.
- Ensure the right type of information about the working of the AI agent is provided to a given stakeholder

Ethical impact artificial moral agents

change the balance of good-bad in society

- How to establish the right relationship between a society forum and AI developer actors.
- Ensure the right type of information about the working of the AI agent is provided to a given stakeholder

Ethical impact artificial moral agents

change the balance of good-bad in society

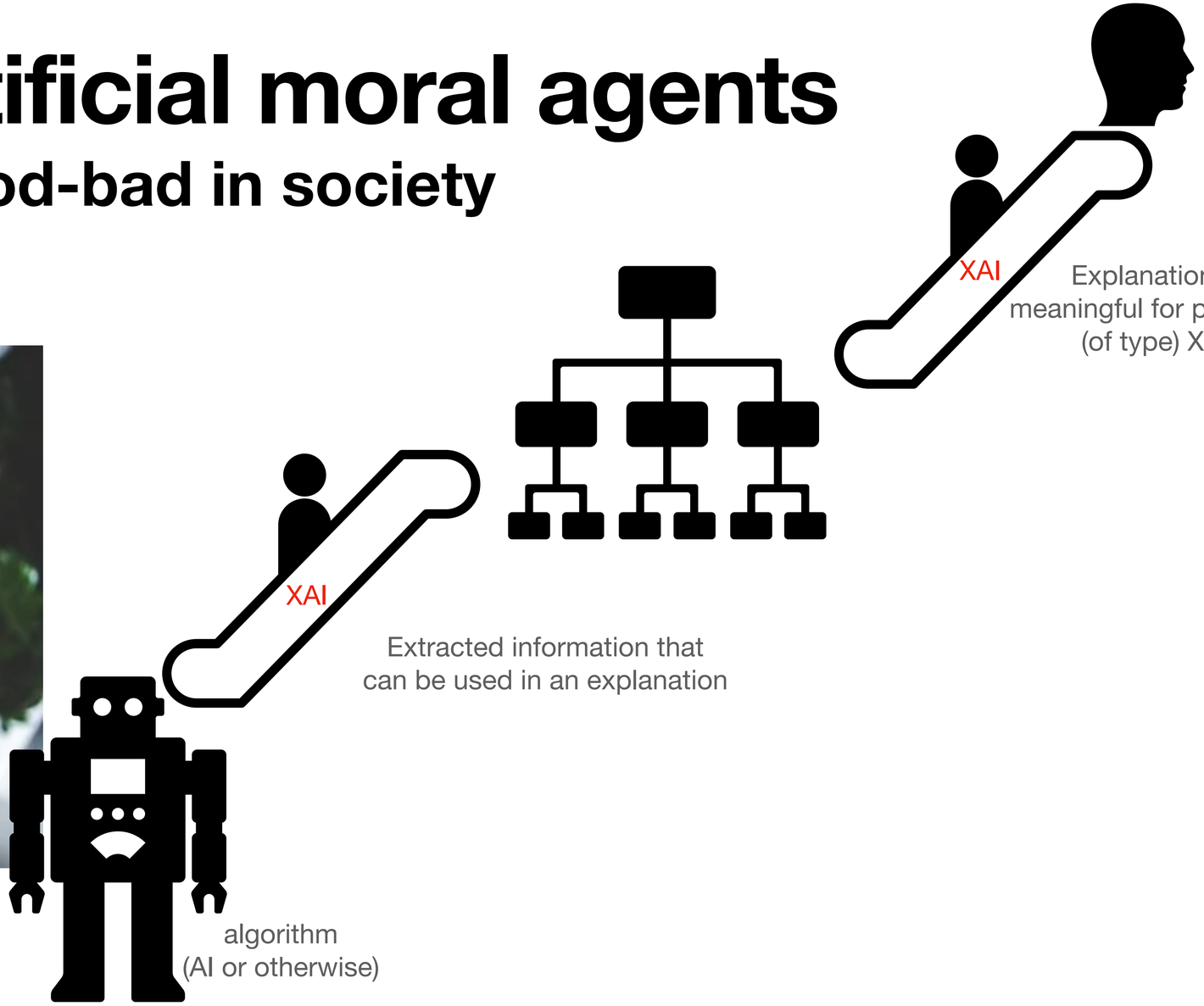
- How to establish the right relationship between a society forum and AI developer actors.
- Ensure the right type of information about the working of the AI agent is provided to a given stakeholder

Ethical impact artificial moral agents

change the balance of good-bad in society

Ethical impact artificial moral agents

change the balance of good-bad in society



Ethical impact artificial moral agents

change the balance of good-bad in society

Ethical impact artificial moral agents

change the balance of good-bad in society

Ethical impact artificial moral agents

change the balance of good-bad in society

- Ensuring that individuals and groups are treated alike by decision making algorithms and in representative data

Ethical impact artificial moral agents

change the balance of good-bad in society

- Ensuring that individuals and groups are treated alike by decision making algorithms and in representative data

Ethical impact artificial moral agents

change the balance of good-bad in society

- Ensuring that individuals and groups are treated alike by decision making algorithms and in representative data

- Privacy is a space in which a person can be/ act without unsolicited scrutiny

Implicit vs explicit artificial agents

Implicit vs explicit artificial agents

Implicit vs explicit artificial agents

- Implicit ethical agents

Implicit vs explicit artificial agents

- Implicit ethical agents 

Implicit vs explicit artificial agents

- Implicit ethical agents 
- Explicit ethical agents

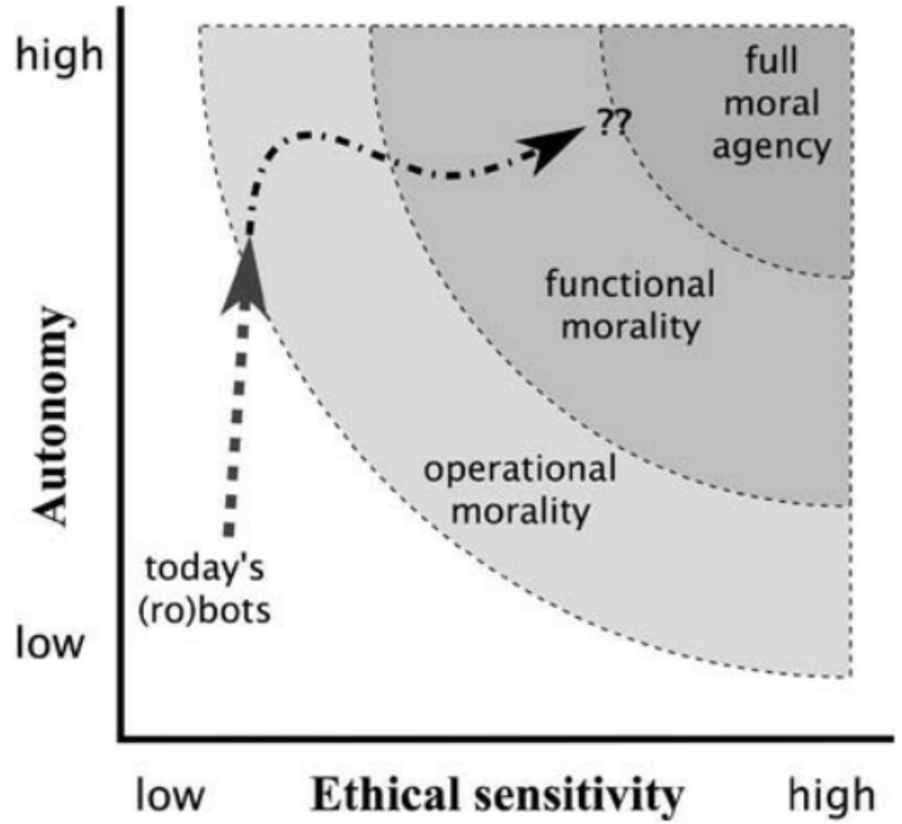
Implicit vs explicit artificial agents

- Implicit ethical agents 
- Explicit ethical agents 

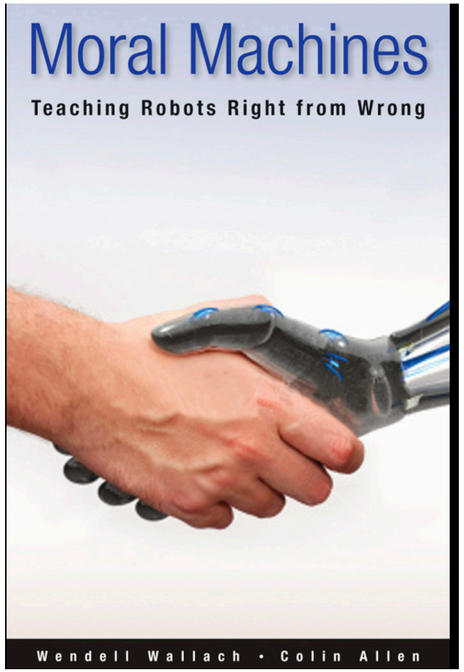
Implicit vs explicit artificial agents

- Implicit ethical agents

- Explicit ethical agents

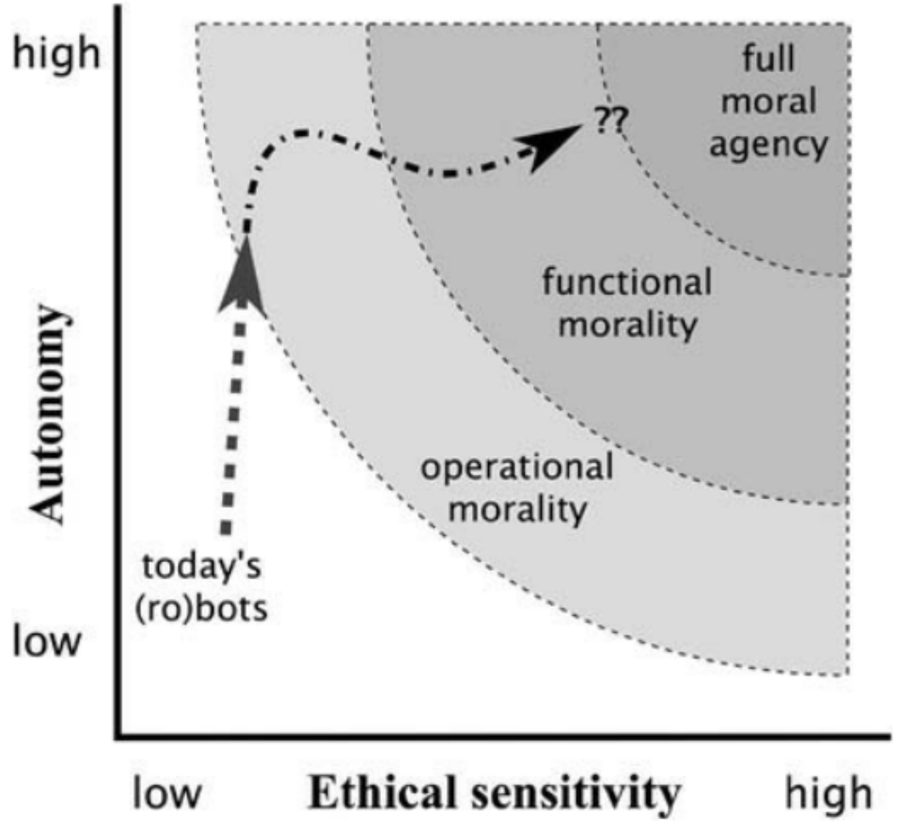


Implicit vs explicit artificial agents



- Implicit ethical agents

- Explicit ethical agents



What to choose

What to choose

What to choose

- Implicit ethical agents: predictable environment and choices

What to choose

- Implicit ethical agents: predictable environment and choices

What to choose

- Implicit ethical agents: predictable environment and choices
- Explicit ethical agents: unpredictable environment or choices

What to choose

- Implicit ethical agents: predictable environment and choices
- Explicit ethical agents: unpredictable environment or choices

It is the AI way

Simplify the environment

Simplify environments and build complex reasoning systems for these simple environments. For example, factory robots can do sophisticated tasks in the engineered environment of a factory, but they may be hopeless in a natural environment. Much of the complexity of the task can be reduced by simplifying the environment.

Simplify the agent

Simple agents in natural environments. This is inspired by seeing how **insects** can survive in complex environments even though they have very limited reasoning abilities. Researchers then make the agents have more reasoning abilities as their tasks become more complicated. Example:

It is the AI way

Simplify the environment

Simplify environments and build complex reasoning systems for these simple environments. For example, factory robots can do sophisticated tasks in the engineered environment of a factory, but they may be hopeless in a natural environment. Much of the complexity of the task can be reduced by simplifying the environment.

Simplify the agent

Simple agents in natural environments. This is inspired by seeing how **insects** can survive in complex environments even though they have very limited reasoning abilities. Researchers then make the agents have more reasoning abilities as their tasks become more complicated. Example:

It is the AI way

Simplify the environment

Simplify environments and build complex reasoning systems for these sim
factory robots can do sophisticated tasks in the engineered environment o
hopeless in a natural environment. Much of the complexity of the task can
environment.

Robotstøvsuger som ruller og samler støv

Mocoro Mop Ball | Art. nr.: 102680

Simplify the agent

Simple agents in natural environments. This is inspired by seeing how **ins**
environments even though they have very limited reasoning abilities. Researc
reasoning abilities as their tasks become more complicated. Example:

So.. you want to make an agent that does good and not evil?

what is good?

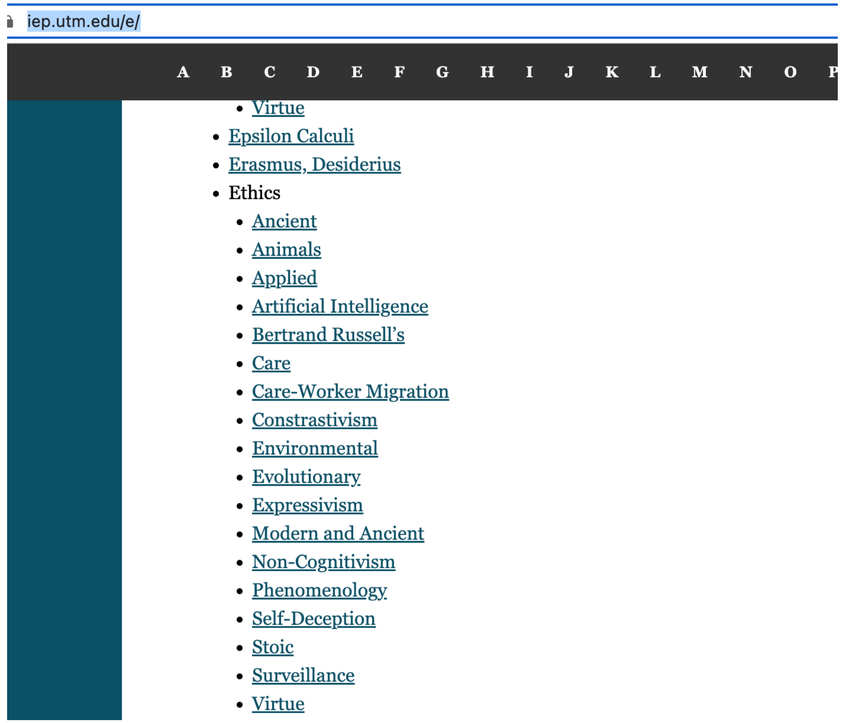
What is ethics?

What is ethics?

- **Ethics** or **moral philosophy** is a branch[1] of philosophy that "involves systematizing, defending, and recommending concepts of right and wrong behavior".

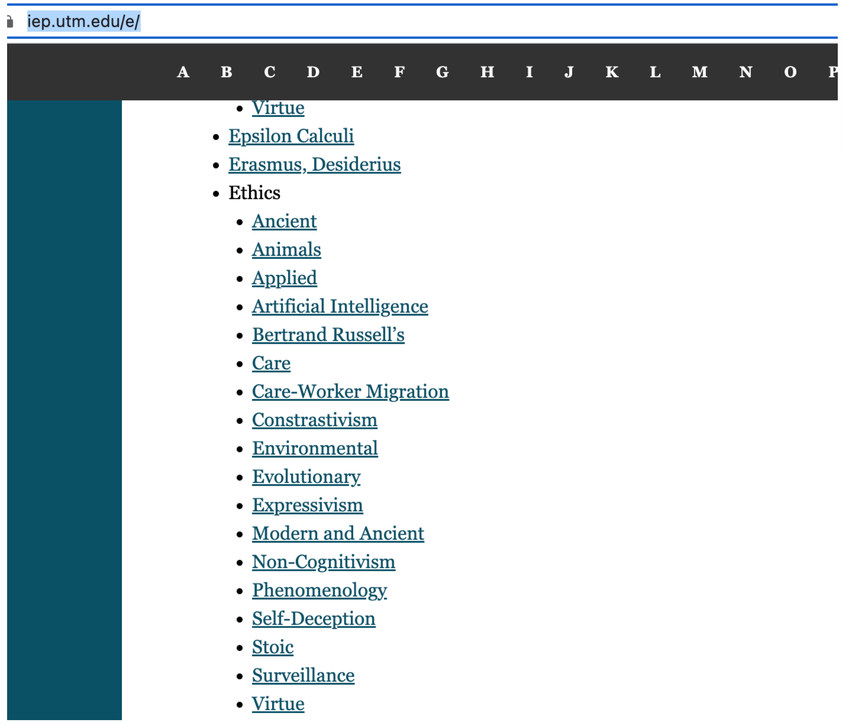
What is ethics?

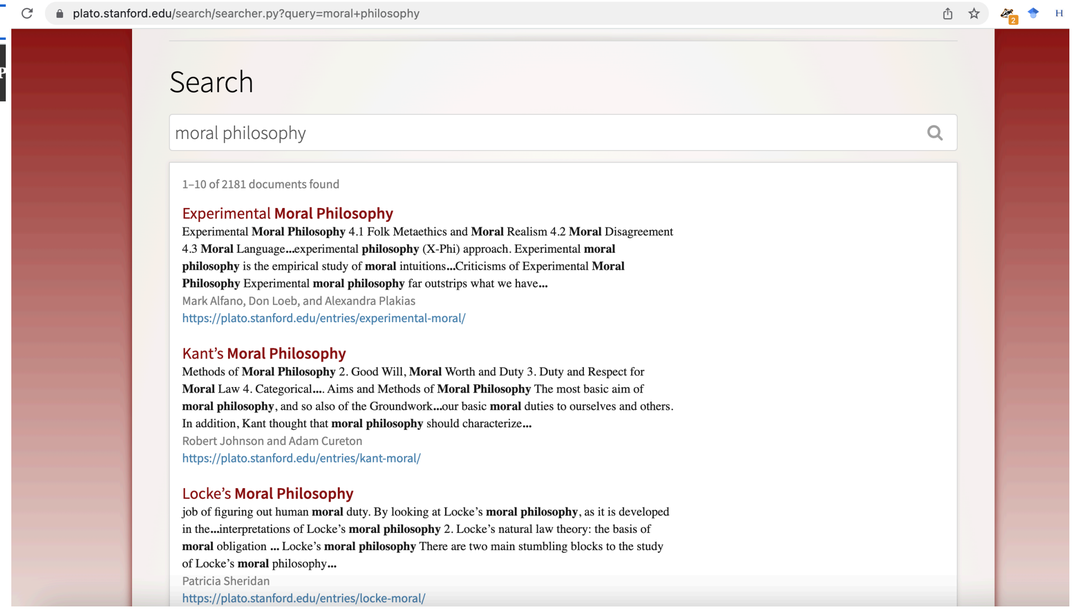
- **Ethics** or **moral philosophy** is a branch[1] of philosophy that "involves systematizing, defending, and recommending concepts of right and wrong behavior".



What is ethics?

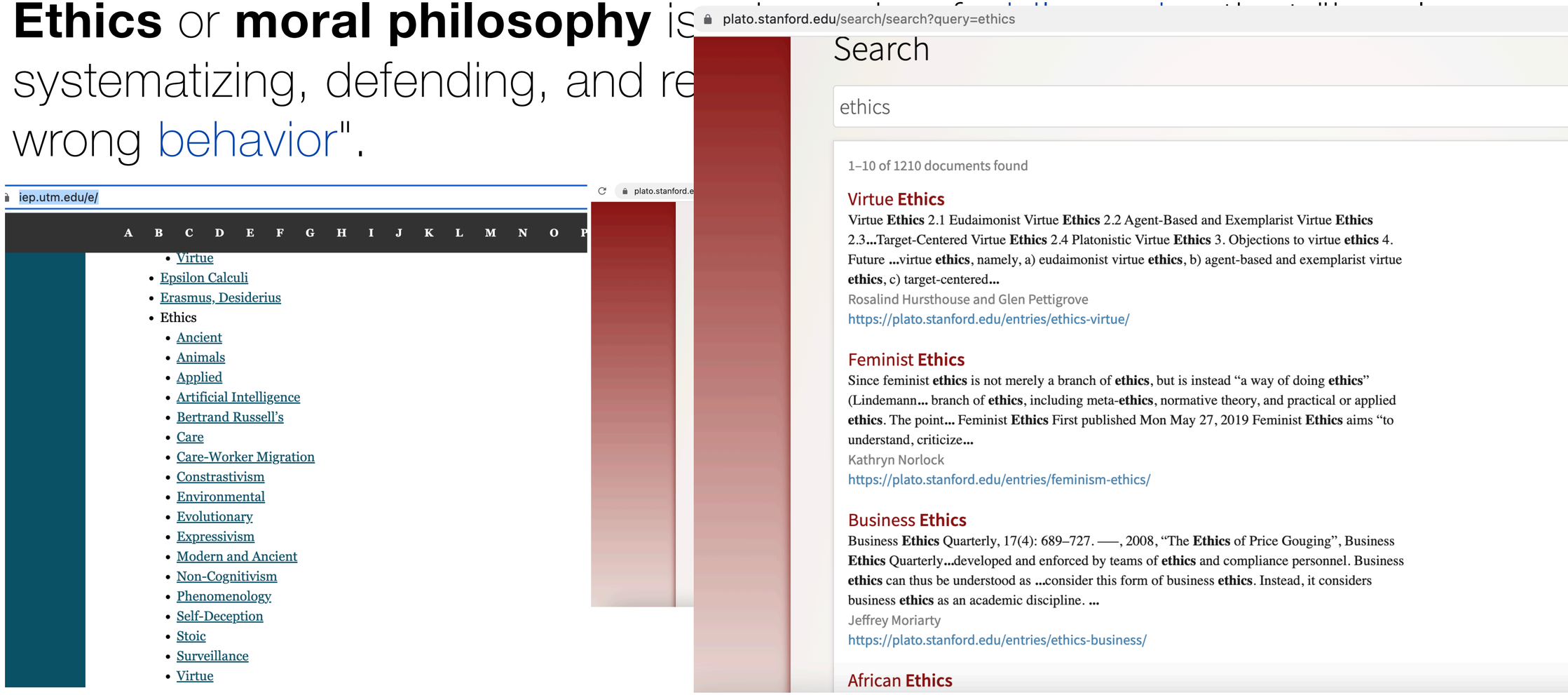
- **Ethics** or **moral philosophy** is a branch[1] of philosophy that "involves systematizing, defending, and recommending concepts of right and wrong behavior".





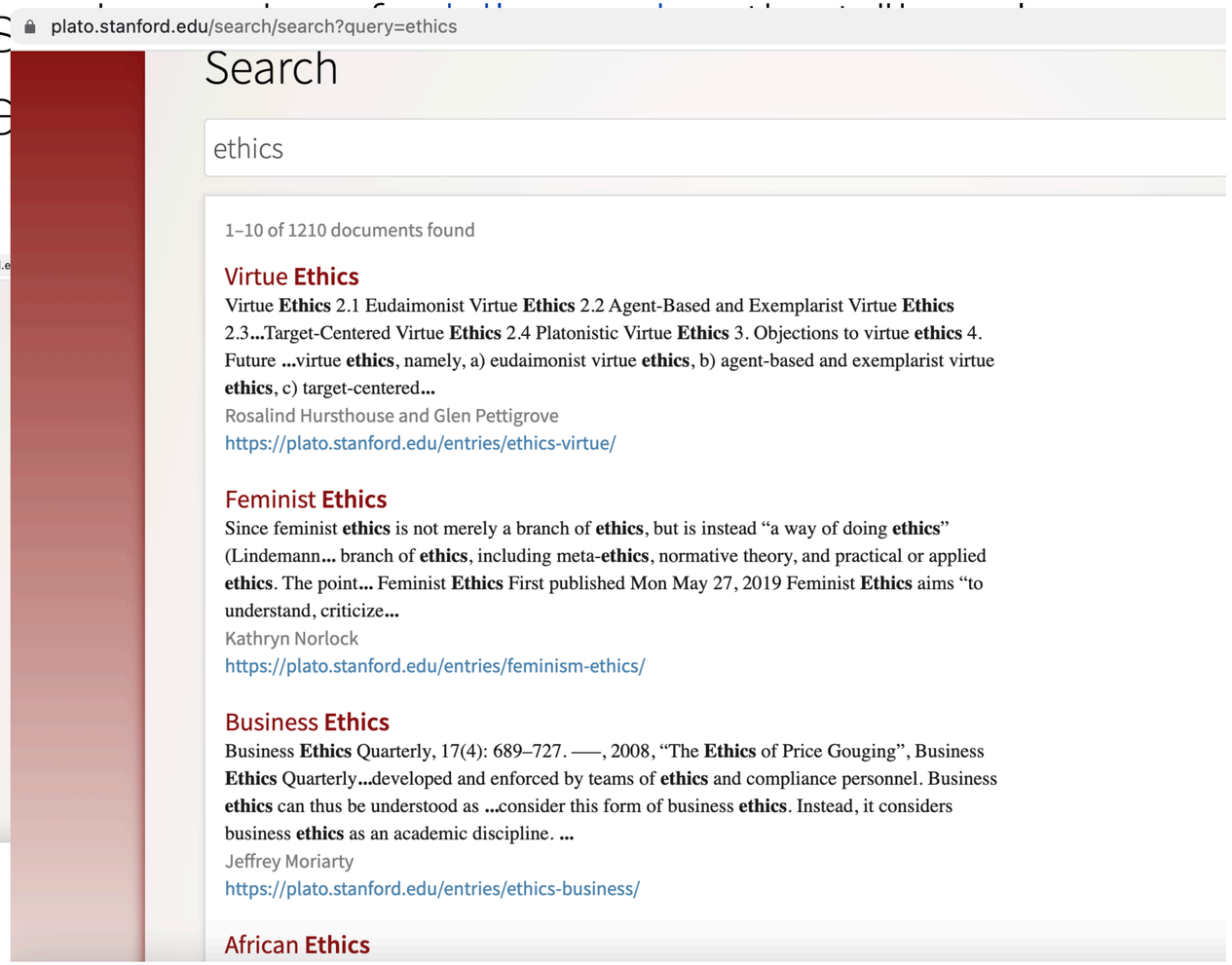
What is ethics?

- **Ethics** or **moral philosophy** is
  systematizing, defending, and re
  wrong behavior".

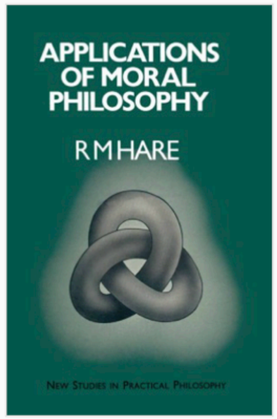


What is ethics?

- **Ethics** or **moral philosophy** is
  systematizing, defending, and re
  wrong behavior".



What is morality?



Applications of Moral Philosophy

Authors (view affiliations)

R. M. Hare

Textbook

18 Citations

71 Downloads

Part of the New Studies in Practical Philosophy book series

the essence of morality is to treat the interests of others as of equal weight with ones own

What does moral philosophy do?

What does moral philosophy do?

- **Ethics** or **moral philosophy** is a branch[1] of philosophy that "involves systematizing, defending, and recommending concepts of right and wrong behavior".

What does moral philosophy do?

- **Ethics** or **moral philosophy** is a branch[1] of philosophy that "involves systematizing, defending, and recommending concepts of right and wrong behavior".
- **Meta-ethics**: concerned with the concepts of right and wrong themselves

What does moral philosophy do?

- **Ethics** or **moral philosophy** is a branch[1] of philosophy that "involves systematizing, defending, and recommending concepts of right and wrong behavior".
- **Meta-ethics**: concerned with the concepts of right and wrong themselves
- **Normative ethics**: develop means to identify what are the right and wrong decisions, actions, states of the world etc.

What does moral philosophy do?

- **Ethics** or **moral philosophy** is a branch[1] of philosophy that "involves systematizing, defending, and recommending concepts of right and wrong behavior".
- **Meta-ethics**: concerned with the concepts of right and wrong themselves
- **Normative ethics**: develop means to identify what are the right and wrong decisions, actions, states of the world etc.
- **Applied ethics**: issue recommendations to professionals on what is the right thing to do by a given person role, in a given situation.

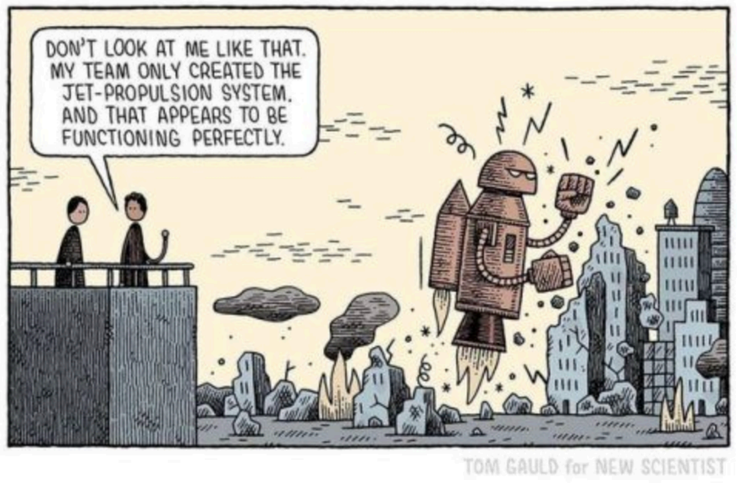
Normative ethics

Normative ethics

- **Moral theory** is an explanation of what makes action right, a state good, or an agent good.

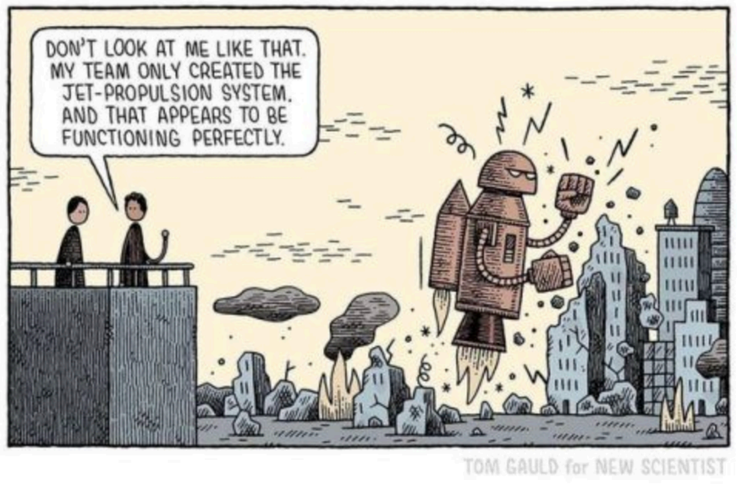
Normative ethics

- **Moral theory** is an explanation of what makes action right, a state good, or an agent good.



Normative ethics

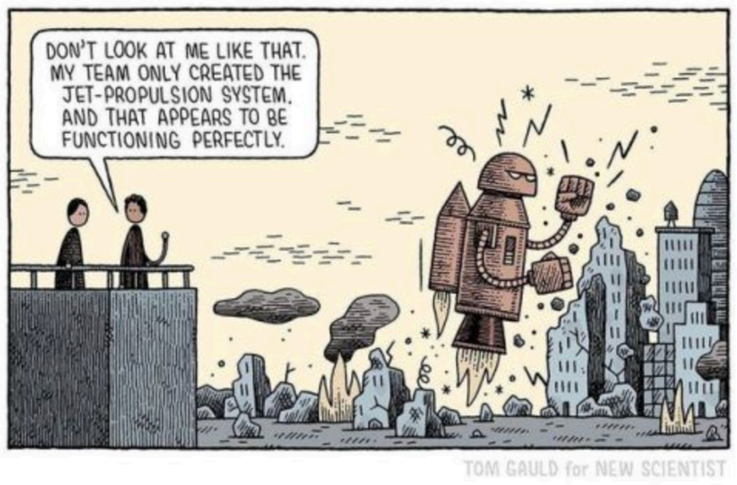
- **Moral theory** is an explanation of what makes action right, a state good, or an agent good.



An **agent** makes a **choice**. A **choice** has **consequences**.

Normative ethics

- **Moral theory** is an explanation of what makes action right, a state good, or an agent good.

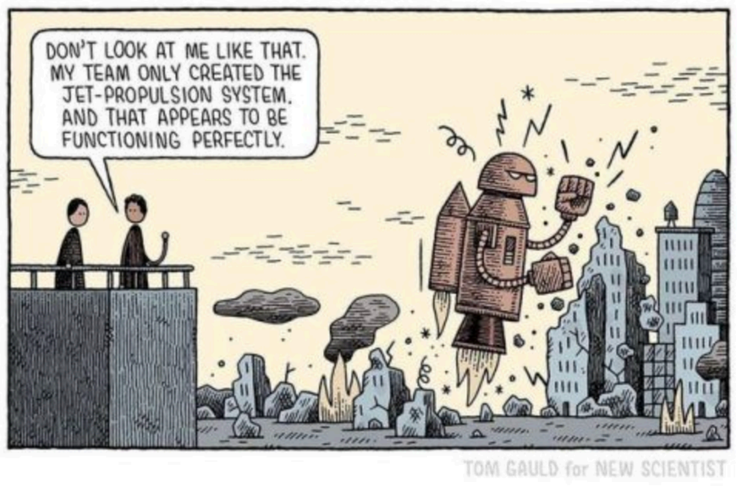


An **agent** makes a **choice**. A **choice** has **consequences**.

Virtue theories

Normative ethics

- **Moral theory** is an explanation of what makes action right, a state good, or an agent good.



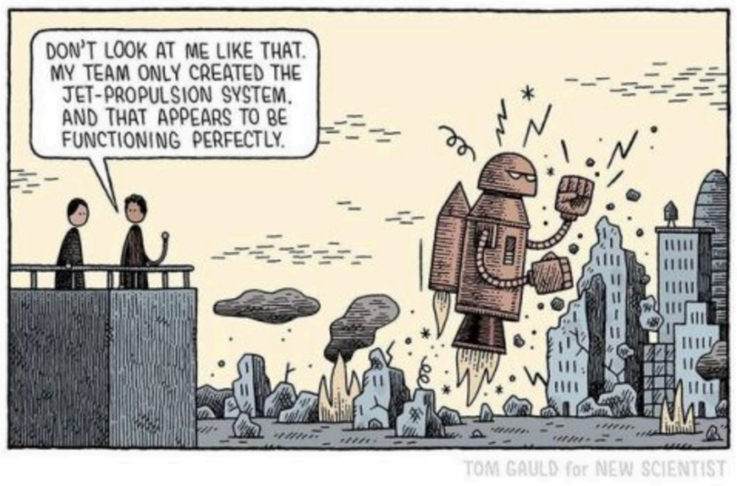
An **agent** makes a **choice**. A **choice** has **consequences**.

Virtue theories

Deontological theories

Normative ethics

- **Moral theory** is an explanation of what makes action right, a state good, or an agent good.



An **agent** makes a **choice**. A **choice** has **consequences**.

Virtue theories Deontological theories Consequentialist theories

Virtue ethics

Virtue ethics

- Put most relevance on the properties of the agent that makes the decisions

Virtue ethics

- Put most relevance on the properties of the agent that makes the decisions
- Prescribe not how to make a decision but what intentions, objectives and preferences, i.e. virtues, the agent should have in order to choose right.

Virtue ethics

- Put most relevance on the properties of the agent that makes the decisions
- Prescribe not how to make a decision but what intentions, objectives and preferences, i.e. virtues, the agent should have in order to choose right.
- A virtue is a stable disposition to act and feel according to some ideal model of excellence.

Virtue ethics

- Put most relevance on the properties of the agent that makes the decisions
- Prescribe not how to make a decision but what intentions, objectives and preferences, i.e. virtues, the agent should have in order to choose right.
- A virtue is a stable disposition to act and feel according to some ideal model of excellence.
- A notable virtue theory is **Aristotelian ethics.**

Deontological ethics

Deontological ethics

- Prescribe that the righteousness of a decision should be based on whether the chosen option is itself right or wrong under a series of rules.

Deontological ethics

- Prescribe that the righteousness of a decision should be based on whether the chosen option is itself right or wrong under a series of rules.
- Prescribe obligations, permissions, prohibitions that the agent should follow when choosing between alternatives

Deontological ethics

- Prescribe that the righteousness of a decision should be based on whether the chosen option is itself right or wrong under a series of rules.
- Prescribe obligations, permissions, prohibitions that the agent should follow when choosing between alternatives
- Deontological theories also prescribe ethical values or ethical principles that one should follow in order to identify the right choices: non-malevolence, justice, autonomy etc.

Deontological ethics

- Prescribe that the righteousness of a decision should be based on whether the chosen option is itself right or wrong under a series of rules.
- Prescribe obligations, permissions, prohibitions that the agent should follow when choosing between alternatives
- Deontological theories also prescribe ethical values or ethical principles that one should follow in order to identify the right choices: non-malevolence, justice, autonomy etc.
- A notable deontological theory is **Kantian ethics.**

Consequentialist ethics

Consequentialist ethics

- Prescribe that a decision is moral if it is motivated by assessing the consequences of the available options, namely what kind of states of affairs they bring about.

Consequentialist ethics

- Prescribe that a decision is moral if it is motivated by assessing the consequences of the available options, namely what kind of states of affairs they bring about.

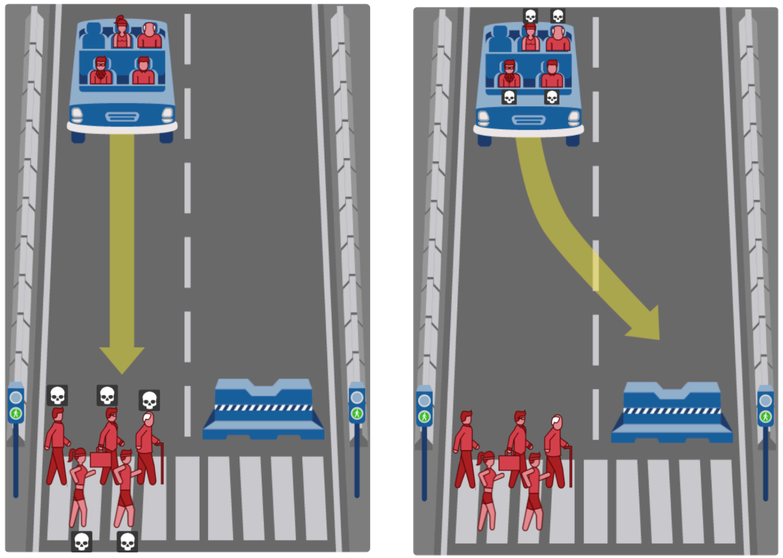
- A notable consequentialist theory is **Utilitarian ethics.**

Consequentialist ethics

- Prescribe that a decision is moral if it is motivated by assessing the consequences of the available options, namely what kind of states of affairs they bring about.
- A notable consequentialist theory is **Utilitarian ethics.**
- Utilitarianism is the theory asserting that the morally right action is the one that produces the most favourable balance of good over evil, everyone considered.

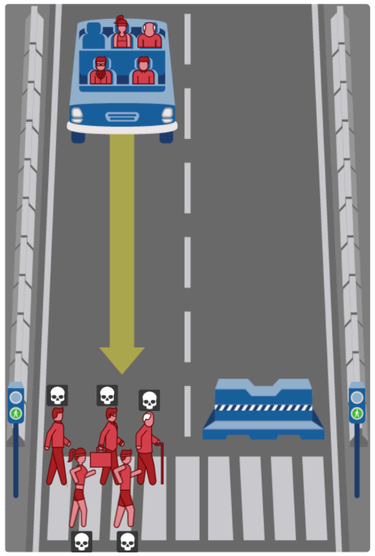
Ethical dilemmas in moral philosophy

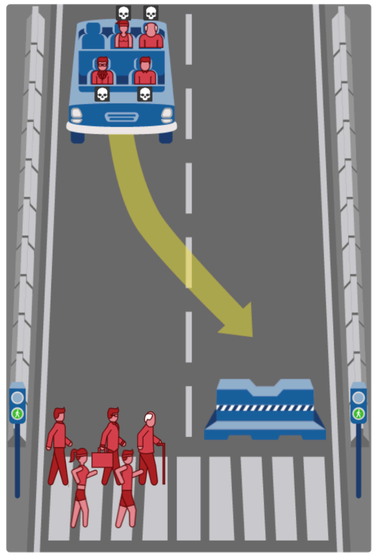
What should the self-driving car do?



Ethical dilemmas in moral philosophy

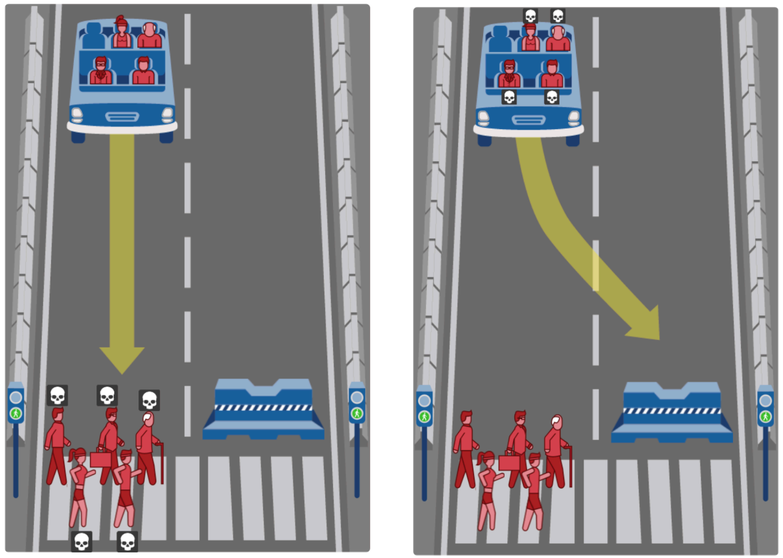
What should the self-driving car do?





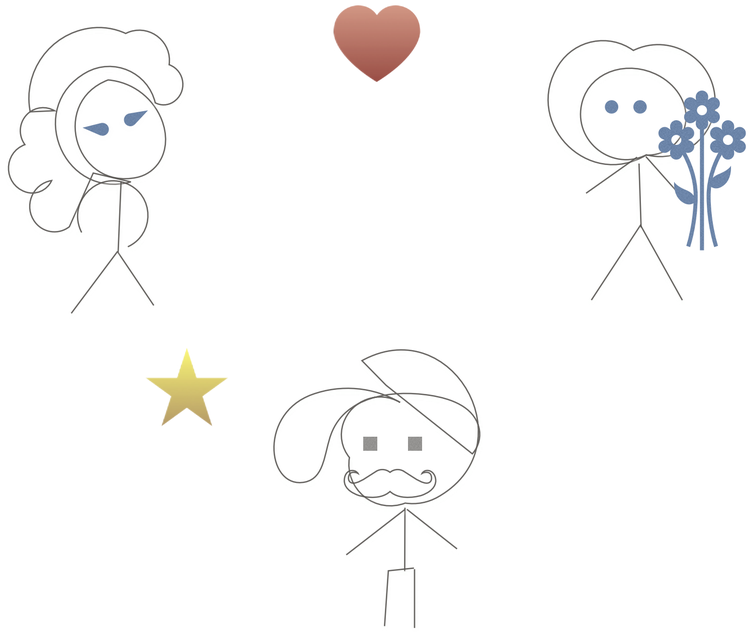
Ethical dilemmas in moral philosophy

What should the self-driving car do?



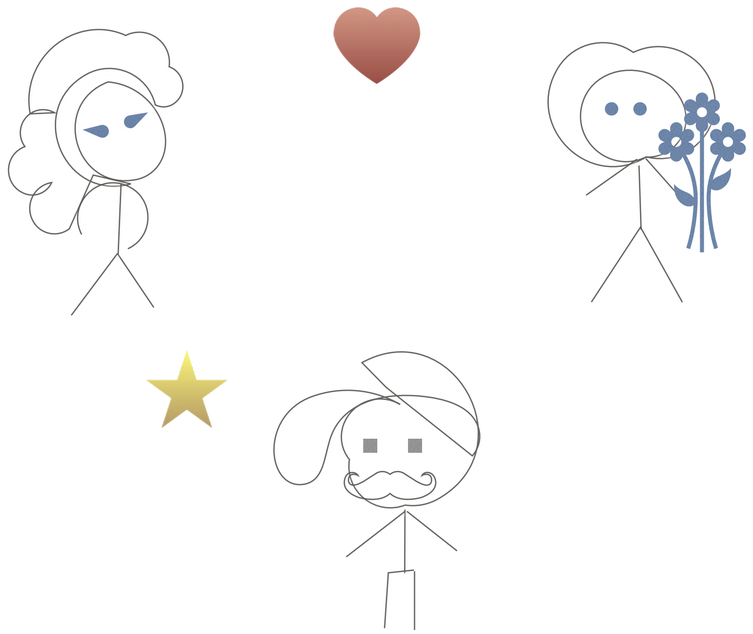
Ethical dilemmas in moral philosophy

The principle of double effect



Ethical dilemmas in moral philosophy

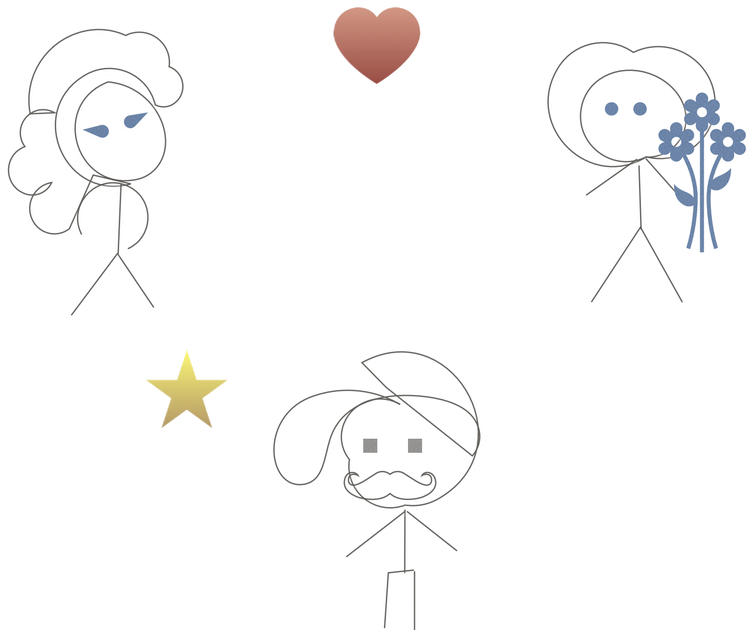
The principle of double effect



The categorical imperative, Kantian ethics

Ethical dilemmas in moral philosophy

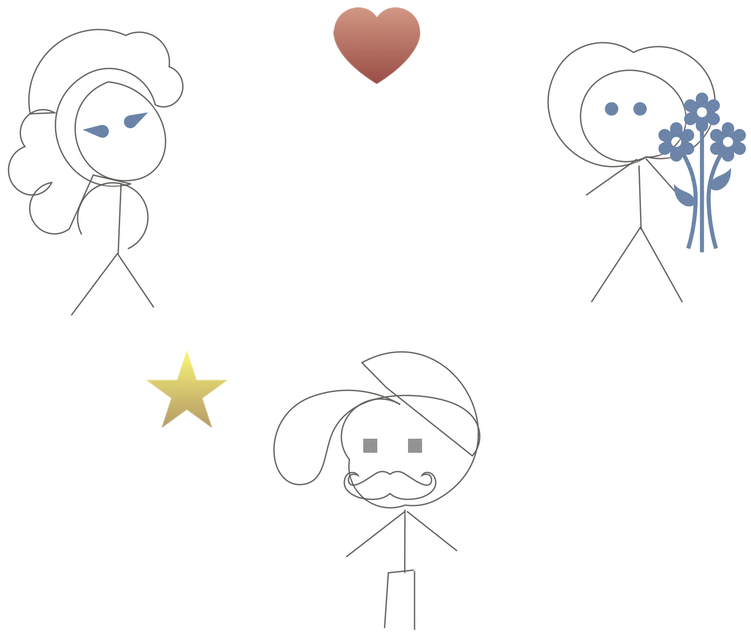
The principle of double effect



The categorical imperative, Kantian ethics

Ethical dilemmas in moral philosophy

The principle of double effect

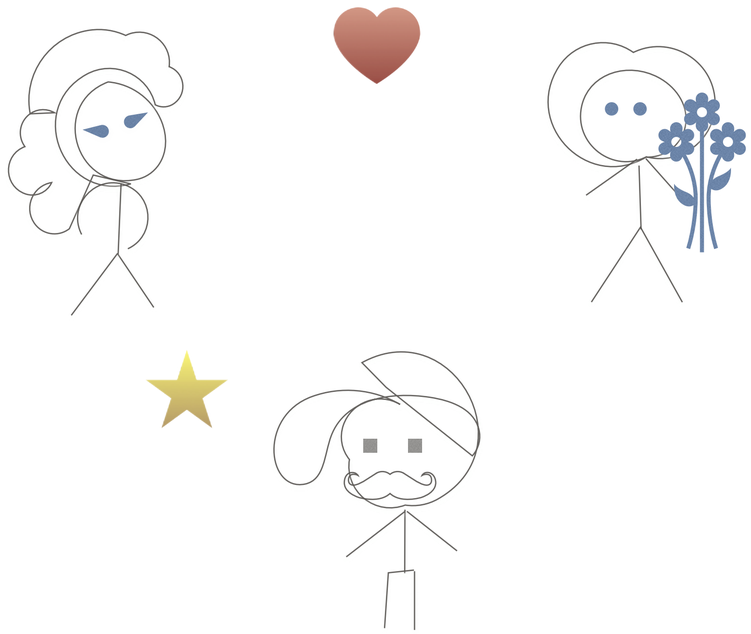


The categorical imperative, Kantian ethics

- “Act only on that maxim through which you can at the same time will that it should become a universal law”

Ethical dilemmas in moral philosophy

The principle of double effect

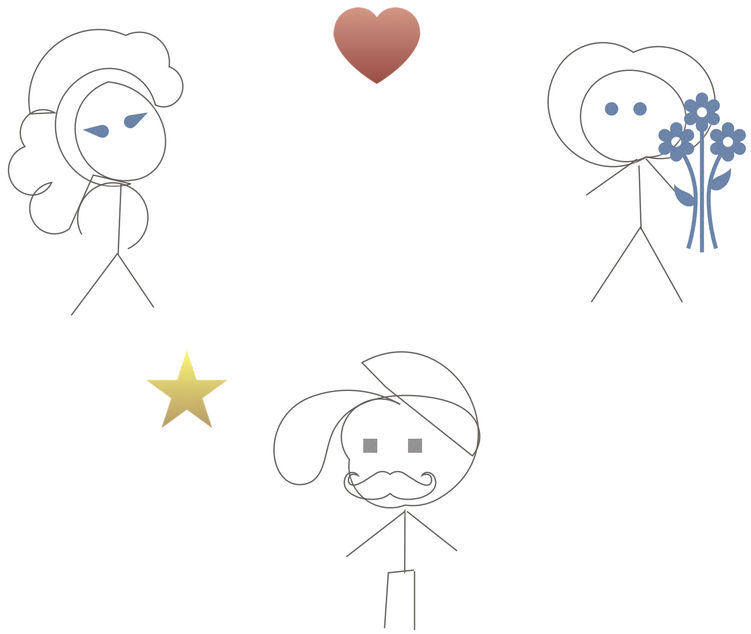


The categorical imperative, Kantian ethics

- “Act only on that maxim through which you can at the same time will that it should become a universal law”
- An action is permissible if:

Ethical dilemmas in moral philosophy

The principle of double effect



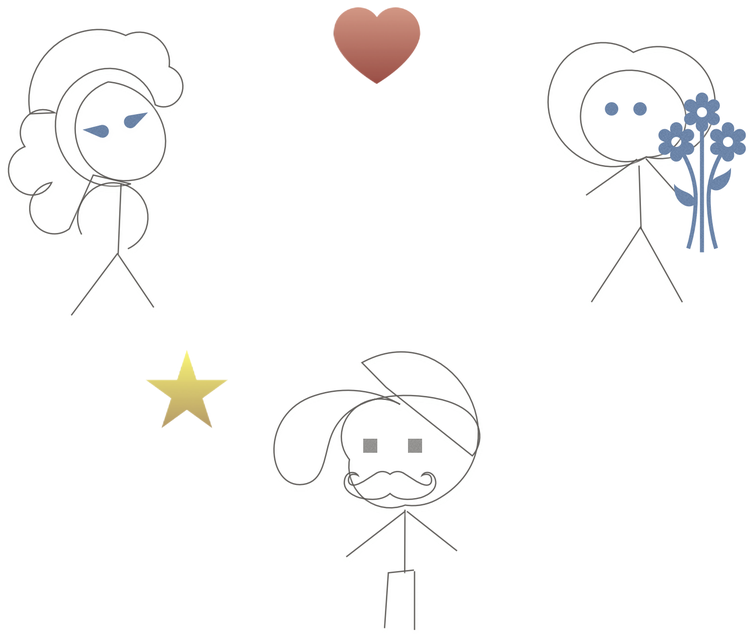
The categorical imperative, Kantian ethics

- “Act only on that maxim through which you can at the same time will that it should become a universal law”
- An action is permissible if:

1) its maxim can be universalised (if everyone can use it in similar situations)

Ethical dilemmas in moral philosophy

The principle of double effect



The categorical imperative, Kantian ethics

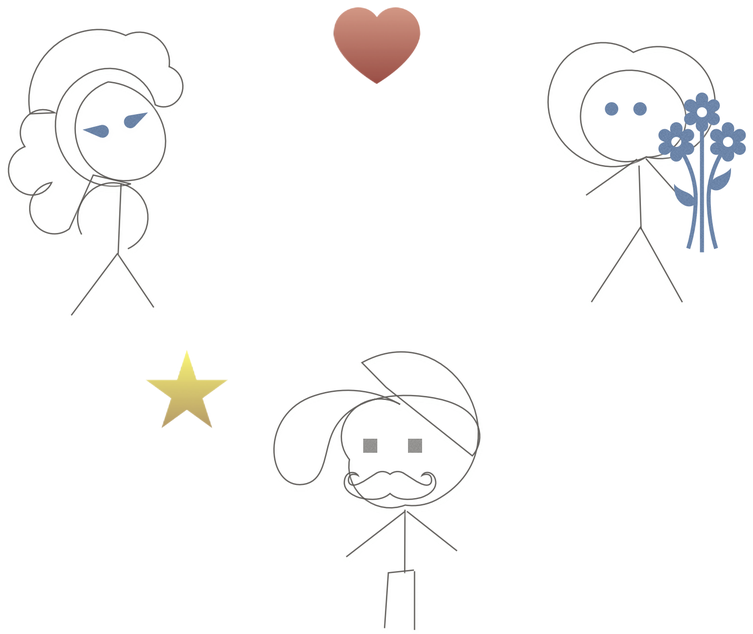
- “Act only on that maxim through which you can at the same time will that it should become a universal law”
- An action is permissible if:

  1) its maxim can be universalised (if everyone can use it in similar situations)

  2) you would be willing to let 1) happen

Ethical dilemmas in moral philosophy

The principle of double effect



The categorical imperative, Kantian ethics

- “Act only on that maxim through which you can at the same time will that it should become a universal law”
- An action is permissible if:

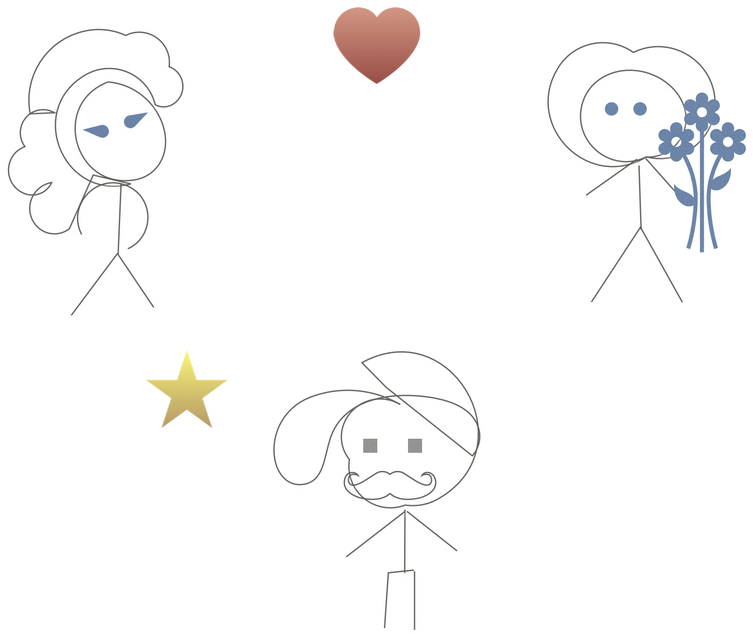
  1) its maxim can be universalised (if everyone can use it in similar situations)

  2) you would be willing to let 1) happen

- “So act to treat humanity, whether in thine own person or in that of any other, in every case as an end withal, never as means only.”

Ethical dilemmas in moral philosophy

The principle of double effect



The categorical imperative, Kantian ethics

- “Act only on that maxim through which you can at the same time will that it should become a universal law”
- An action is permissible if:

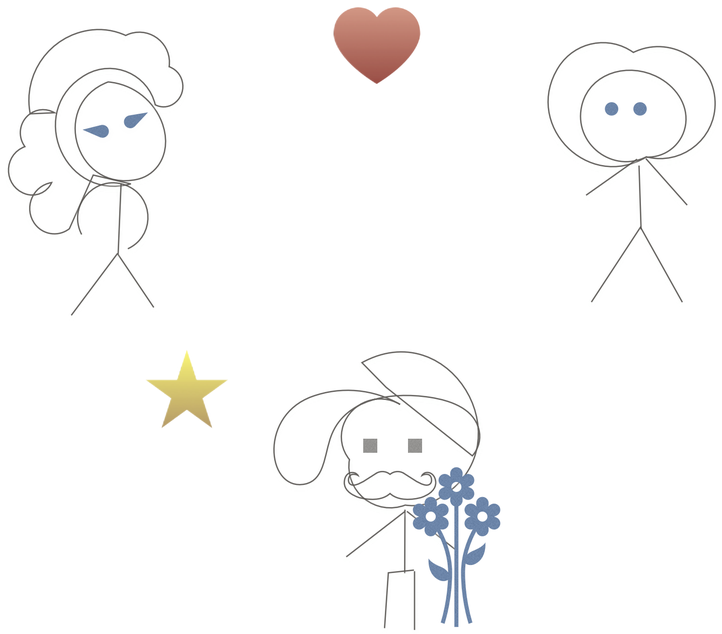
  1) its maxim can be universalised (if everyone can use it in similar situations)

  2) you would be willing to let 1) happen

- “So act to treat humanity, whether in thine own person or in that of any other, in every case as an end withal, never as means only.”

Ethical dilemmas in moral philosophy

The principle of double effect



The categorical imperative, Kantian ethics

- “Act only on that maxim through which you can at the same time will that it should become a universal law”
- An action is permissible if:

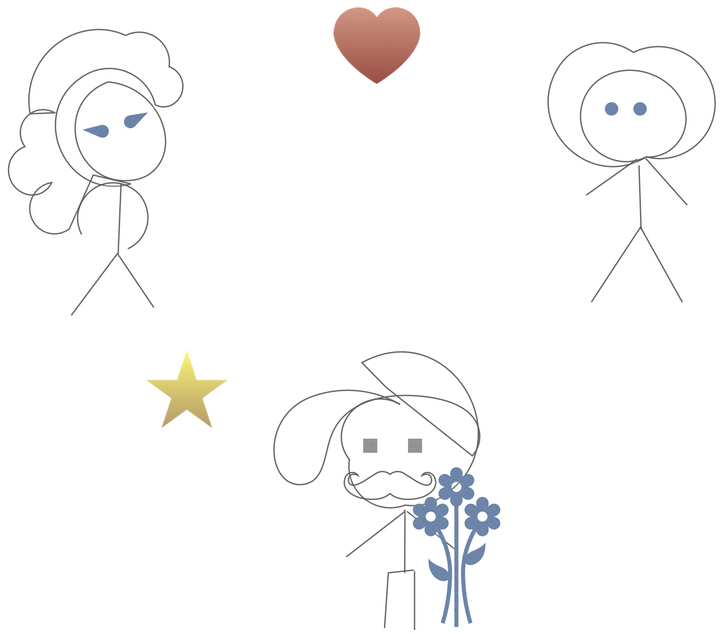
  1) its maxim can be universalised (if everyone can use it in similar situations)

  2) you would be willing to let 1) happen

- “So act to treat humanity, whether in thine own person or in that of any other, in every case as an end withal, never as means only.”

Ethical dilemmas in moral philosophy

The principle of double effect



The categorical imperative, Kantian ethics

- “Act only on that maxim through which you can at the same time will that it should become a universal law”
- An action is permissible if:

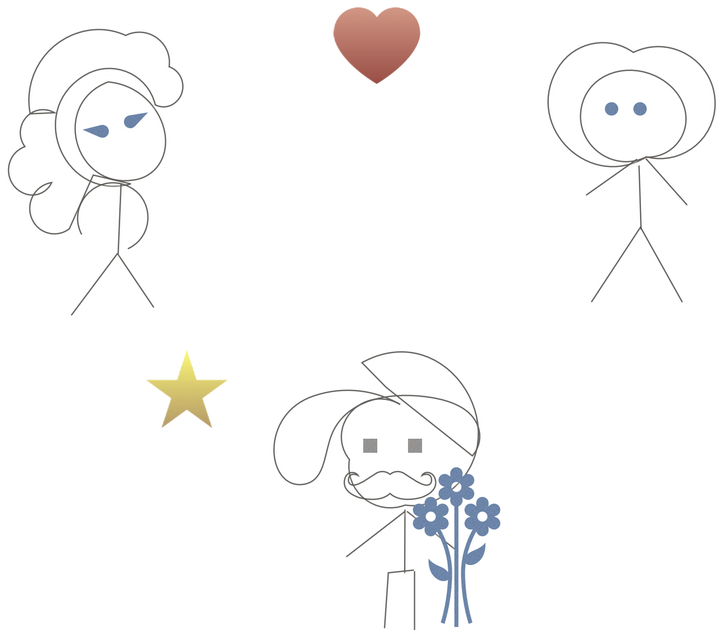
  1) its maxim can be universalised (if everyone can use it in similar situations)

  2) you would be willing to let 1) happen

- “So act to treat humanity, whether in thine own person or in that of any other, in every case as an end withal, never as means only.”

Ethical dilemmas in moral philosophy

The principle of double effect



The categorical imperative, Kantian ethics

- “Act only on that maxim through which you can at the same time will that it should become a universal law”
- An action is permissible if:

  1) its maxim can be universalised (if everyone can use it in similar situations)

  2) you would be willing to let 1) happen

- “So act to treat humanity, whether in thine own person or in that of any other, in every case as an end withal, never as means only.”

Ethical dilemmas in AI

Ethical dilemmas in AI

Ethical dilemmas in AI

Misleading Information	Label	Removal
Disputed Claim	Label	Warning
Unverified Claim	No action	No action*
	Moderate	Severe
Propensity for Harm		

Ethical dilemmas in AI

Misleading Information	Label	Removal
Disputed Claim	Label	Warning
Unverified Claim	No action	No action*
	Moderate	Severe
Propensity for Harm		

I ate an elephant

Ethical dilemmas in AI

Misleading Information	Label	Removal
Disputed Claim	Label	Warning
Unverified Claim	No action	No action*
	Moderate	Severe
Propensity for Harm		

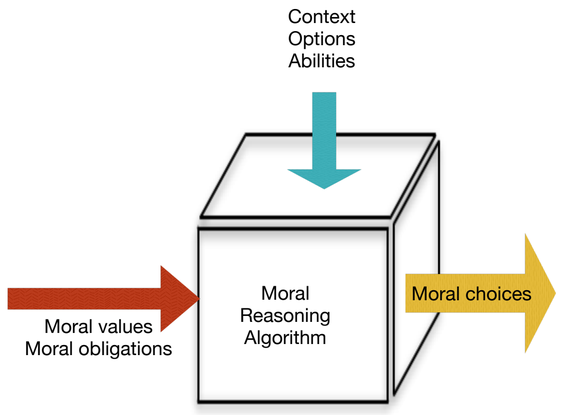
I ate an elephant

The Covid virus does not exist and here is why

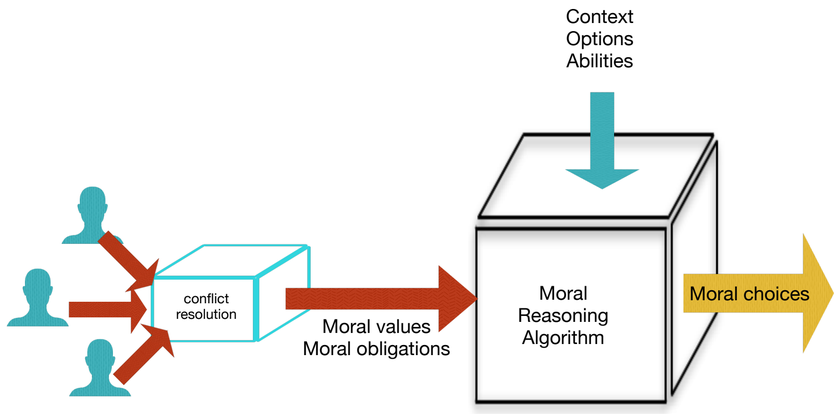
So.. you want to make an agent that does good and not evil?

How do we use operationalise moral theories?

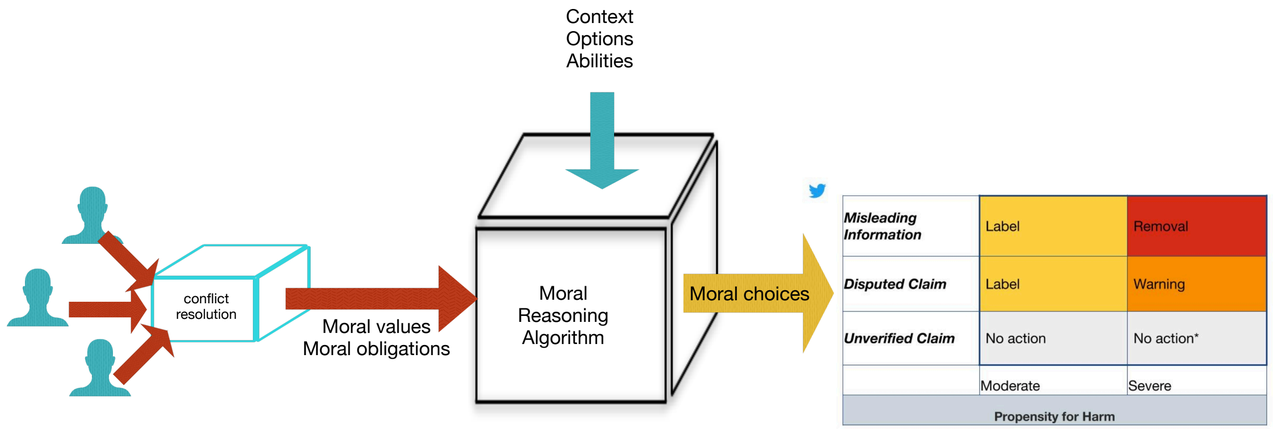
Automating moral reasoning



Automating moral reasoning

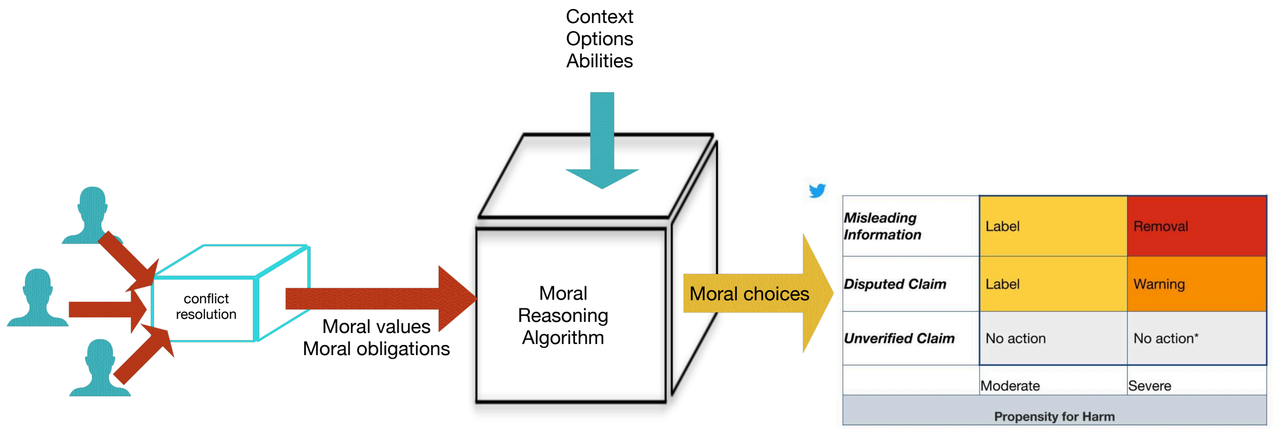


Automating moral reasoning



Misleading Information	Label	Removal
Disputed Claim	Label	Warning
Unverified Claim	No action	No action*
	Moderate	Severe
Propensity for Harm		

Automating moral reasoning



Misleading Information	Label	Removal
Disputed Claim	Label	Warning
Unverified Claim	No action	No action*
	Moderate	Severe
Propensity for Harm		

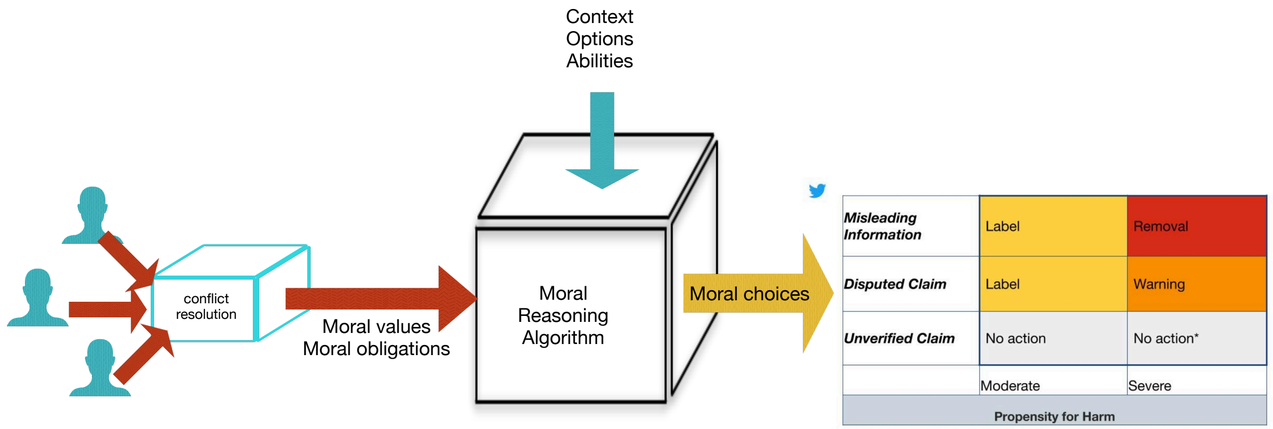
Original Article | Published: 30 September 2017

Social choice ethics in artificial intelligence

Seth D. Baum

AI & SOCIETY **35**, 165–176 (2020) | Cite this article

Automating moral reasoning



Original Article | Published: 30 September 2017

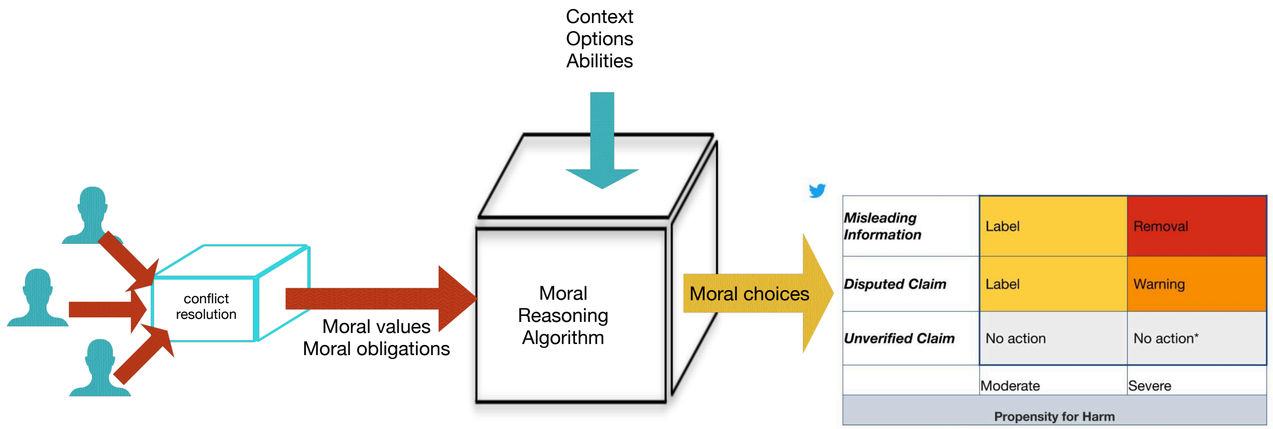
Social choice ethics in artificial intelligence

Seth D. Baum

AI & SOCIETY **35**, 165–176 (2020) | Cite this article

6155 Accesses | **32** Citations | **5** Altmetric | Metrics

Automating moral reasoning



Original Article | Published: 30 September 2017

Social choice ethics in artificial intelligence

Seth D. Baum

AI & SOCIETY **35**, 165–176 (2020) | Cite this article

6155 Accesses | **32** Citations | **5** Altmetric | Metrics

Implementations of social choice ethics must make three types of choices, each of which create their own set of ethical dilemmas (Baum 2009):

1. *Standing* Who or what is included in the group to have its values factored into the AI?
2. *Measurement* What procedure is used to obtain values from each member of the selected group?
3. *Aggregation* How are the values of individual group members combined to form the aggregated group values?

Top-down approach

- In engineering: a problem is iteratively divided into smaller problems until a problem small enough to be solved is reached
- In machine ethics: given is an ethical theory, how can we implement it?
- For explicit agents: how can we build the agent to follow a given moral theory?
- For implicit agents: how can we follow a given moral theory to build an agent.

Top-down approach

- For implicit agents: how can we follow a given moral theory to build an agent.
- Operational morality: the moral significance of their actions lies entirely in the humans involved in their design and use
- Implicit AMA: constrain the machine's actions to avoid unethical outcomes.

Top-down approach

- For explicit agents: how can we build the agent to follow a given moral theory?
- Functional morality: enable AMA to make moral judgments when deciding a course of action without direct instructions from humans.
- Explicit AMA: Represent ethics explicitly and then operate effectively on the basis of this knowledge.

Top-down approach

- For explicit agents: how can we build the agent to follow a given moral theory?
- Functional morality: enable AMA to make moral judgments when deciding a course of action without direct instructions from humans.
- Explicit AMA: Represent ethics explicitly and then operate effectively on the basis of this knowledge.

An **agent** makes a **choice**. A **choice** has **consequences**.

Top-down approach

- For explicit agents: how can we build the agent to follow a given moral theory?
- Functional morality: enable AMA to make moral judgments when deciding a course of action without direct instructions from humans.
- Explicit AMA: Represent ethics explicitly and then operate effectively on the basis of this knowledge.

An **agent** makes a **choice**. A **choice** has **consequences**.

Top-down approach

- For explicit agents: how can we build the agent to follow a given moral theory?
- Functional morality: enable AMA to make moral judgments when deciding a course of action without direct instructions from humans.
- Explicit AMA: Represent ethics explicitly and then operate effectively on the basis of this knowledge.

An **agent** makes a **choice**. A **choice** has **consequences**.

Top-down approach

- For explicit agents: how can we build the agent to follow a given moral theory?
- Functional morality: enable AMA to make moral judgments when deciding a course of action without direct instructions from humans.
- Explicit AMA: Represent ethics explicitly and then operate effectively on the basis of this knowledge.

An **agent** makes a **choice**. A **choice** has **consequences**.

Top-down explicit agents

Top-down explicit agents

Limitations

- Idealistic standards that are hard to meet even by people
- Requires hard AI problems to be solved: situational awareness, prediction of consequences
- Context dependency due to the high AI requirements?
- Which theory should be used?

Top-down explicit agents

Limitations

- Idealistic standards that are hard to meet even by people
- Requires hard AI problems to be solved: situational awareness, prediction of consequences
- Context dependency due to the high AI requirements?
- Which theory should be used?

Advantages

- No surprises: the agent follows a “tried and tested” theory, so we know what to expect
- Behaviour can be verified: we can check if the action of the agent complies with the theory recommendation for the given situation.

Bottom-up approach

- In engineering: describe the solution, not how to reach it
- In machine ethics: train an agent to follow a moral theory (or behave morally)
- For implicit agents: how to build an agent that learns to reason correctly according to a moral theory
- For explicit agents: how to build an agent that learns how to follow a moral theory

Bottom-up approach

- In engineering: describe the solution, not how to reach it
- In machine ethics: train an agent to follow a moral theory (or behave morally)
- For implicit agents: how to build an agent that learns to reason correctly according to a moral theory
- For explicit agents: how to build an agent that learns how to follow a moral theory
- Operational morality: the moral significance of their actions lies entirely in the humans involved in their design and use
- Implicit AMA: constrain the machine's actions to avoid unethical outcomes.

Bottom-up approach

- In engineering: describe the solution, not how to reach it
- In machine ethics: train an agent to follow a moral theory (or behave morally)
- For implicit agents: how to build an agent that learns to reason correctly according to a moral theory
- For explicit agents: how to build an agent that learns how to follow a moral theory

- Operational morality: the moral significance of their actions lies entirely in the humans involved in their design and use
- Implicit AMA: constrain the machine's actions to avoid unethical outcomes.
- Functional morality: enable AMA to make moral judgments when deciding a course of action without direct instructions from humans.
- Explicit AMA: Represent ethics explicitly and then operate effectively on the basis of this knowledge.

Bottom-up approach

- People make moral decisions without following a specific moral theory
- A strength of bottom-up engineering lies in the assembly of components to achieve a goal.
- The idea is that instead of learning how to always be moral, an artificial agent will learn little by little as people do
- No clear approach has been put forward how to do this.

Bottom-up approach

- People make moral decisions without following a specific moral theory
- A strength of bottom-up engineering lies in the assembly of components to achieve a goal.
- The idea is that instead of learning how to always be moral, an artificial agent will learn little by little as people do
- No clear approach has been put forward how to do this.

Michael Anderson and Susan Leigh Anderson GenEth: a general ethical dilemma analyzer
https://doi.org/10.1515/pjbr-2018-0024 (Links to an external site.)

Bottom-up approach

- People make moral decisions without following a specific moral theory
- A strength of bottom-up engineering lies in the assembly of components to achieve a goal.
- The idea is that instead of learning how to always be moral, an artificial agent will learn little by little as people do
- No clear approach has been put forward how to do this.

Michael Anderson and Susan Leigh Anderson GenEth: a general ethical dilemma analyzer
https://doi.org/10.1515/pjbr-2018-0024 (Links to an external site.)

Reinforcement Learning As a Framework for Ethical Decision Making David Abel and James MacGlashan and Michael L. Littman
https://david-abel.github.io/papers/wkshp_aaai2016_rl_ethics.pdf

Bottom-up approach limitations and advantages

Limitations:

- No safeguards. How to be sure that a reasonable behaviour will be learned.
- Example: Tay https://en.wikipedia.org/wiki/Tay_(bot)
- Requires hard AI problems to be solved: situational awareness, prediction of consequences, learning from environment
- How to specify the right examples or objective function?
- The training data is very much **not** free
- How to verify/prove that ethical behaviour really has been reached?

Bottom-up approach limitations and advantages

Limitations:

- No safeguards. How to be sure that a reasonable behaviour will be learned.
- Example: Tay https://en.wikipedia.org/wiki/Tay_(bot)
- Requires hard AI problems to be solved: situational awareness, prediction of consequences, learning from environment
- How to specify the right examples or objective function?
- The training data is very much **not** free
- How to verify/prove that ethical behaviour really has been reached?

Advantages:

- Robust, supportive of autonomous behaviour
- We do not have to worry that we have forgotten to specify some rule

Ethical Turing Test - prove that you are a moral explicit agent

Article | Full-text available

Prolegomena to any future artificial moral agent

July 2000 · Journal of Experimental & Theoretical Artificial Intelligence 12(3):251-261

DOI · 10.1080/09528130050111428

Source · DBLP

Colin Allen
34.42 · Indiana University Bloomingt...

Gary Varner
Not on ResearchGate

Jason Zinser
Not on ResearchGate

Ethics and Information Technology

June 2016, Volume 18, Issue 2, pp 103–115 | Cite as

Against the moral Turing test: accountable design and the moral reasoning of autonomous systems

Authors | Authors and affiliations

Thomas Arnold, Matthias Scheutz

Original Paper

First Online: 29 March 2016

More material

http://slavkovik.com/ijcaitutorial2020.html

https://www.youtube.com/watch?v=H7n1W8J1vWo

Why so little machine learning

- So far used: symbolic learning
- Reinforcement learning: requires simulations
- Supervised learning: requires labeled examples+