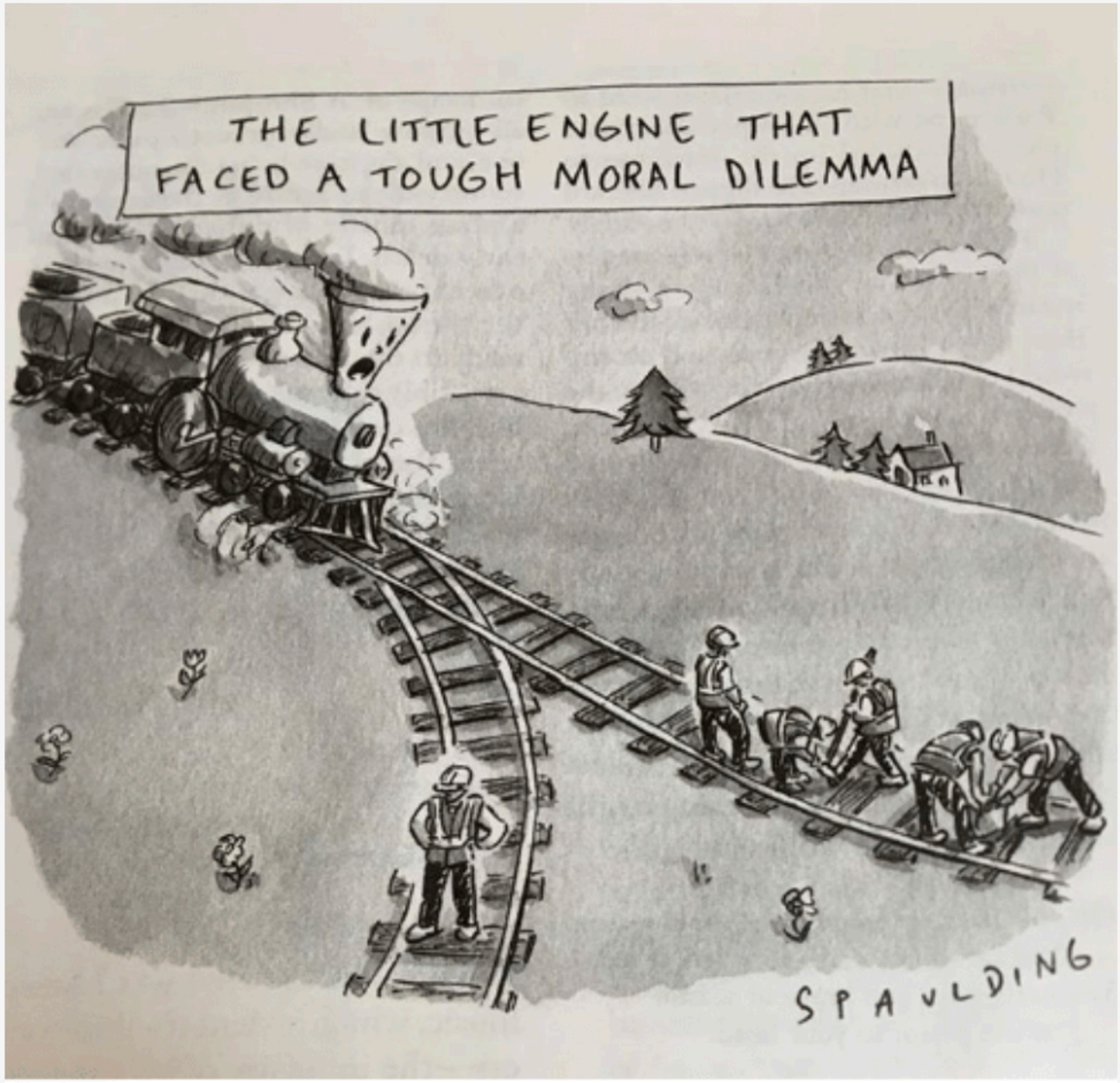
MACHINES THAT KNOW RIGHT, AND CAN NOT DO WRONG

THE THEORY AND PRACTICE OF MACHINE ETHICS

LOUISE DENNIS & MARIJA SLAVKOVIK

IJCAI-ECAI 2018 TUTORIAL

Machine ethics



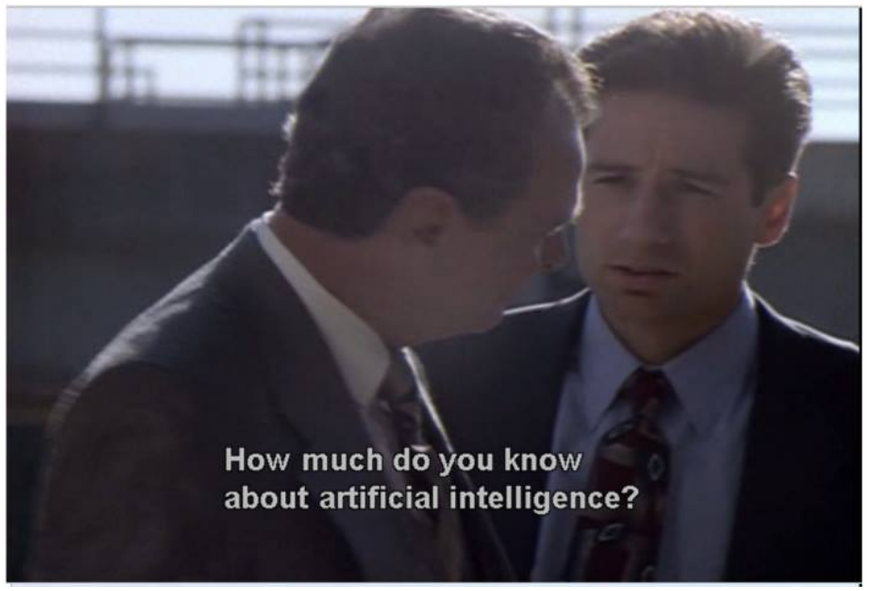
The New Yorker just made a Trolley Problem Meme! What a time to be alive. o.O

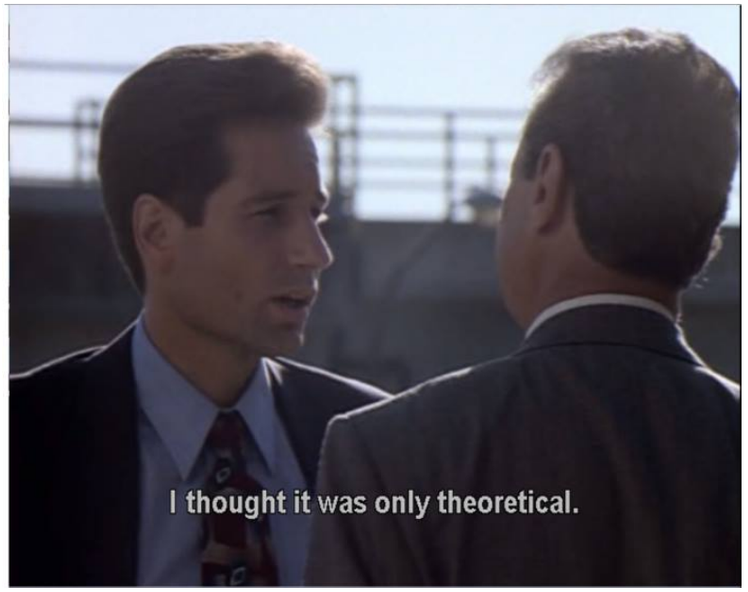
Machine ethics

Machine ethics

How to enable autonomous and/or intelligent systems to not violate the ethical norms of the environment they occupy?

Why now?


How much do you know
about artificial intelligence?

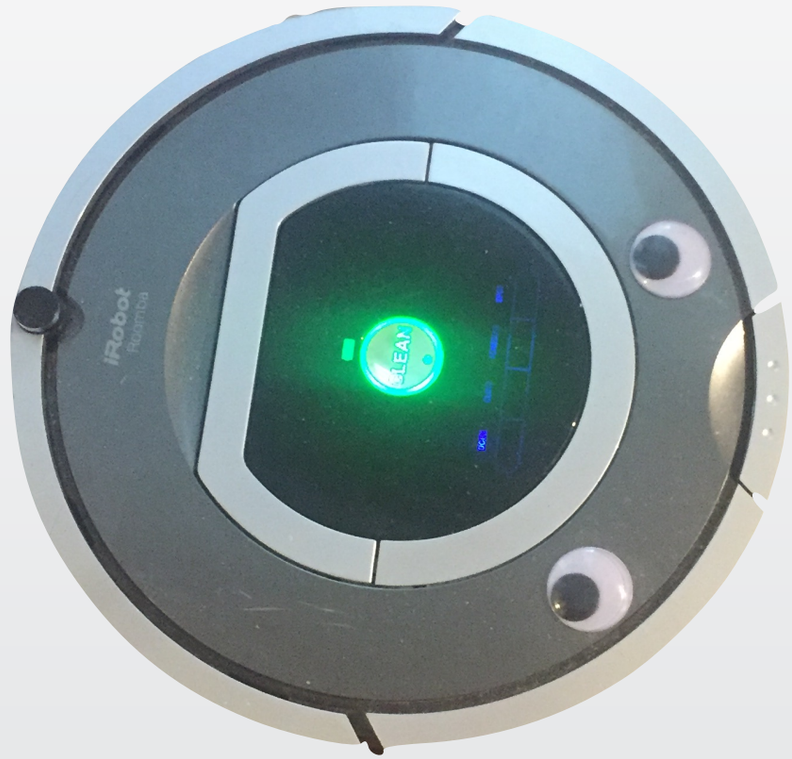

I thought it was only theoretical.


It was... until two years ago.

Why now?

Why now?

Why now?




SEAT
RENAULT
NISSAN BOUTIQUE
NISSAN
Autonomous Drive

Avgang Departures 10 48

10:47	L2	Stabekk	8		NSB
10:49	R10	Drammen	5		NSB
10:50		Oslo Lufthavn	13		Flytoget
10:54	L12	Eidsvoll	11		NSB
10:56	L1	Lillestrøm	9		NSB
10:57	L1	Asker	7		NSB
10:59	L14	Asker	5		NSB
11:00		Oslo Lufthavn	13		Flytoget
11:01	R20	Halden	18		NSB
11:02	R30	Gjøvik			NSB Gjøvikbanen
11:05	L22	Sk			
11:09	L12	Ko			

Avgang Departures 10:48

11:10		Oslo Lufthavn	13		Flytoget
11:11	L1	Spikkestad	7		NSB
11:11	L1	Lillestrøm	9		NSB
11:14		Dal	11		NSB
11:17	L2	Stabekk	8		NSB
11:18		Moss	9		NSB
11:19		Drammen	5		NSB
11:20		Oslo Lufthavn	13		Flytoget
11:21	L2	Stabekk	7		NSB
11:24	R11	Eidsvoll	11		NSB
11:27	L1	Asker			
11:29	R10	Drammen			

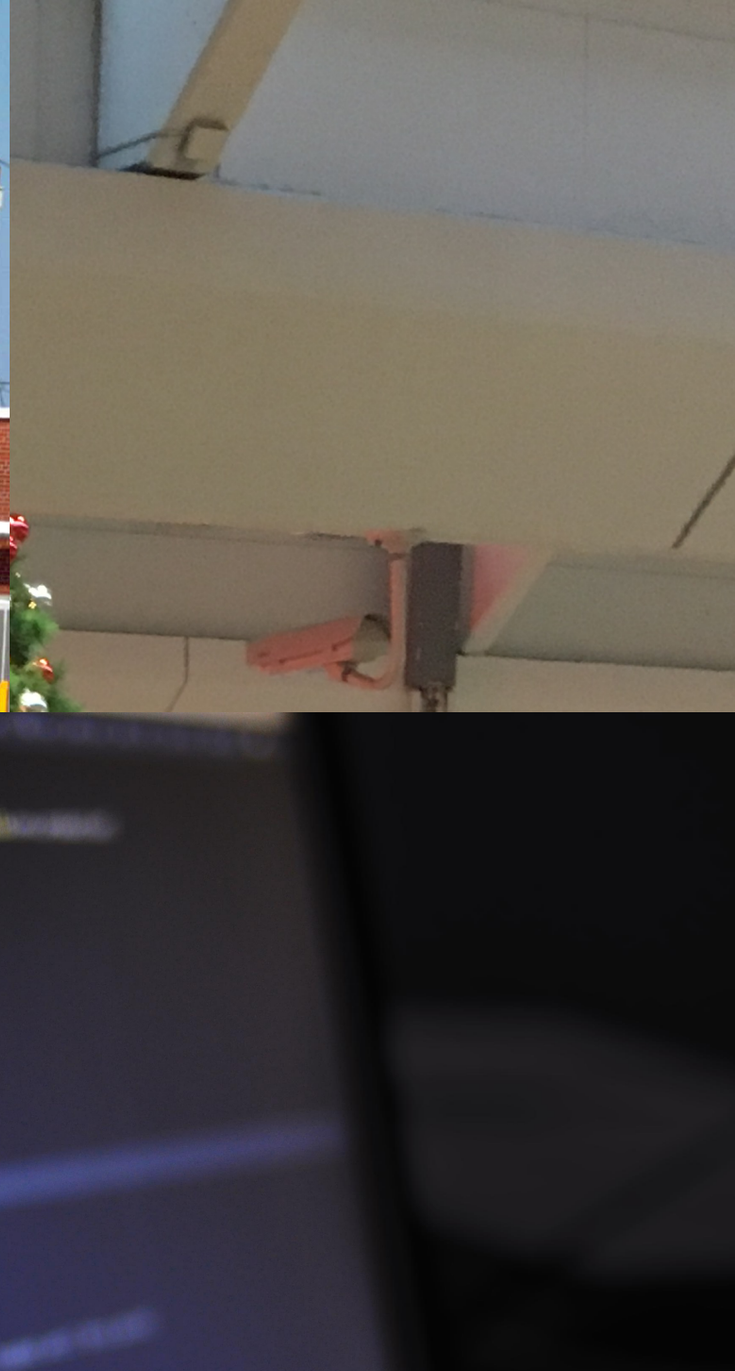
Ankomst Arrivals 10 48

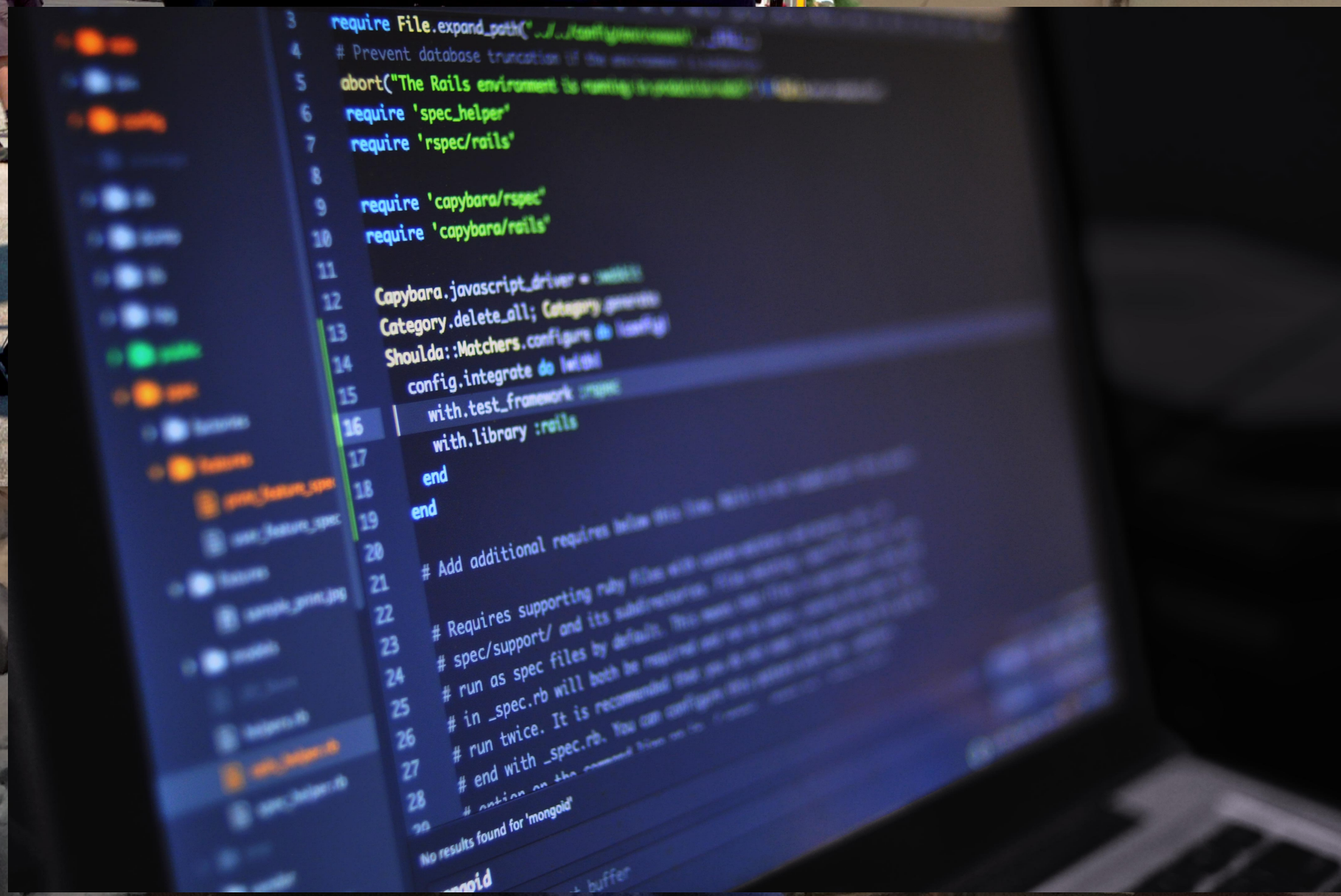
10:39		Oslo Lufthavn	14	
10:45	L2	Stabekk	8	10:48
10:46		Dal	11	
10:47		Stabekk	13	
10:50	L1	Asker	14	10:57
10:51	L1	Spikkestad	11	
10:52	L1	Asker	10	
10:52		Oslo Lufthavn	2	
10:53	L1	Asker	9	
10:54		Oslo Lufthavn	2	
10:59		Oslo Lufthavn	2	
11:01	L1	Asker	11	

9 10


Ankomst Arrivals
10 48


GIVE




require File.expand_path(
Prevent database
abort("The Rails
require 'spec_helper'
require 'rspec/rails'
require 'capybara/rspec'
require 'capybara/rails'
Capybara.javascript_driver
Category.delete_all;
Shoulda::Matchers.
config.integrate
with.test_framework
with.library :rails
end
end
Add additional
Requires supporting
spec/support/ and its
run as spec files by
in _spec.rb will
run twice. It is
end with _spec.rb.
No results found for 'mongoid'

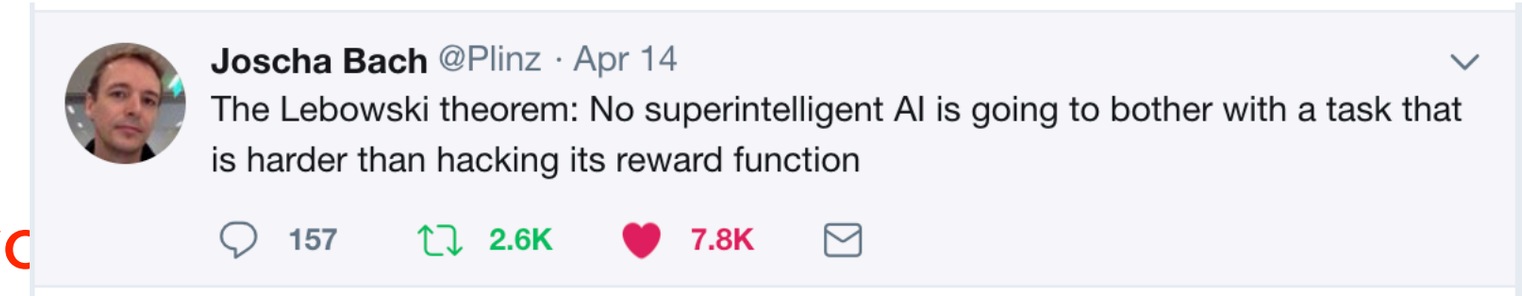
Machine ethics challenges

Machine ethics challenges

- What happens when you generalise from human agent in moral theories?

Machine ethics challenges

- What happens when yo
moral theories?



Machine ethics challenges

- What happens when you generalise from human agent in moral theories?

Machine ethics challenges

- What happens when you generalise from human agent in moral theories?
- What is a good moral theory for artificial agents?

Machine ethics challenges

- W
- W

Sacrifice One For the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents

Full Text: PDF

Authors:
Bertram F. Malle Brown University, Providence, RI, USA
Matthias Scheutz Tufts University, Medford, MA, USA
Thomas Arnold Harvard University, Cambridge, MA, USA
John Voiklis Brown University, Providence, RI, USA
Corey Cusimano Brown University, Providence, RI, USA

2015 Article

Bibliometrics
· Citation Count: 7
· Downloads (cumulative): 796
· Downloads (12 Months): 200
· Downloads (6 Weeks): 21

Published in:

· Proceeding
HRI '15 Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction
Pages 117-124

Portland, Oregon, USA — March 02 - 05, 2015
ACM New York, NY, USA ©2015
table of contents ISBN: 978-1-4503-2883-8 doi>10.1145/2696454.2696458

Machine ethics challenges

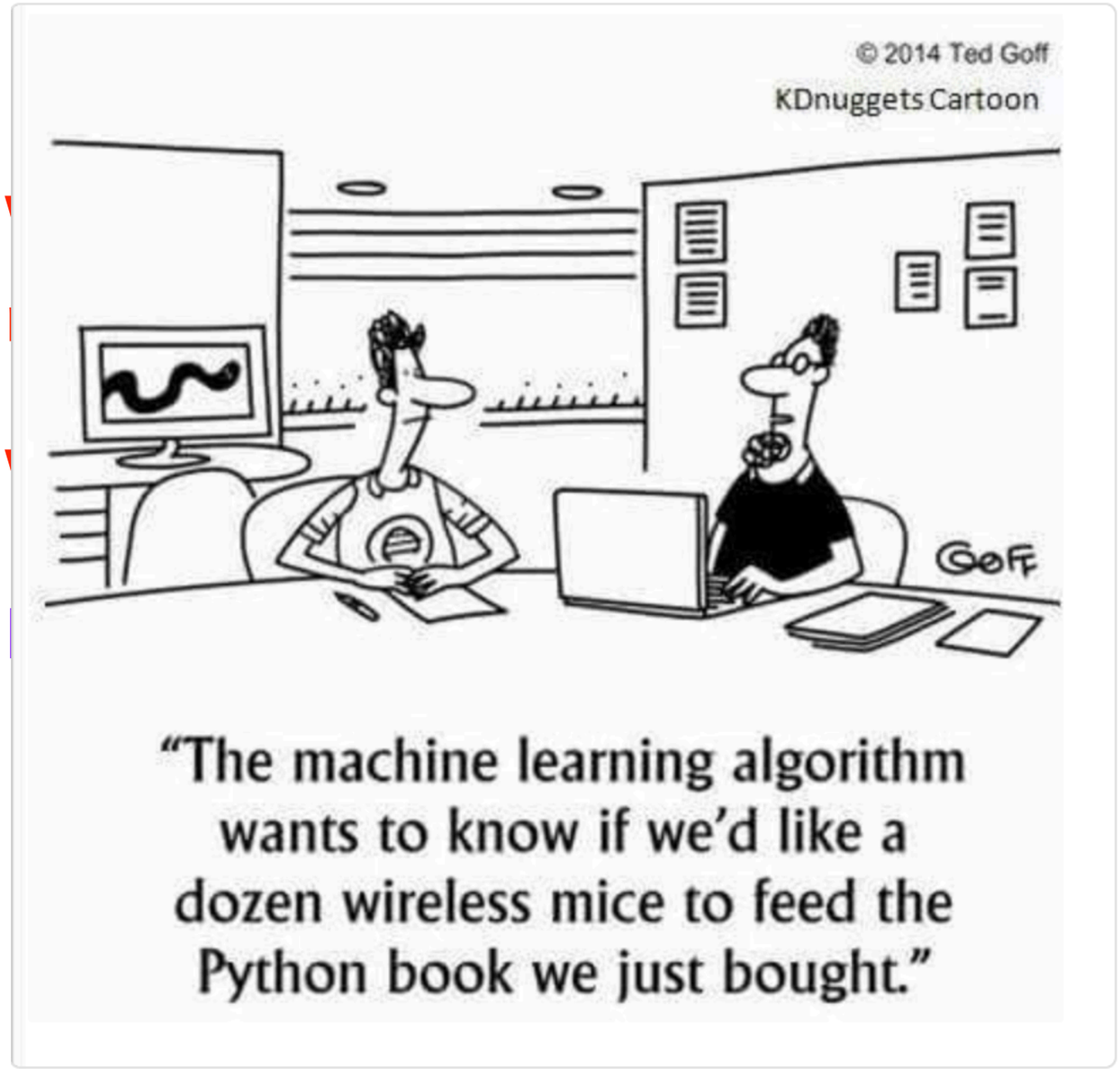
- What happens when you generalise from human agent in moral theories?
- What is a good moral theory for artificial agents?

Machine ethics challenges

- What happens when you generalise from human agent in moral theories?
- What is a good moral theory for artificial agents?
- How to formalise common sense?

Machine ethics challenges

- V e from human agent in
- V tificial agents?
-



Machine ethics challenges

- What happens when you generalise from human agent in moral theories?
- What is a good moral theory for artificial agents?
- How to formalise common sense?

Machine ethics challenges

- What happens when you generalise from human agent in moral theories?
- What is a good moral theory for artificial agents?
- How to formalise common sense?
- Given a moral theory, how do we implement moral reasoning and decision making?

Machine ethics challenges

- What happens when you generalise from human agent in moral theories?
- What is a good moral theory for artificial agents?
- How to formalise common sense?
- Given a moral theory, how do we implement moral reasoning and decision making?
- How do we do it so that we can verify and certify moral behaviour before a product is deployed?

Liability as a function of ability

Liability as a function of ability

- In Rome, citizens and barbarians were held differently liable for the same crime

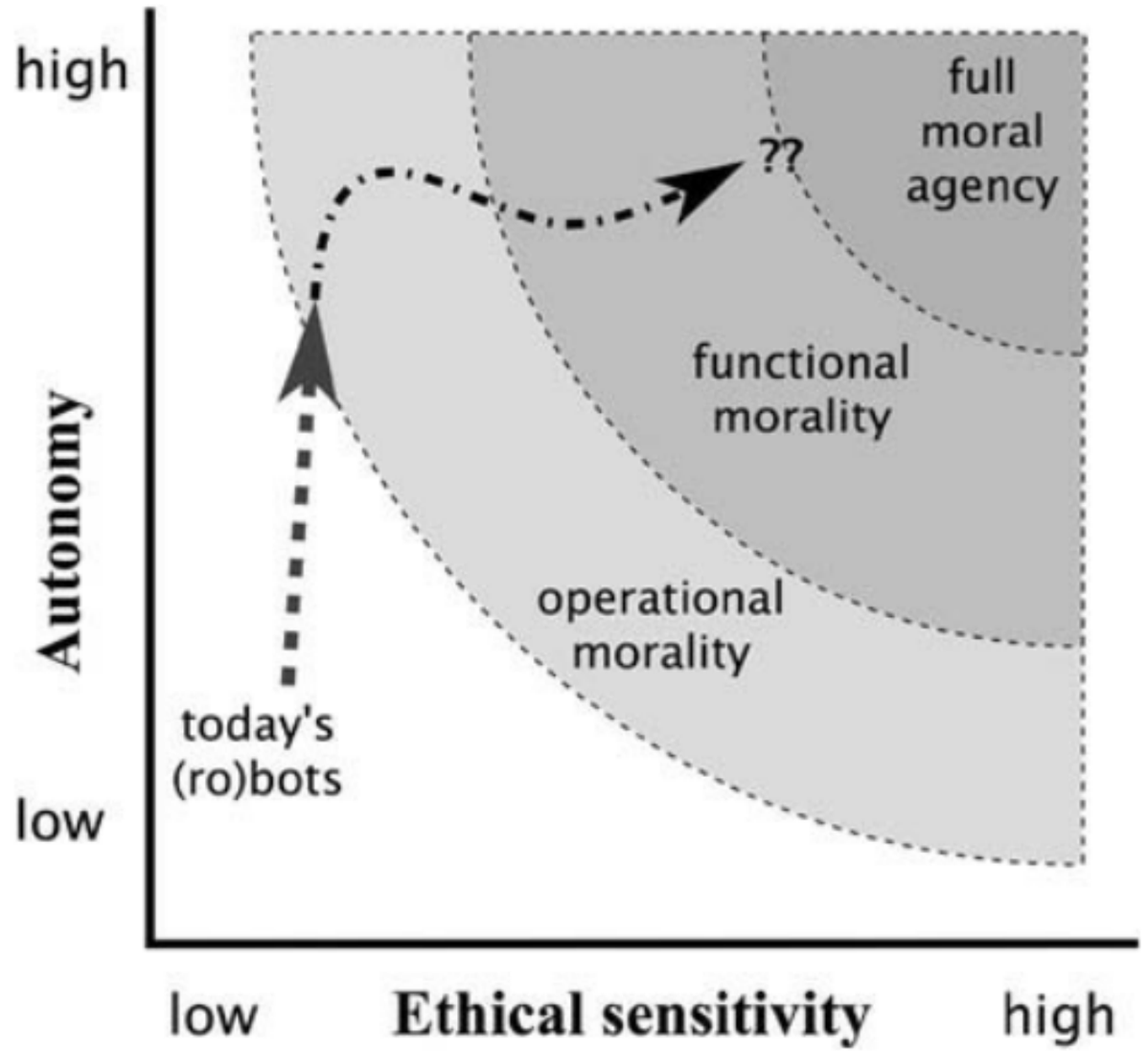
Liability as a function of ability

- In Rome, citizens and barbarians were held differently liable for the same crime
- Today the law distinguishes between crimes committed by children and by able-minded adults

Liability as a function of ability

- In Rome, citizens and barbarians were held differently liable for the same crime
- Today the law distinguishes between crimes committed by children and by able-minded adults
- The question is how much ethical sensitivity should we expect from an agent with a given ability?

Wallach and Allen in Moral



Moor [2006]

Moor [2006]

- Ethical-impact agents

Moor [2006]

- Ethical-impact agents
- Implicit ethical agents

Moor [2006]

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

Moor [2006]

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

Moor [2006]

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

Dyrkolbotn, Pedersen & Slavkovik [2018]

Dyrkolbotn, Pedersen & Slavkovik [2018]

- Is the agent relying on its ability to make decisions autonomously for fulfilling ethical objectives?

Dyrkolbotn, Pedersen & Slavkovik [2018]

- Is the agent relying on its ability to make decisions autonomously for fulfilling ethical objectives?
- If an implicit ethical agent violates an ethical expectation that it is supposed to satisfy, this is evidence of a defect

Dyrkolbotn, Pedersen & Slavkovik [2018]

- Is the agent relying on its ability to make decisions autonomously for fulfilling ethical objectives?

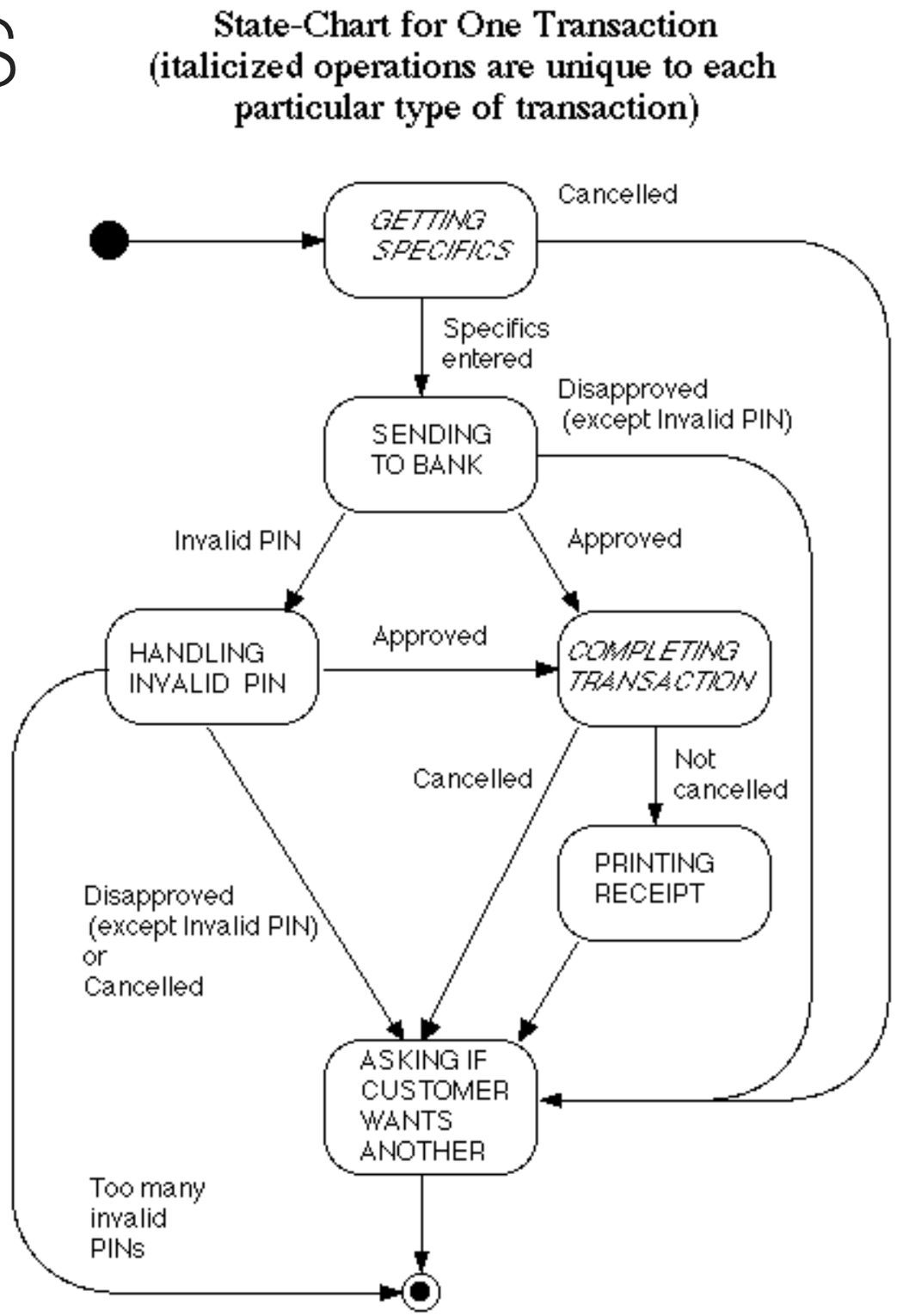
- If an implicit ethical agent violates an ethical expectation that it is supposed to satisfy, this is evidence of a defect

- If the agent can makes ethical decisions explicitly it can both **not** satisfy an ethical expectation and **not** be in defect

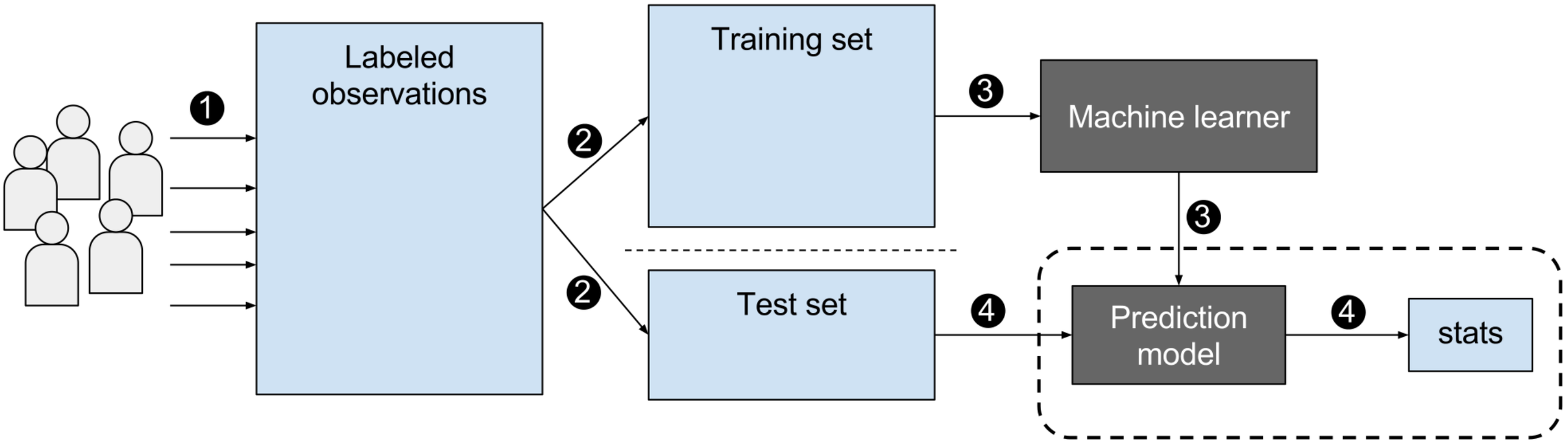
Not all equi-autonomous AI are made the same . . .

Rule-based methods

```
% A solution to the sudoku problem.

#maxint=9.

tab(X,Y,1) v tab(X,Y,2) v tab(X,Y,3) v tab(X,Y,4) v tab(X,Y,5) v
tab(X,Y,6) v tab(X,Y,7) v tab(X,Y,8) v tab(X,Y,9) :- #int(X), 0 <=
X, X <= 8, #int(Y), 0 <= Y, Y <= 8.

% Check rows and columns
:- tab(X,Y1,Z), tab(X,Y2,Z), Y1<>Y2.
:- tab(X1,Y,Z), tab(X2,Y,Z), X1<>X2.

% Check subtable
:- tab(X1,Y1,Z), tab(X2,Y2,Z), Y1 <> Y2,
   div(X1,3,W1), div(X2,3,W1),
   div(Y1,3,W2), div(Y2,3,W2).

:- tab(X1,Y1,Z), tab(X2,Y2,Z), X1 <> X2,
   div(X1,3,W1), div(X2,3,W1),
   div(Y1,3,W2), div(Y2,3,W2).

%Auxiliary: X divided by Y is Z
div(X,Y,Z) :- XminusDelta = Y*Z, X = XminusDelta + Delta, Delta < Y.


% Table positions  X=0..8, Y=0..8

tab(0,1,6). tab(0,3,1). tab(0,5,4). tab(0,7,5).
```

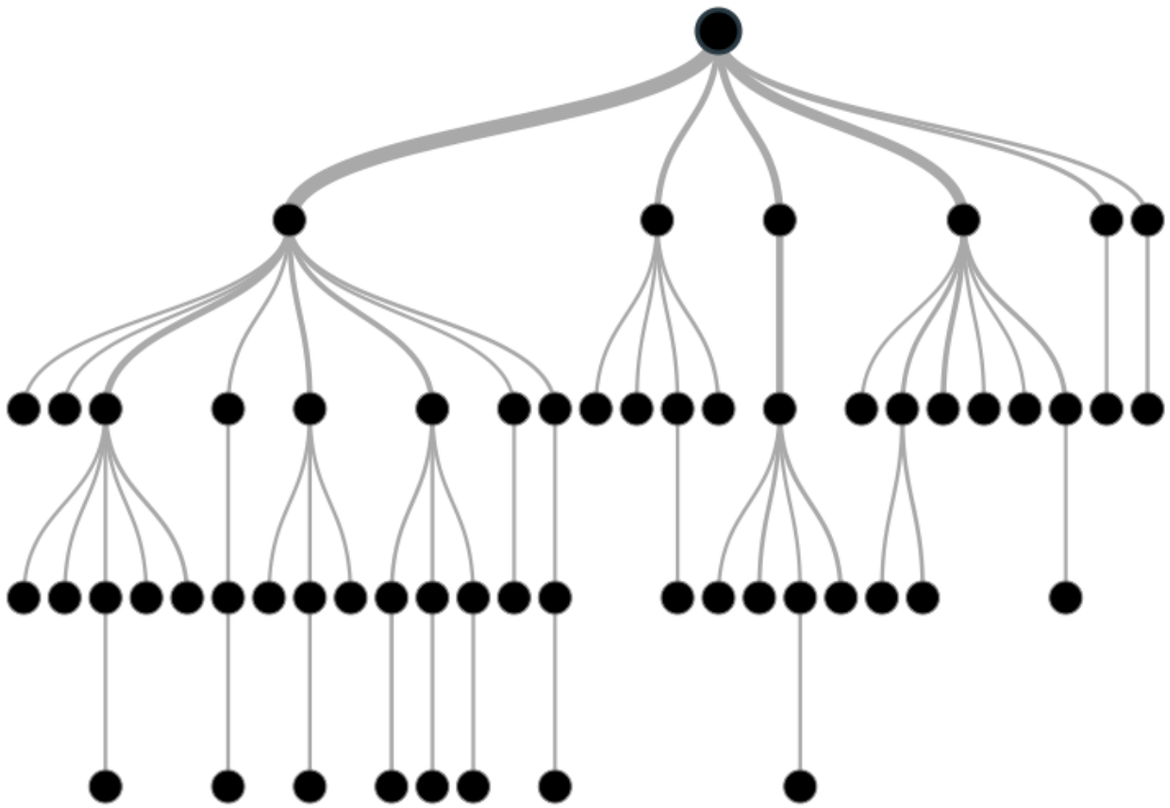


Statistical-based methods

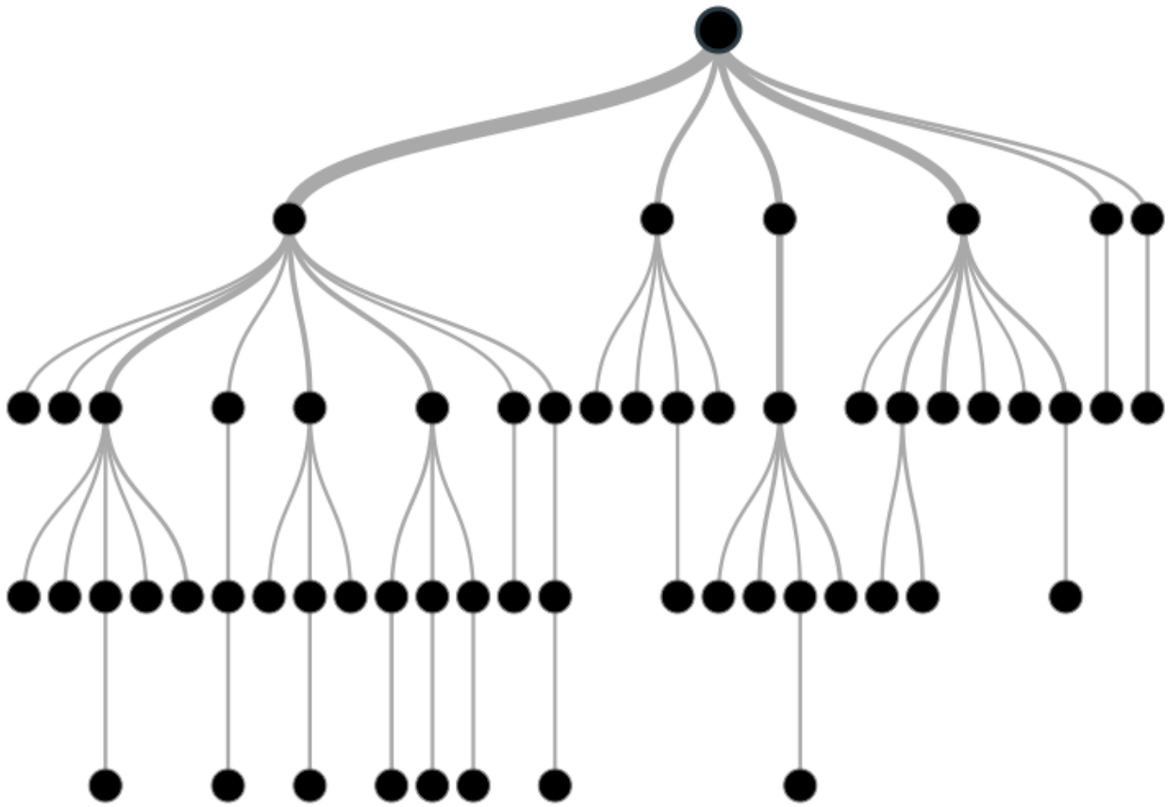


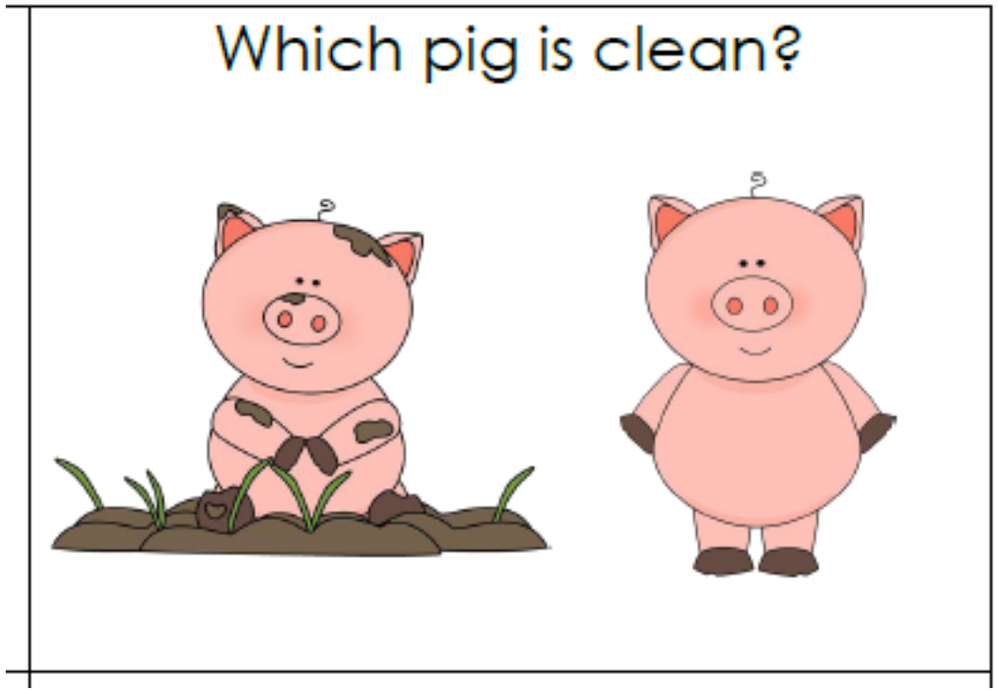
Top-down vs Bottom-up

Top-down vs Bottom-up



Top-down vs Bottom-up





Top-down

Top-down

Download PDF Export

Cognitive Systems Research

Available online 22 May 2017

In Press, Corrected Proof

An architecture for ethical robots inspired by the simulation theory of cognition

Dieter Vanderelst, Alan Winfield

Show more

https://doi.org/10.1016/j.cogsys.2017.04.002

Top-down

Top-down

The Hybrid Ethical Reasoning Agent IMMANUEL

Full Text: PDF

Authors: Felix Lindner — University of Freiburg, Freiburg, Germany
Martin Mose Bentzen — Danish Technical University, Lyngby, Denmark

Published in:

· Proceeding
HRI '17 Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction
Pages 187-188

Vienna, Austria — March 06 - 09, 2017
ACM New York, NY, USA ©2017
table of contents ISBN: 978-1-4503-4885-0
doi>10.1145/3029798.3038404

Top-down

Top-down

Download PDF Export

Robotics and Autonomous Systems

Volume 77, March 2016, Pages 1-14

Formal verification of ethical choices in autonomous systems

… Louise Dennis [a] ✉, Michael Fisher [a] ✉, Marija Slavkovik [b] ✉, Matt Webster [a] ✉

Show more

https://doi.org/10.1016/j.robot.2015.11.012

Get rights and content

Open Access funded by Engineering and Physical Sciences Research Council

Top-down

Top-down

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

Toward a General Logicist Methodology for Engineering Ethically Correct Robots

3
Author(s)

S. Bringsjord ; K. Arkoudas ; P. Bello

Top-down

Top-down

Bottom-up

Bottom-up

AAAI Publications, Workshops at the Thirtieth AAAI Conference on Artificial Intelligence

Ensuring Ethical Behavior from Autonomous Systems

Michael Anderson, Susan Anderson, Vincent Berenz

Last modified: 2016-03-29

Abstract

We advocate a case-supported principle-based behavior paradigm coupled with the Fractal robot architecture as a means to control an eldercare robot. The most ethically preferable action at any given moment is determined using a principle, abstracted from cases where a consensus of ethicists exists.

Keywords

machine ethics; application

Full Text: PDF

Bottom-up

Bottom-up

The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence
AI, Ethics, and Society: Technical Report WS-16-02

Reinforcement Learning as a Framework for Ethical Decision Making

David Abel and **James MacGlashan** and **Michael L. Littman**
Brown University, Computer Science Department
115 Waterman Street
Providence, RI 02912-1910

Bottom-up

Bottom-up

Browse Conferences > Computational Intelligence in...

An artificial neural network approach for creating an ethical artificial agent

Sign In or Purchase to View Full Text

6 Paper Citations

310 Full Text Views

Related Articles

Global path planning using artificial potential fields

Technology challenges for building Internet-scale ubiquitous computing

View All

2 Author(s)

Ali Reza Honarvar ; Nasser Ghasem-Aghaee

View All Authors

Abstract | Authors | Figures | References | Citations | Keywords | Metrics | Media

Abstract:
Autonomous robotic systems and intelligent artificial agents' capability have advanced dramatically. Since the intelligent artificial agents have been developing more autonomous and human-like, the capability of them to make moral decisions becomes an important issue. In this work we developed an artificial neutral network which considered various effective factors for ethical assessment of an action to determine that if a behavior or an action is ethically permissible or not. We integrated this net to the BDI-Agent model as a part of its reasoning process to behave ethically in various environments.

Published in: Computational Intelligence in Robotics and Automation (CIRA), 2009 IEEE International Symposium on

Validation

Validation

- the process of confirming that the final system has the intended behaviour once it is active in its target environment

Validation

- the process of confirming that the final system has the intended behaviour once it is active in its target environment
- done for external stake-holders

Validation

- the process of confirming that the final system has the intended behaviour once it is active in its target environment
- done for external stake-holders
- assessment of accuracy, repeatability, trust, usability, resilience, etc.

Different roles of certification

Different roles of certification

- For users: allocate the right amount of trust

Different roles of certification

- For users: allocate the right amount of trust
  - a machine may have the appearance of “experienced”, “benevolent”, “sympathetic” without being any of these things

Different roles of certification

- For users: allocate the right amount of trust
  - a machine may have the appearance of “experienced”, “benevolent”, “sympathetic” without being any of these things
  - children and elderly around robots

Different roles of certification

- For users: allocate the right amount of trust
  - a machine may have the appearance of “experienced”, “benevolent”, “sympathetic” without being any of these things
  - children and elderly around robots
- For company: clear liability distribution

Different roles of certification

- For users: allocate the right amount of trust
  - a machine may have the appearance of “experienced”, “benevolent”, “sympathetic” without being any of these things
  - children and elderly around robots
- For company: clear liability distribution
- For regulators: promote welfare in society

Different roles of certification

- For users: allocate the right amount of trust
  - a machine may have the appearance of “experienced”, “benevolent”, “sympathetic” without being any of these things
  - children and elderly around robots
- For company: clear liability distribution
- For regulators: promote welfare in society
  - BS8611 - issues over which ethical issues should be considered

Who decides what is ethical?

Who decides what is ethical?

- User

W

Robohub

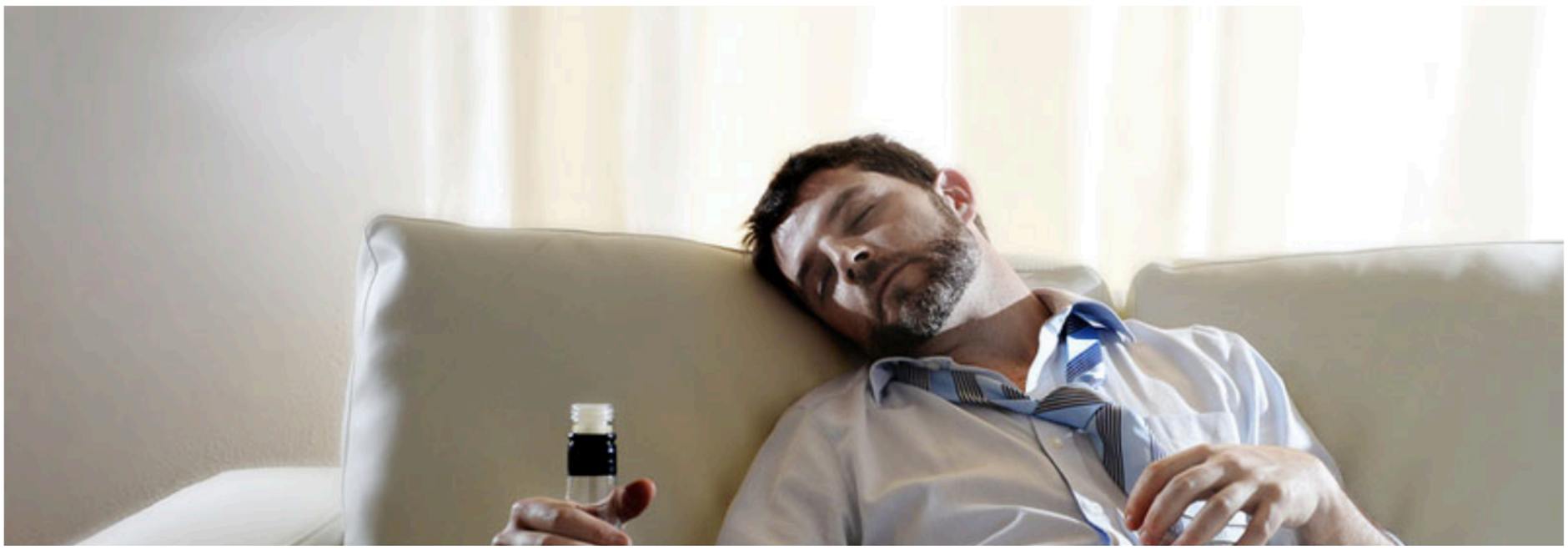
news views podcast learn

- Like Tweet LinkedIn View Comments

Should a care robot bring an alcoholic a drink? It depends on who owns the robot

by Open Roboethics Initiative

Politics-Law-Society Reader Poll February 24, 2015



↑ up

Who decides what is ethical?

- User

Who decides what is ethical?

- User
- Public opinion, society

Who decides what is ethical?

- User
- Public opinion, society
- Manufacturer

Who decides what is ethical?

- User
- Public opinion, society
- Manufacturer
- Government mandated body

Issues with governments deciding

Issues with governments deciding

- Law is slow

Issues with governments deciding

- Law is slow
- One company produces for many markets

Issues with governments deciding

- Law is slow
- One company produces for many markets
- What is not illegal is not always ethical

Issues with governments deciding

- Law is slow
- One company produces for many markets
- What is not illegal is not always ethical
- What is good enough for people is not good enough for machines

Issues with governments deciding

- Law is slow
- One company produces for many markets
- What is not illegal is not always ethical
- What is good enough for people is not good enough for machines
- How many accidents is too many accidents?

Issues with governments deciding

- Law is slow
- One company produces for many markets
- What is not illegal is not always ethical
- What is good enough for people is not good enough for machines
- How many accidents is too many accidents?
  - Eg. aviation: no. accidents per miles flown.

Issues with governments deciding

- Law is slow
- One company produces for many markets
- What is not illegal is not always ethical
- What is good enough for people is not good enough for machines
- How many accidents is too many accidents?
  - Eg. aviation: no. accidents per miles flown.
  - Autonomous system: no. of persons affected?

Tackling the XAI

Tackling the XAI

- Statistical based methods are used for problems that cannot be explicitly procedurally specified or computed

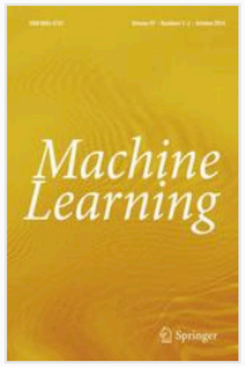
Tackling the XAI

- Statistical based methods are used for problems that cannot be explicitly procedurally specified or computed
- When will it fail?

Tackling the XAI

- Statistical based methods are used for problems that cannot be explicitly procedurally specified or computed
- When will it fail?
- Why will it fail?

Tackling the XAI



Machine Learning
July 1997, Volume 28, Issue 1, pp 41–75 | Cite as

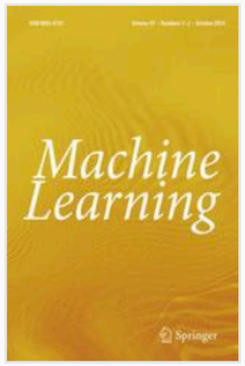
Multitask Learning

Authors | Authors and affiliations

Rich Caruana

Article

Tackling the XAI



Machine Learning
July 1997, Volume 28, Issue 1, pp 41–75 | Cite as

Multitask Learning

Authors | Authors and affiliations

Rich Caruana

Article

2 Shares | 6.7k Downloads | 713 Citations

https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html?_r=1

Ethical Turing test

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

Ethical Turing test

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 Researc

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