#### 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?

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



#### What is AI DD ™?

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



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.



devices that can do some

things by themselves and the humans that use them



devices that can do some

things by themselves and the humans that use them

Utgiftsrefusjon

Utgifter								
	Туре	Vedlegg	Beskrivelse	Beløp	Kurs	Beløp NOK		
	Faglitteratur/bøker	1	Hands-On Explainable AI (XAI)	44,85 GBP	11,629	<b>521,55</b> >		
	Totalt					521,55		

#### Vedlegg

#### Vedlegg fra utgiftsposter

Faglitteratur/bøker (44.85 GBP) - 1 <sub>pdf</sub>



devices that can do some

things by themselves and the humans that use them

 Utgiftsrefusjon
 Rediger

 Type
 Vedlegg
 Beskrivelse
 Polop Kurs
 Beløp NOK

 Faglitteratur/bøker
 1
 Hands-On Explainable AI (XAI)
 44,85 GBP 11,629
 521,55 >

 Totalt
 521,55
 521,55
 521,55
 521,55

Vedlegg

#### Vedlegg fra utgiftsposter

Faglitteratur/bøker (44.85 GBP) - 1 <sub>pdf</sub>



devices that can do some

things by themselves and the humans that use them

Utgiftsrefusjon Utgifter Rediger Beløp NOK Туре Beskrivelse Vedlegg Kurs 44,85 GBP 11,629 521,55 Hands-On Faglitteratur/bøker Explainable AI (XAI) 521,55 Totalt 09.10.2021 Order date 026-3107586-1263557 Order # Invoice details VAT rate Description Qty Unit price Unit price Item subtotal (excl. VAT) (incl. VAT) (incl. VAT) Hands-On Explainable AI (XAI) with Python: Interpret, visualize, explain, and £37.99 £37.99 £37.99 0% integrate reliable AI for fair, secure, and trustworthy AI apps | 1800208138 ASIN: 1800208138 £6.86 £6.86 Shipping Charges £6.86 Invoice total £44.85 VAT rate VAT subtotal Item subtotal (excl. VAT) ----....



devices that can do some

things by themselves and the humans that use them

Utgiftsrefusjon Utgifter Rediger Beløp NOK Туре Beskrivelse Vedlegg Kurs 44,85 GBP 11,629 521,55 Hands-On Faglitteratur/bøker Explainable AI (XAI) 521,55 Totalt 09.10.2021 Order date 026-3107586-1263557 Order # Invoice details VAT rate Description Qty Unit price Unit price Item subtotal (excl. VAT) (incl. VAT) (incl. VAT) Hands-On Explainable AI (XAI) with Python: Interpret, visualize, explain, and £37.99 £37.99 £37.99 0% integrate reliable AI for fair, secure, and trustworthy AI apps | 1800208138 ASIN: 1800208138 £6.86 Shipping Charges £6.86 £6.86 Invoice total £44.85 VAT rate Item subtotal (excl. VAT) -----....



#### 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



286



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



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,



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,



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,



- 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.
  - 4.choose from the identified options following some reasoning.



- 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.



 A decision is ethical if it is made not only on the factual identified societies consideration of what is right or wrong.

objectives, preferences and constraints, but also based on a person's or

- A decision is ethical if it is made not only on the factual identified societies consideration of what is right or wrong.
- weight with ones own"

# objectives, preferences and constraints, but also based on a person's or

• Ethical decisions include considering "the interests of others as of equal

- A decision is ethical if it is made not only on the factual identified societies consideration of what is right or wrong.
- weight with ones own"
- which states of the world are best to be achieved.

# objectives, preferences and constraints, but also based on a person's or

• Ethical decisions include considering "the interests of others as of equal

 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

#### Al, agent, morality

#### Al, agent, morality

• Al does not make decisions, an agent makes decisions!

#### Al, agent, morality

#### • Al does not make decisions, an agent makes decisions!



Charlie bit my finger! ORIGINAL

#### What is an agent? An entity that acts in an environment.



An agent as an input-output system





tasks associated with agency



#### In nature, we tend to call an agent something that is able to do all the

- tasks associated with agency

#### In nature, we tend to call an agent something that is able to do all the

• There are different type of artificial agents: reactive, intelligent, embodied

- tasks associated with agency
- Example: playing chess against a computer

#### In nature, we tend to call an agent something that is able to do all the

• There are different type of artificial agents: reactive, intelligent, embodied

- tasks associated with agency
- Example: playing chess against a computer

#### In nature, we tend to call an agent something that is able to do all the

#### • There are different type of artificial agents: reactive, intelligent, embodied



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



- tasks associated with agency
- Example: playing chess against a computer

#### In nature, we tend to call an agent something that is able to do all the

#### • There are different type of artificial agents: reactive, intelligent, embodied



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

Genres: Board game; Abstract strategy game; Mind sport

Players: 2

Skills required: Strategy, tactics



## 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 😮

J.H. Moor View All Authors

#### The Nature, Importance, and Difficulty of Machine Ethics

er ions	<b>3005</b> Full Text Views		PDF	₹	R	$\checkmark$	ē	©	₫	
stract		Abstract:								
cument S	Sections	<ul> <li>The question of whether machine ethics exists agree on what counts as machine ethics. Some</li> </ul>	or might exist in th e might argue that	ie futu machi	re is d ne eth	ifficult 1 ics ob\	o ansv viouslv	ver if v exists	ve can becau	't Jse
Varieties achine E	s of thics	humans are machines and humans have ethics exist because ethics is simply emotional expres positions on machine ethics are possible, and a	s. Others could arg ssion and machine a discussion of the	jue tha s can' issue	at mac t have could	hine et emotio rapidly	hics ot ons. A	ovious wide ra	ly does ange c to dee	sn't of p and
Ethical-Impact gents		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								
Implicit Ethical gents		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 deriving								
Explicit	Ethical	asserting of denying								
gents		Published in: IEEE Intelligent Systems (Volur	me: 21,Issue: 4,	July-A	ug. 20	06)				
Full Ethi	cal Agents	<b>Page(s):</b> 18 - 21	INSPEC A	cces	sion N	umbe	<b>:</b> 9065	956		
hors		Date of Publication: 07 August 2006 🕜	DOI: 10.1	109/M	IS.200	6.80				








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

Authors

### The Nature, Importance, and Difficulty of Machine Ethics

Author(s)	J.H. Moor	View All

	3005
er	Full
ions	Text Viev



### 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)

5. Full Ethical Agents

Page(s): 18 - 21

Date of Publication: 07 August 2006 3

**INSPEC Accession Number:** 9065956







69 Pape

**Document Sections** 

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

Authors

• Ethical-impact AMA





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

Authors

### The Nature, Importance, and Difficulty of Machine Ethics

Author(s)	J.H. Moor	View All

	3005
er	Full
ions	Text Viev



### Abstract

### 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)

5. Full Ethical Agents

Page(s): 18 - 21

Date of Publication: 07 August 2006 3

**INSPEC Accession Number:** 9065956



- Ethical-impact AMA
- Implicit ethical AMA





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

Authors

### The Nature, Importance, and Difficulty of Machine Ethics

Author(s)	J.H. Moor	View All

	3005
er	Full
ions	Text Viev



### Abstract

**Document Sections** 

1. Varieties of

Machine Ethics

2. Ethical-Impact

3. Implicit Ethical

4. Explicit Ethical

Agents

Agents

Agents

Authors

69

Pape

### 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)

5. Full Ethical Agents

Page(s): 18 - 21

Date of Publication: 07 August 2006 3

**INSPEC Accession Number:** 9065956



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





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

Authors

### The Nature, Importance, and Difficulty of Machine Ethics

Author(s)	J.H. Moor	View All

	3005
er	Full
ions	Text Viev



### Abstract

**Document Sections** 

1. Varieties of

Machine Ethics

2. Ethical-Impact

3. Implicit Ethical

4. Explicit Ethical

Agents

Agents

Agents

Authors

69

Pape

### 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)

5. Full Ethical Agents

Page(s): 18 - 21

Date of Publication: 07 August 2006 3

**INSPEC Accession Number:** 9065956



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





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

Authors

### The Nature, Importance, and Difficulty of Machine Ethics

Author(s)	J.H. Moor	View All

	3005
er	Full
ions	Text Viev



### Abstract

**Document Sections** 

1. Varieties of

Machine Ethics

2. Ethical-Impact

3. Implicit Ethical

4. Explicit Ethical

Agents

Agents

Agents

Authors

69 Pape

### 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)

5. Full Ethical Agents

Page(s): 18 - 21

Date of Publication: 07 August 2006 3

**INSPEC Accession Number:** 9065956



• can operate autonomously,

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

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

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



Camel Races, Al Sheehaniya - 2008



H Humanium

When technology helps end child labour: the crazy race of robot jockeys -Humanium

Images may be subject to copyright. Learn more

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







Wikipedia

H Humanium

When technology helps end child labour: the crazy race of robot jockeys -Humanium

Images may be subject to copyright. Learn more



### Jacquard machine

The Jacquard machine is a device fitted to a simplifies the process of manufacturing text such complex patterns as brocade, damask matelassé. The resulting ensemble of the lc Jacquard machine is then called a Jacquard

### Ser 1

Sosiale tjenester Økonomisk sosialhjelp

### Søknad om økonomisk sosialhjelp

Du skal søke til NAV-kontoret der du bor. Stadig flere kommuner kan ta i mot digitale søknader. Hvis du ikke skal søke digitalt, kan du søke med kommunens papirskjema.

Skal du ettersende dokumentasjon? Hvis du har søkt digitalt kan du gå til oversikten over dine digitale søknader for å ettersende dokumentasjon.

### **Relatert informasjon**

Informasjon om økonomisk sosialhjelp

<u>Slik klager du på vedtak</u>









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



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



- 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





- 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





- 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





accountability transparency explainability interpretability













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



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



- 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 ethical agents



- Implicit ethical agents
- Explicit ethical agents

- Implicit ethical agents
- Explicit ethical agents



- Implicit ethical agents
- Explicit ethical 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 Al 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 Al 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 simplify environments and build complex reasoning systems for these simplify factory robots can do sophisticated tasks in the engineered environment of **Robotstøvsuger som ruller og samler støv** hopeless in a natural environment. Much of the complexity of the task can Mocoro Mop Ball | Art. nr.: 102680 environment.

### Simplify the agent

Simple agents in natural environments. This is inspired by seeing how **in** 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?



wrong behavior".

• Ethics or moral philosophy is a branch of philosophy that "involves systematizing, defending, and recommending concepts of right and

. . . .

wrong behavior".

lep.utm.edu	i/e/															
	А	В	С	D	E	F	G	н	I	J	К	L	М	N	0	P
		•	Virtu	<u>1e</u>												
		• <u>E</u> ţ	osilon	Calc	<u>uli</u>											
		• <u>Er</u>	asmu	<u>15, De</u>	eside	rius										
		• Et	hics													
		•	<u>Anci</u>	ent												
		•	<u>Anin</u>	<u>nals</u>												
		•	<u>Appl</u>	lied												
		•	Artif	icial	Intel	ligenc	<u>ee</u>									
		•	Bert	rand	Russ	<u>ell's</u>										
		•	<u>Care</u>													
		•	<u>Care</u>	-Wor	<u>ker I</u>	<u>Migra</u>	tion									
		•	Cons	strast	ivisn	<u>1</u>										
		•	<u>Envi</u>	ronn	ienta	1										
		•	Evol	ution	<u>ary</u>											
		•	<u>Expr</u>	essiv	<u>vism</u>											
		•	Mod	ern a	nd A	ncien	<u>t</u>									
		•	Non	-Cogi	nitivi	<u>sm</u>										
		•	<u>Pher</u>	nome	nolog	ЗУ.										
		•	<u>Self-</u>	Dece	ption	L										
		•	<u>Stoic</u>	2												
		•	<u>Surv</u>	eillar	<u>ice</u>											
		•	Virtu	<u>1e</u>												

• Ethics or moral philosophy is a branch of philosophy that "involves systematizing, defending, and recommending concepts of right and

# • Ethics or moral philosophy is a branch. of philosophy that "involves systematizing, defending, and recommending concepts of right and wrong behavior".



n/searcher.py?query=moral+philosophy	Û	☆	æ <mark>2</mark>	÷ F	1
h					
al philosophy	Q				
of 2181 documents found					
erimental Moral Philosophy					
rimental Moral Philosophy 4.1 Folk Metaethics and Moral Realism 4.2 Moral Disagreement					
sonby is the empirical study of moral intuitions. Criticisms of Experimental Moral					
sophy is the empirical study of <b>moral</b> intuitionsentitisms of Experimental <b>Moral</b>					
Alfano, Don Loeb, and Alexandra Plakias					
://plato.stanford.edu/entries/experimental-moral/					
t's Moral Philosophy					
ods of Moral Philosophy 2. Good Will, Moral Worth and Duty 3. Duty and Respect for			(		
al Law 4. Categorical Aims and Methods of Moral Philosophy The most basic aim of					
I philosophy, and so also of the Groundworkour basic moral duties to ourselves and others.					
dition, Kant thought that moral philosophy should characterize					
://plato.stanford.edu/entries/kant-moral/					
ke's Moral Philosophy					
f figuring out human <b>moral</b> duty. By looking at Locke's <b>moral philosophy</b> , as it is developed					
interpretations of Locke's moral philosophy 2. Locke's natural law theory: the basis of					
l obligation Locke's moral philosophy There are two main stumbling blocks to the study					
cke's moral philosophy					
cia Sheridan					
://plato.stanford.edu/entries/locke-moral/					

### • Ethics Or moral philosophy systematizing, defending, and re wrong behavior".



### Search

### ethics

1-10 of 1210 documents found

### Virtue **Ethics**

Virtue Ethics 2.1 Eudaimonist Virtue Ethics 2.2 Agent-Based and Exemplarist Virtue Ethics 2.3...Target-Centered Virtue Ethics 2.4 Platonistic Virtue Ethics 3. Objections to virtue ethics 4. Future ...virtue ethics, namely, a) eudaimonist virtue ethics, b) agent-based and exemplarist virtue ethics, c) target-centered... **Rosalind Hursthouse and Glen Pettigrove** 

. . . .

https://plato.stanford.edu/entries/ethics-virtue/

### Feminist Ethics

Since feminist ethics is not merely a branch of ethics, but is instead "a way of doing ethics" (Lindemann... branch of ethics, including meta-ethics, normative theory, and practical or applied ethics. The point... Feminist Ethics First published Mon May 27, 2019 Feminist Ethics aims "to understand, criticize...

Kathryn Norlock https://plato.stanford.edu/entries/feminism-ethics/

### Business Ethics

Business Ethics Quarterly, 17(4): 689–727. —, 2008, "The Ethics of Price Gouging", Business Ethics Quarterly...developed and enforced by teams of ethics and compliance personnel. Business ethics can thus be understood as ... consider this form of business ethics. Instead, it considers business ethics as an academic discipline. ...

Jeffrey Moriarty https://plato.stanford.edu/entries/ethics-business/

### African Ethics

### • Ethics Or moral philosophy systematizing, defending, and re wrong behavior".



### Search

### ethics

1-10 of 1210 documents found

### Virtue **Ethics**

Virtue Ethics 2.1 Eudaimonist Virtue Ethics 2.2 Agent-Based and Exemplarist Virtue Ethics 2.3...Target-Centered Virtue Ethics 2.4 Platonistic Virtue Ethics 3. Objections to virtue ethics 4. Future ...virtue ethics, namely, a) eudaimonist virtue ethics, b) agent-based and exemplarist virtue ethics, c) target-centered... **Rosalind Hursthouse and Glen Pettigrove** 

. . . .

https://plato.stanford.edu/entries/ethics-virtue/

### Feminist Ethics

Since feminist ethics is not merely a branch of ethics, but is instead "a way of doing ethics" (Lindemann... branch of ethics, including meta-ethics, normative theory, and practical or applied ethics. The point... Feminist Ethics First published Mon May 27, 2019 Feminist Ethics aims "to understand, criticize...

Kathryn Norlock https://plato.stanford.edu/entries/feminism-ethics/

### Business Ethics

Business Ethics Quarterly, 17(4): 689–727. —, 2008, "The Ethics of Price Gouging", Business Ethics Quarterly...developed and enforced by teams of ethics and compliance personnel. Business ethics can thus be understood as ... consider this form of business ethics. Instead, it considers business ethics as an academic discipline. ...

Jeffrey Moriarty https://plato.stanford.edu/entries/ethics-business/

### African Ethics

## What is morality?



RMHARE



### Applications of Moral Philosophy

Authors (view affiliations)

R. M. Hare



Part of the <u>New Studies in Practical Philosophy</u> book series



systematizing, defending, and recommending concepts of right and wrong behavior".

• Ethics or moral philosophy is a branch of philosophy that "involves

- Ethics or moral philosophy is a branch 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



- Ethics or moral philosophy is a branch 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.



- Ethics or moral philosophy is a branch 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.



or an agent good.

### • Moral theory is an explanation of what makes action right, a state good,

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



TOM GAULD for NEW SCIENTIST

## • 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.

TOM GAULD for NEW SCIENTIST

## • 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

TOM GAULD for NEW SCIENTIST

## • 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

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



Virtue theories

Deontological theories

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

Consequentialist theories

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

- 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.

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

 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

- Put most relevance on the properties of the agent that makes the decisions
- model of excellence.
- A notable virtue theory is **Aristotelian ethics**.

 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

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 that the righteousness of a decision should be based on
- Prescribe obligations, permissions, prohibitions that the agent should follow when choosing between alternatives

## whether the chosen option is itself right or wrong under a series of rules.

- Prescribe that the righteousness of a decision should be based on
- Prescribe obligations, permissions, prohibitions that the agent should follow when choosing between alternatives
- that one should follow in order to identify the right choices: nonmalevolence, justice, autonomy etc.

## whether the chosen option is itself right or wrong under a series of rules.

• Deontological theories also prescribe ethical values or ethical principles

- Prescribe that the righteousness of a decision should be based on
- Prescribe obligations, permissions, prohibitions that the agent should follow when choosing between alternatives
- that one should follow in order to identify the right choices: nonmalevolence, justice, autonomy etc.
- A notable deontological theory is **Kantian ethics**.

## whether the chosen option is itself right or wrong under a series of rules.

• Deontological theories also prescribe ethical values or ethical principles
Prescribe that a decision is moral if it is motivated by assessing the affairs they bring about.

consequences of the available options, namely what kind of states of

- Prescribe that a decision is moral if it is motivated by assessing the affairs they bring about.
- A notable consequentialist theory is Utilitarian ethics.

# consequences of the available options, namely what kind of states of

- Prescribe that a decision is moral if it is motivated by assessing the affairs they bring about.
- A notable consequentialist theory is Utilitarian ethics.
- considered.

consequences of the available options, namely what kind of states of

• Utilitarianism is the theory asserting that the morally right action is the one that produces the most favourable balance of good over evil, everyone

# 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?













The categorical imperative, Kantian ethics



The categorical imperative, Kantian ethics



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"



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:



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)



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



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



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



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



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



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









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







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









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





The Covid virus does not exist and here is why



Wilt u meer van ons horen? Meedenken en meepraten?

Scan dan de onderstaande QR-code of like onze







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

How do we use operationalise moral theories?









	Removal
	Warning
	No action*
	Severe
larm	



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

	Removal
	Warning
	No action*
	Severe
arm	







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

	Removal
	Warning
	No action*
	Severe
arm	







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

	Removal
	Warning
	No action*
	Severe
larm	

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

- 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?





- In engineering: a problem is iteratively c 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.

In engineering: a problem is iteratively divided into smaller problems until a problem small

eory, how can we implement it? ne agent to follow a given moral theory? a given moral theory to build an agent.

- For implicit agents: how can we follow a given moral theory to build an agent. lacksquare
- involved in their design and use
- Implicit AMA: constrain the machine's actions to avoid unethical outcomes.  $\bullet$

Operational morality: the moral significance of their actions lies entirely in the humans

- 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.

- 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.

- 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  $\bullet$ 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  $\bullet$ 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.

- 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

- 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.

- 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.

- 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.

- 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.)



- 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 <u>https://en.wikipedia.org/wiki/Tay (bot)</u>
- 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 <u>https://en.wikipedia.org/wiki/Tay (bot)</u>
- 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

• Requires hard AI problems to be solved: situational awareness, prediction of consequences, learning from

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

Full-text available Article

### 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



-	÷	~
-	Т	ρ
-	•	~

**Ethics and Information Technology** June 2016, Volume 18, <u>Issue 2</u>, pp 103–115 | <u>Cite as</u>

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



### More material



Original Research/Scholarship Published: 08 July 2020

Landscape of Machine Implemented Ethics

Vivek Nallur

Science and Engineering Ethics 26, 2381–2399(2020) Cite this article

http://slavkovik.com/ijcaitutorial2020.html https://www.youtube.com/watch?v=H7n1W8J1vWo

RESEARCH-ARTICLE OPEN ACCESS

### **Implementations in Machine Ethics: A Survey**



Publication: ACM Computing Surveys • December 2020 • Article No.: 132 • https://doi.org/10.1145/3419633



# Why so little machine learning

- So far used: symbolic learning lacksquare
- Reinforcement learning: requires simulations  $\bullet$
- Supervised learning: requires labeled examples+  $\bullet$