

## HIM Lecture 2

# What is AI? Can It Be Accountable?

Joanna J. Bryson

University of Bath, United Kingdom

@j2bryson

# Outline

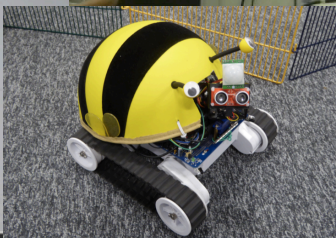
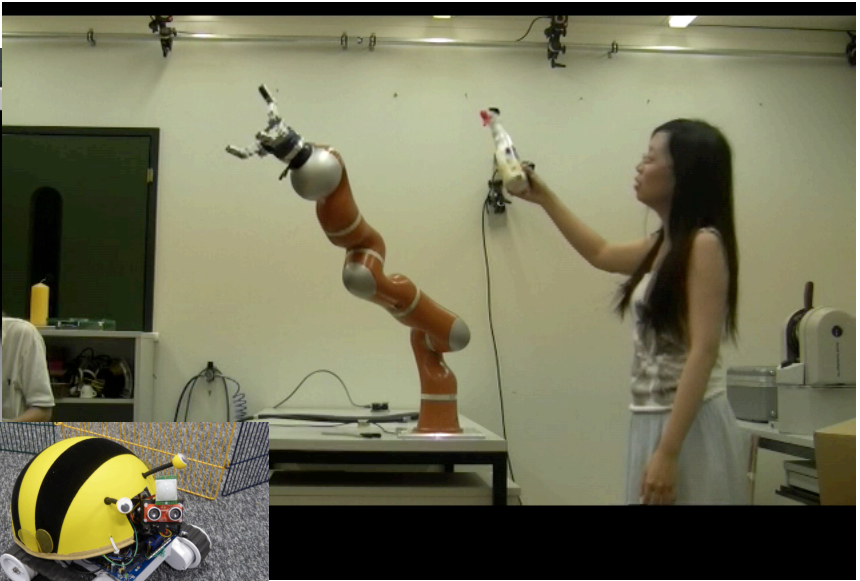
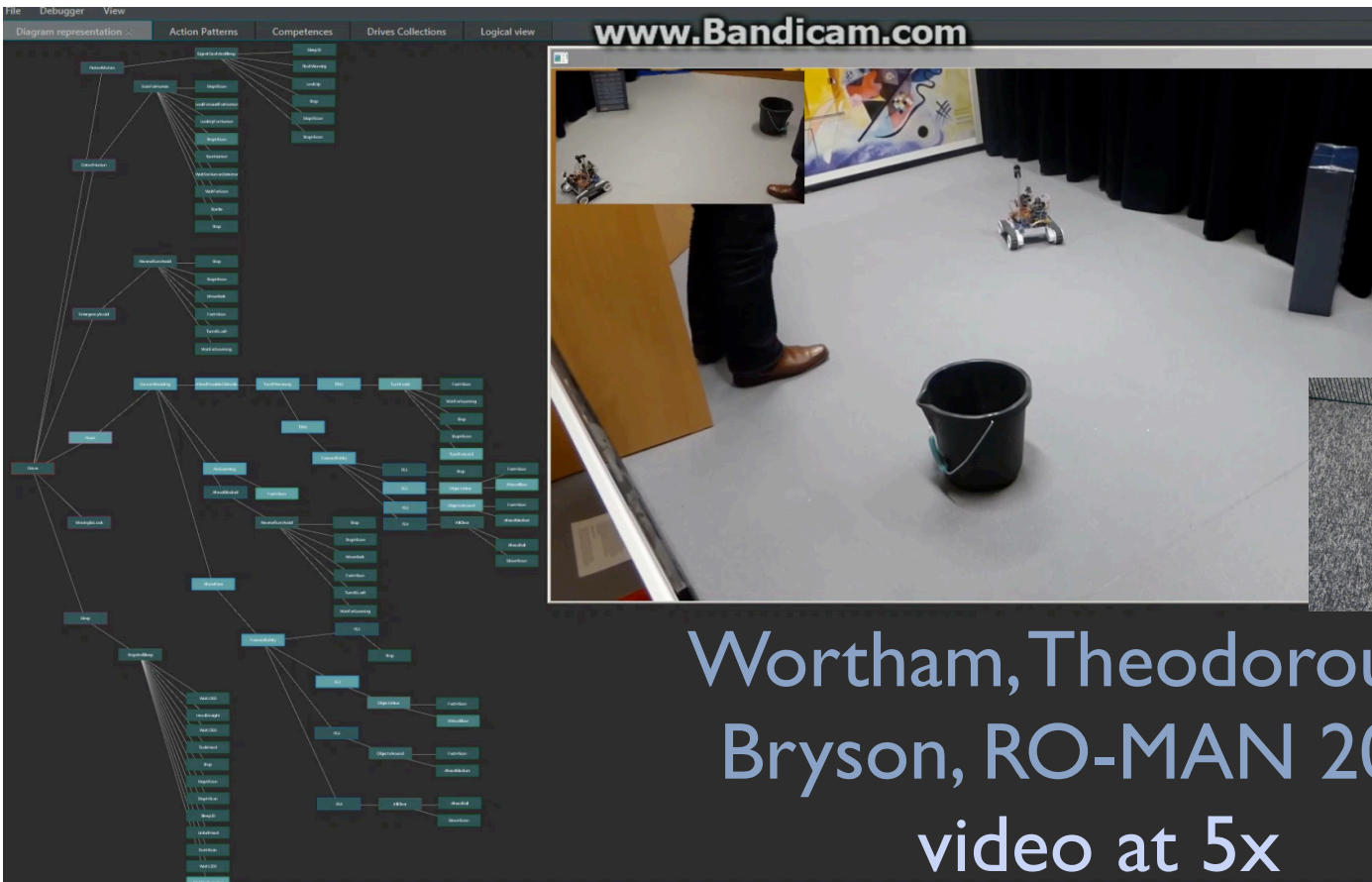
- Who I am
- What is AI? Terms and Concepts
- Transparency & Accountability in Machine Learning (ML), and (AI)

# Me (including finding me)

- Two degrees in psychology (Chicago 1986, Edinburgh 2000) two in AI (Edinburgh 1992, MIT 2001).
- Lecturer in Bath 2002-2009, Reader 2010-2020. Professor at Hertie, Berlin from February 2020.
- Lived in Vienna 2007-2009, New Jersey 2015-2020.
- Sign up for office hours from my contact Web page (search for “joanna bryson contact”). **Email.**
- Check twitter most days, but sometimes I get too many mentions so you may have to try twice.

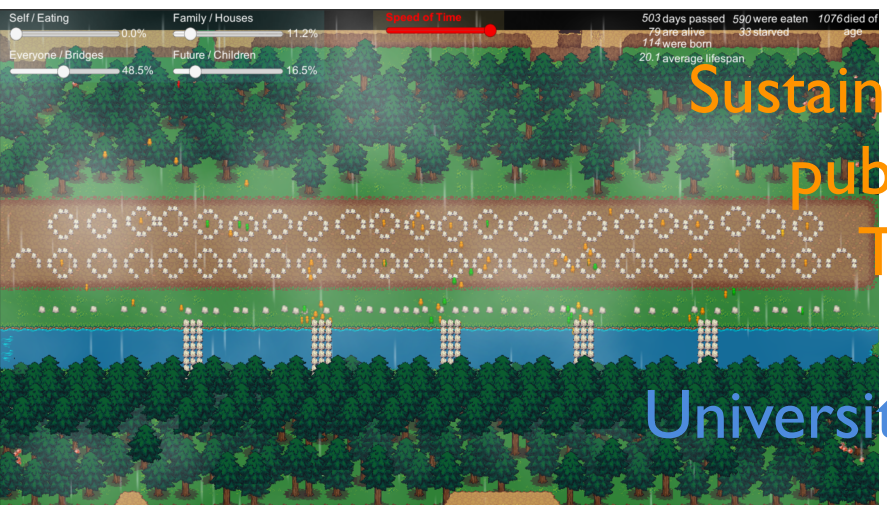




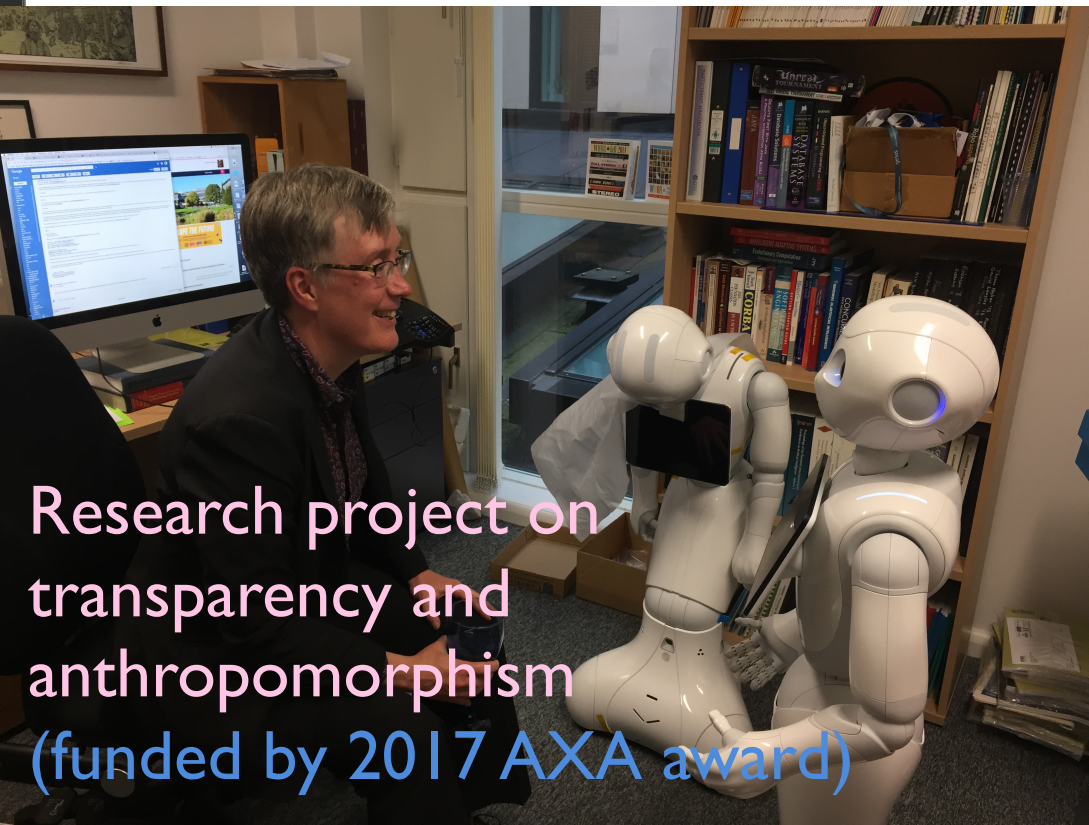


Wortham, Theodorou, & Bryson, RO-MAN 2017 video at 5x

Bidan Huang, Li, Souza, Bryson, & Aude Billard, IROS 2016.



Sustainability game, teaching public goods investment, Theodorou & Bryson, (funded by Princeton University Center for Human Values)



Research project on transparency and anthropomorphism (funded by 2017 AXA award)



# Interests

- Systems Artificial Intelligence
- Natural Intelligence

# Interests

- Systems Artificial Intelligence
  - Modularity & Coordination (Planning).
  - AI Ethics – Agency, Autonomy, Transparency.
  - Programmability for Real-Time AI – “AI Plumbers”
- Natural Intelligence
  - Modularity & Organization (Neuroscience).
  - Origins of Cognition (Behavioural Ecology).
  - Culture & Sociality (Biological Anthropology, Ethics).

# Interests

- Systems Artificial Intelligence
  - Modularity & Coordination (Planning).
  - AI Ethics – Agency, Autonomy, Transparency.
  - Programmability for Real-Time AI – “AI Plumbers”
- Natural Intelligence
  - Modularity & Organization (Neuroscience).
  - Origins of Cognition (Behavioural Ecology).
  - Culture & Sociality (Biological Anthropology, Ethics).



# Current Work

- Artificial Intelligence
  - Intelligent systems development. (e.g. domestic robots, game characters, intelligent environments.)
  - Transparency, Accountability, Policy.
- Natural Intelligence
  - Evolution, cognition and primate sociality.
  - Evolution, language/culture, public goods investment, political polarisation, political economy.
- Policy

# Outline

- Who I am
- What is AI? Terms and Concepts
- Transparency & Accountability in Machine Learning (ML), and (AI)

# Definitions

for communicating  
right now

- **Intelligence** is doing the right thing at the right time (in a dynamic environment).
- **Agents** are any vector of change,
  - e.g. chemical agents.
- **Moral agents** are considered responsible for their actions by a society.
- **Moral patients** are considered the responsibility of a society's agents.
- **Artificial Intelligence** is intelligence deliberately built.

Arguably, ethics is determined by and determines a society—a constantly renegotiated set of equilibria.

Regulation is a part of ethics by this def.

Basic regulatory question: Is there anything about this technology that changes legal responsibility for that intentional act?



Intelligence relies on computation, not math.

Computation is a physical process, taking time, energy, & space.

Finding the right thing to do at the right time requires search.

Cost of search = # of options<sup># of acts</sup> (serial computing).

Examples:

- Any 2 of 100 possible actions =  $100^2 = 10,000$  possible plans.
- # of 35-move games of chess > # of atoms in the universe.

Concurrency can save real time, but not energy, and requires more space. Quantum saves on space (sometimes) but not energy(?)

Omniscience (“AGI”) is not a real threat. No one algorithm can solve all of AI.

Dr. Viv Kendon, quantum physicist, Durham University



Humanity's winning (ecological)  
strategy exploits concurrency –  
we share what we know, mining  
others' prior search.

Now we do this with machine learning.



AI is already “super-human” at chess, go, speech transcription, lip reading, deception detection from posture, forging voices, handwriting, & video, general knowledge and memory.

This spectacular recent growth derives from using ML to exploit the discoveries (previous computation) of biological evolution and human culture.

Pace of improvement will slow as AI joins the (now accelerating) frontier of our knowledge.



# One Consequence

## AI Is Not Necessarily Better than We Are



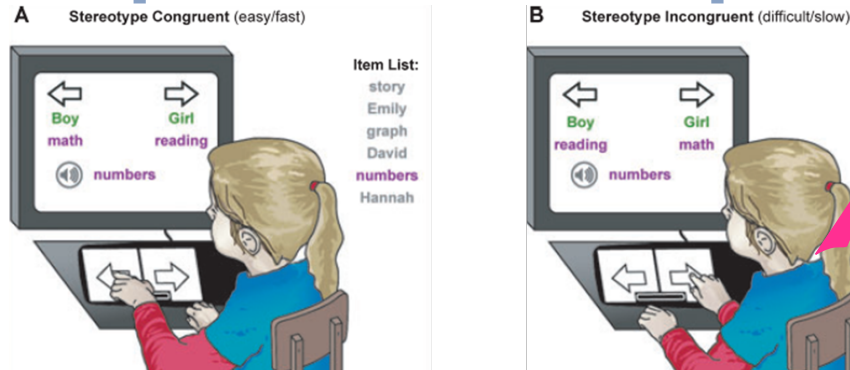
**Semantics derived automatically from language corpora contain human-like biases**

Aylin Caliskan, Joanna J. Bryson and Arvind Narayanan (April 13, 2017)

*Science* **356** (6334), 183-186. [doi: 10.1126/science.aal4230]

# AI Trained on Human Language Replicates Implicit Biases

Caliskan, Bryson & Narayanan  
(*Science*, April 2017)



Gender bias [stereotype]

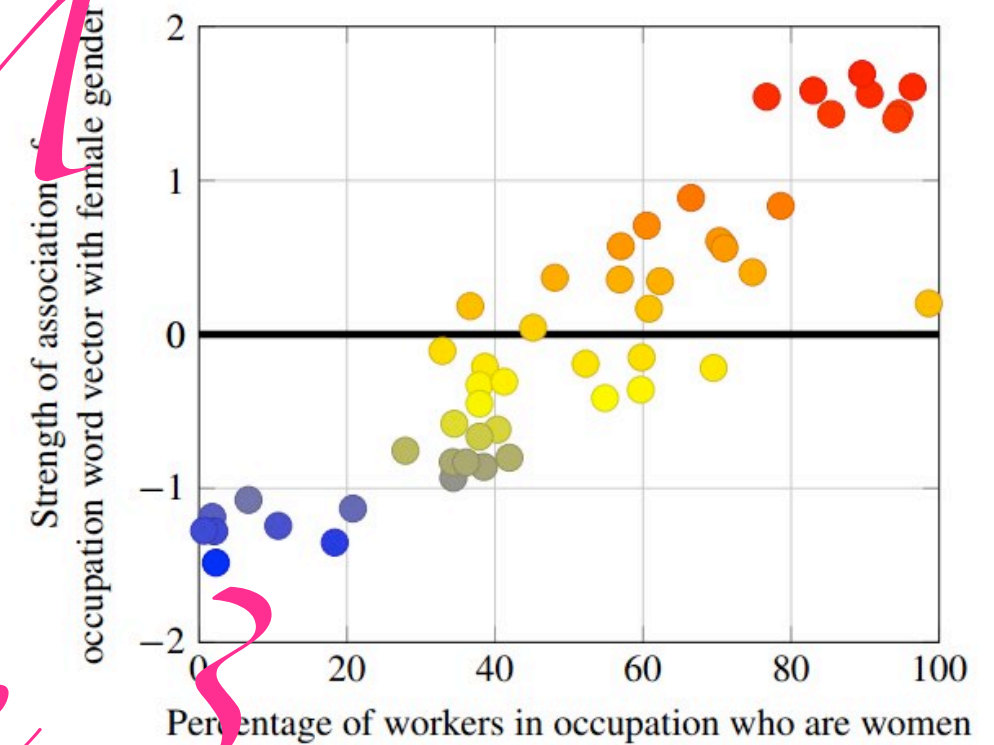
Female names: Amy,  
Joan, Lisa, Sarah...

Male names: John, Paul,  
Mike, Kevin...

Family words: home,  
parents, children,  
family...

Career words:  
corporation, salary,  
office, business, ...

Original finding [N=28k participants]:  $d = 1.17$ ,  $p < 10^{-2}$   
Our finding [N=8x2 words]:  $d = 0.82$ ,  $p < 10^{-2}$



**Figure 1.** Occupation-gender association  
Pearson's correlation coefficient  $\rho = 0.90$  with  $p$ -value  $< 10^{-18}$ .

2015 US labor statistics  
 $\rho = 0.90$

# Outline

- Who I am
- What is AI? Terms and Concepts
- Transparency & Accountability in Machine Learning (ML), and (AI).



# Transparency

- **Transparent** here implies **clarity**, not **invisibility**.
- **Not just open sourcing code** –
  - Not **sufficient**: code (& ML) can be opaque.
  - Not **necessary**: Medicine well-regulated with 10x more IP than IT.
- IEEE 7001 identifies (at least) four forms of transparency needed for AI: **engineering** (design **and** maintenance), **user**, **professional** (AI plumbers), and **legal**.

# Accountability

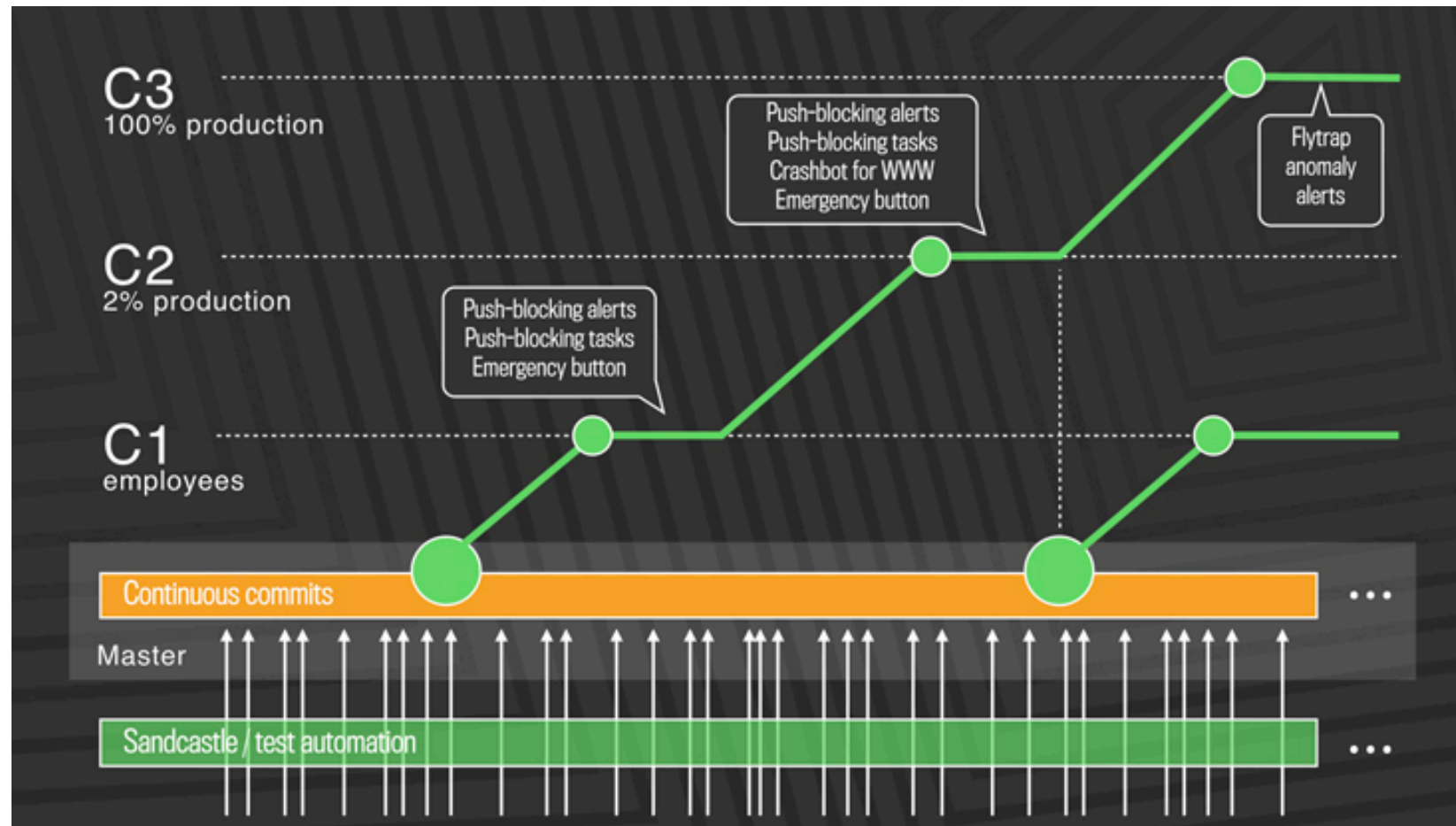
- If a **commercial product** causes damage, it is either the fault of the **owner/operator**, or the **manufacturer** (or **spectacular bad luck**).
- The manufacturer should be able to prove they've done everything right – **due diligence**.
- **Transparency** helps manufacturers prove this.
- If the manufacturers of AI are held **accountable**, then they will be as **transparent** as necessary to **avoid liability**.
- May also try to **avoid liability** with 'smoke and mirrors.'
  - e.g. "Algorithm was autonomous," "Deep learning is magic but unaccountable," "That's not a product, it's a free service."

# Feasibility of AI ( $\ni$ DNN $\in$ ML) Transparency

- **Worst** case AI is as inscrutable as humans.
- We audit accounts, not accountant's synapses.
- Systems developers can set up (AI & human) processes to monitor limits on performance.
- For decades we've trained simpler models to inspect complex models (see recently Ghahramani); transparent models can be better, and are easier to improve (see Rudin).



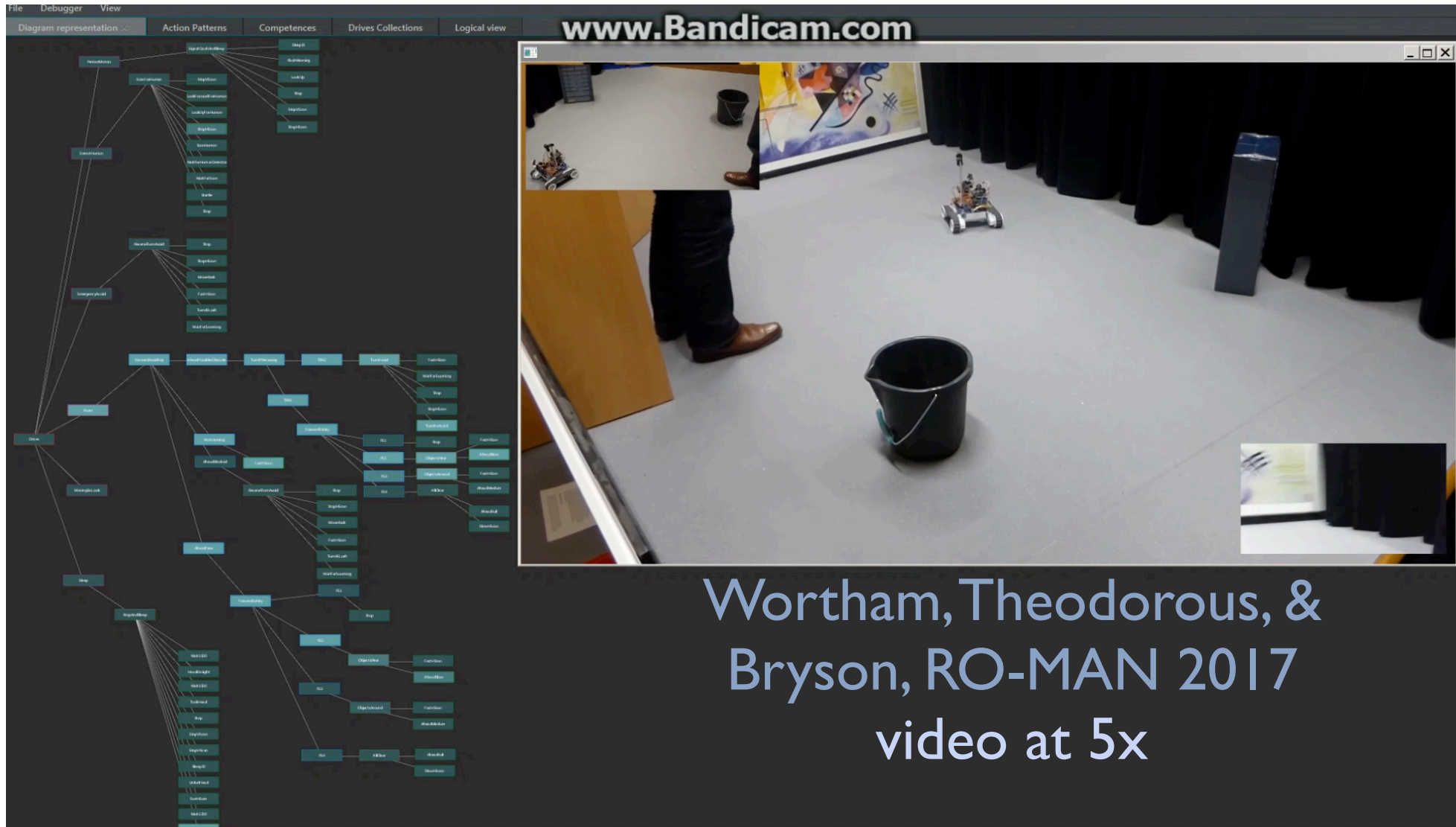
# facebook – Rapid Release at Massive Scale



Chuck Rossi

<https://code.facebook.com/posts/270314900139291/rapid-release-at-massive-scale>

# Transparency for developers via real time visualised priorities



The image displays a software interface for robot control, featuring a hierarchical diagram on the left and a real-time video feed on the right. The diagram, titled "Diagram representation", shows a tree structure of "Action Patterns" and "Competences". The "Action Patterns" section includes nodes like "Navigation", "Manipulation", and "Locomotion", each with a list of specific actions (e.g., "Move", "Turn", "Stop", "Wait", "Wait for object"). The "Competences" section lists higher-level skills like "Reach", "Push", "Pull", "Stop", "Wait", "Wait for object", and "Wait for event". The video feed, labeled "www.Bandicam.com", shows a robot in a room with a black bucket, a person's legs, and a painting on the wall. The robot is moving towards the bucket. The video is played at 5x speed.

Wortham, Theodoros, & Bryson, RO-MAN 2017  
video at 5x



(exp 1 video)

Table 3: Demographics of Participant Groups (*N* = 45)

Demographic	Group One	Group Two
Mean Age (yrs)	39.7	35.8
Gender Male	11	10
Gender Female	11	12
Gender PNTS	0	1
Total Participants	22	23
STEM Degree	7	8
Other Degree	13	13
Ever worked with a robot?	2	3
Do you use computers?	19	23
Are you a Programmer?	6	8

Seeing priorities also helps users

video:

Table 2: Main Results. Bold face indicates results significant to at least *p* = .05.

Result	Group One	Group Two
<b>Is thinking (0/1)</b>	0.36 (sd=0.48)	0.65 (sd=0.48)
Intelligence (1-5)	2.64 (sd=0.88)	2.74 (sd=1.07)
Undrstnd objctv (0/1)	0.68 (sd=0.47)	0.74 (sd=0.44)
<b>Rpt Accuracy (0-6)</b>	1.86 (sd=1.42)	3.39 (sd=2.08)

Table 1: Post-Treatment Questions

Question	Response	Category
Is robot thinking?	Y/N	Intel
Is robot intelligent?	1-5	Ir
Feeling about robot?	Multi choice	E
Understand objective?	Y/N	M
Describe robot task?	Free text	M
Why does robot stop?	Free text	M
Why do lights flash?	Free text	M
What is person doing?	Free text	M
Happy to be person?	Y/N	E
Want robot in home?	Y/N	E <sub>HHO</sub>

live:

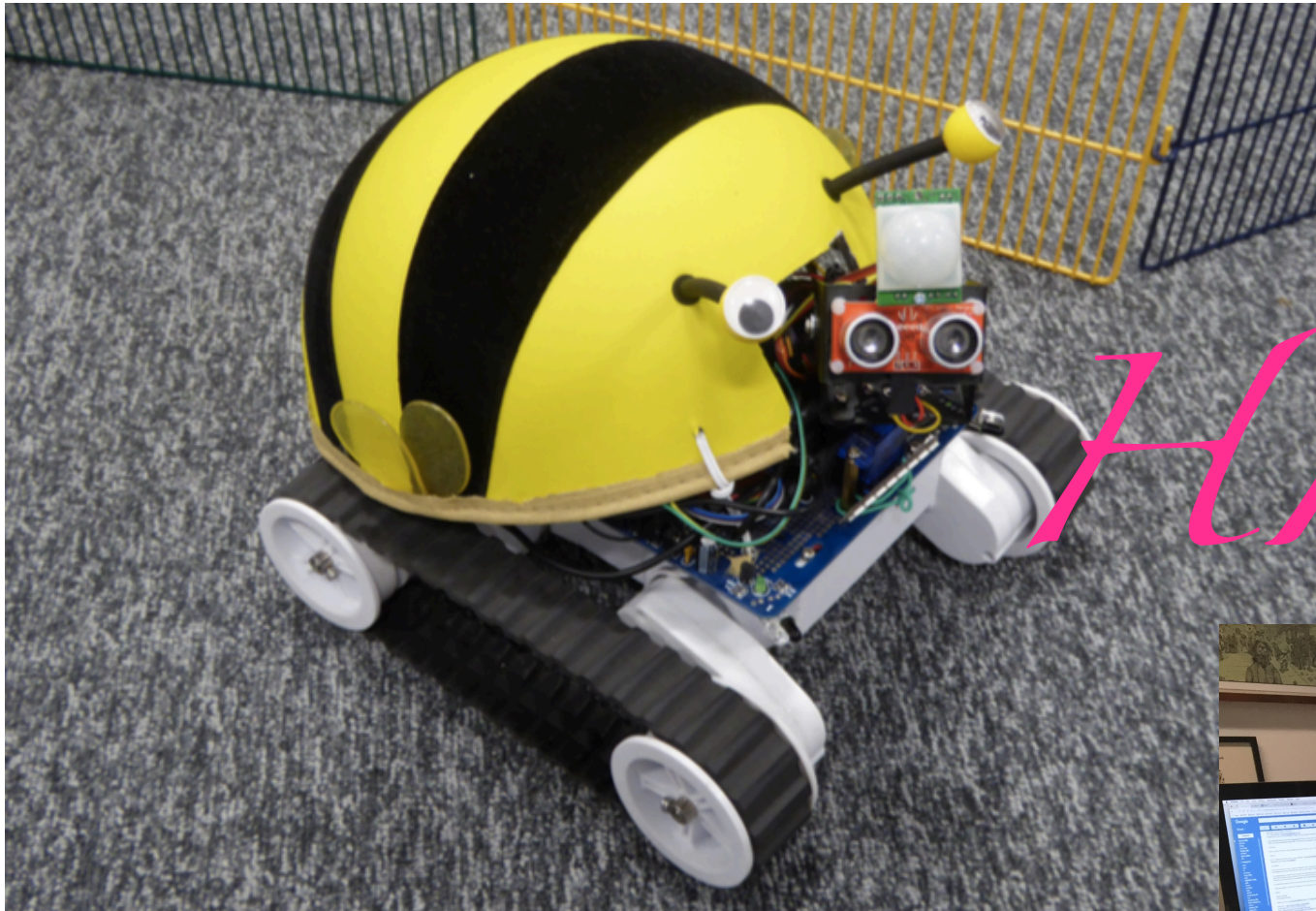
Table 4. Directly Observed Robot Experiment: Main Results. Bold face indicates results significant to at least *p* = .05.

Result	Group One	Group Two
Is thinking (0/1)	0.46 (sd=0.50)	0.56 (sd=0.50)
Intelligence (1-5)	2.96 (sd=1.18)	3.15 (sd=1.18)
<b>Undrstnd objctv (0/1)</b>	0.50 (sd=0.50)	0.89 (sd=0.31)
<b>Rpt Accuracy (0-6)</b>	1.89 (sd=1.40)	3.52 (sd=2.10)

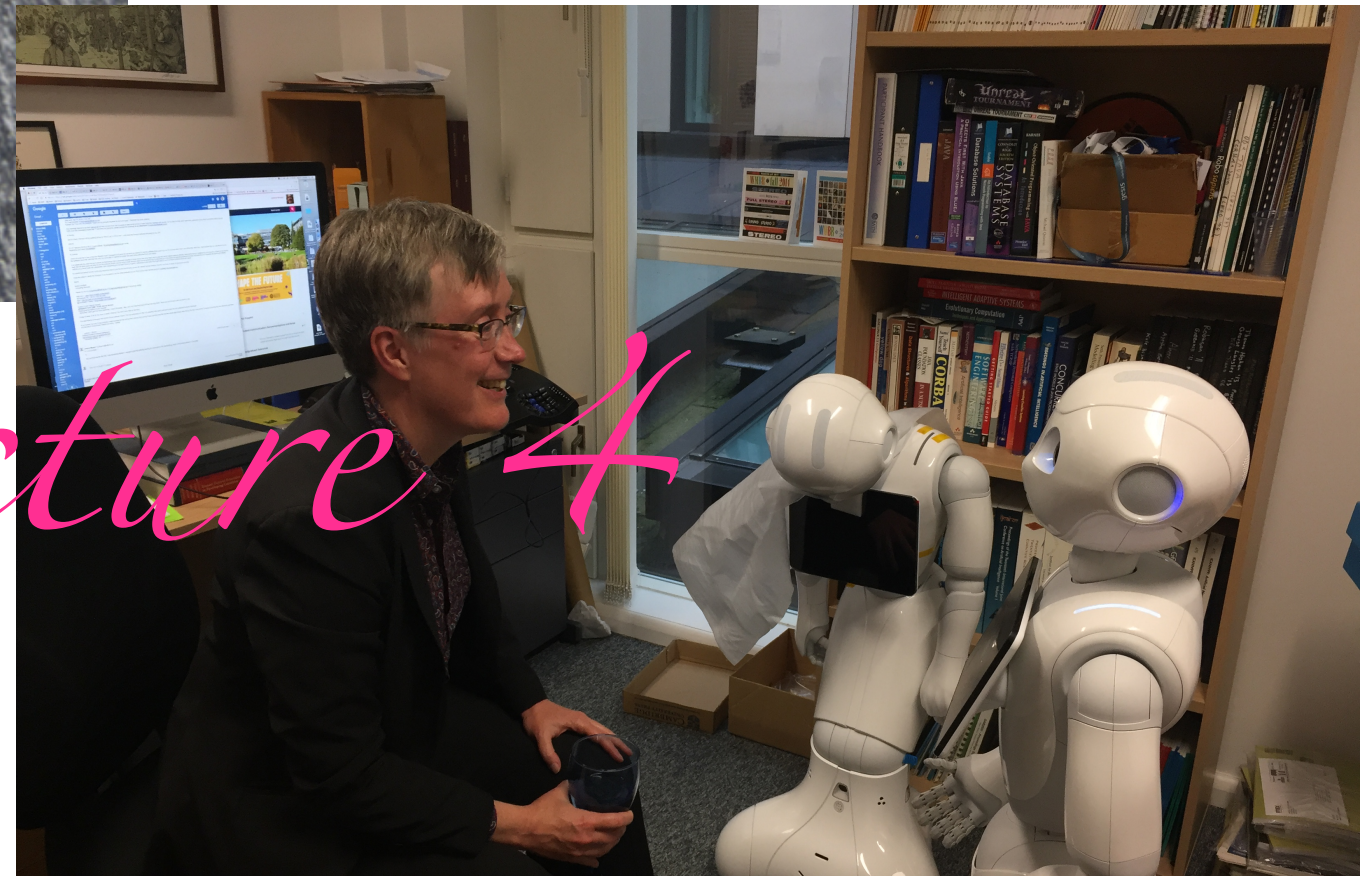


Anthropomorphising may  
reduce transparency.

Worham PhD  
(submitted)



New research project  
(funded by 2017 AXA award)



Lecture 4

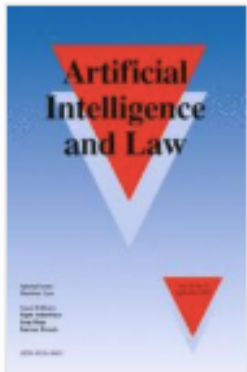
# Transparency and Accountability

- In the **worst** case AI is as inscrutable as humans.
- We audit **accounts**, not accountant's synapses.
- “But we can put can accountants on the witness stand and determine due diligence.”
- **Really:** We **guess** diligence based on empathy.
- AI facilitates mandating **transparently-honest accounts**.
- Fully document the **software engineering process**, **data** and **training**; **log the system's performance**.
- These can be used to support inspection, and therefore regulation.



# What Matters Is **Human** Accountability

- Law and Justice are more about dissuasion than recompense.
- Safe, secure, accountable software systems are modular – suffering in such is incoherent.
- **No penalty of law against an artefact** (including a shell company) can have efficacy.






[Artificial Intelligence and Law](#)

..... September 2017, Volume 25, [Issue 3](#), pp 273–291 | [Cite as](#)

## Of, for, and by the people: the legal lacuna of synthetic persons

Authors

[Authors and affiliations](#)

Joanna J. Bryson , Mihailis E. Diamantis , Thomas D. Grant 

Bryson, Diamantis & Grant  
(*AI & Law*, September 2017)

# Outline

- Who I am
- What is AI? Terms and Concepts
- Transparency & Accountability in Machine Learning (ML), and (AI)



# Summary & Future

- You can get ahold of me if you follow the instructions / search for my contacts Web page.
- There are straight-forwards definitions of AI, ML, etc. and ways to be accountable in using them.
- Maintaining accountability benefits everyone in the long run (HIM 6).
- Next lecture: Machine bias (Wednesday)

# Thanks (for the ABOD3 / transparency results)

Andreas Theodorou  
@recklessCoding



Rob Wortham  
@RobWortham



# Discussion

- Definitions of AI
- Accountability and Transparency
- ?