

Intelligent Control  
and Cognitive Systems

brings you...

# Regulation and Transparency for AI

Joanna J. Bryson

University of Bath, United Kingdom  
@j2bryson

with help from ...



CENTER FOR INFORMATION TECHNOLOGY POLICY  
AT PRINCETON UNIVERSITY

# Outline

- Terms and Concepts
- Transparency & Accountability in Machine Learning (ML), and (AI).
- Regulating AI
- The Moral, Legal, and Economic Hazard of Anthropomorphising AI

# Notes

- This is an area of VERY active “research” and negotiation.
- Many of these slides I gave last week at the Financial Conduct Authority, one of the UK’s regulatory bodies for protecting citizens and the economy.
- Much of it is therefore necessarily opinion, but we can still have a discussion.
- It’s also a chance to revise the basic court content.

- **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.

# Definitions

for communicating  
right now

Ethics is  
determined by and  
determines a  
society—a constantly  
renegotiated set of  
equilibria.

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(?)

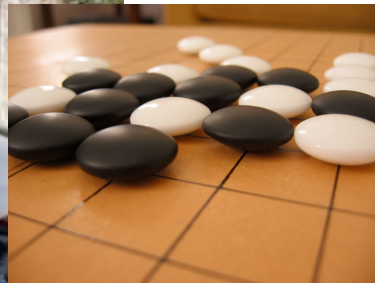
Omniscience (“AGI”) is not a real threat. No one algorithm can solve all of AI.

Viv Kendon, Durham



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



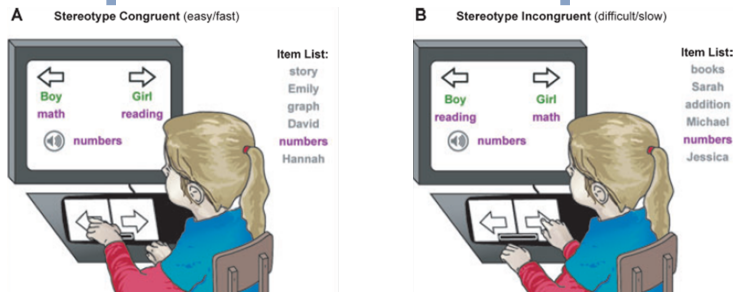
**Semantics derived automatically from language corpora contain human-like biases**

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...

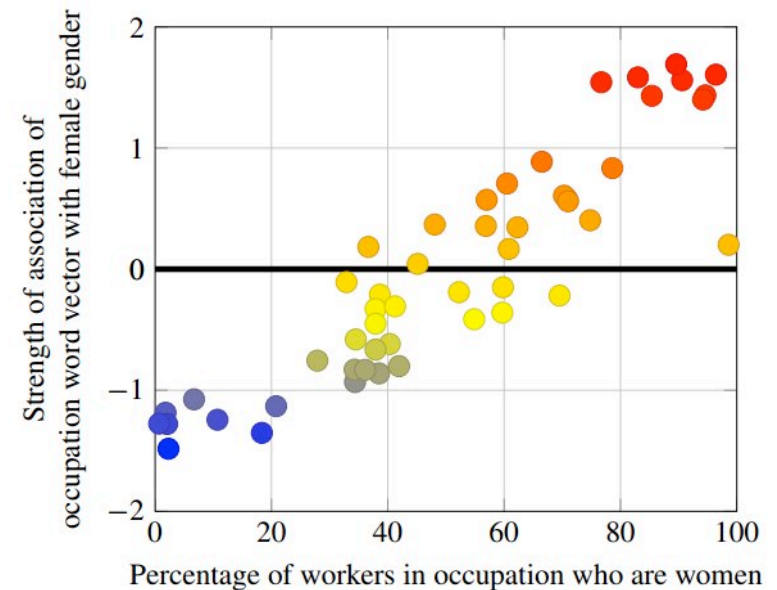
Male names: John, Paul,  
Mike, Kevin...

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

$\rho = 0.90$

# Basic Definitions

Caliskan, Bryson & Narayanan 2017

- **Bias**: expectations derived from experience regularities in the world.
- **Stereotype**: biases based on regularities we do not wish to persist.
- **Prejudice**: acting on stereotypes.

# Example

Caliskan, Bryson & Narayanan 2017

- **Bias:** expectations derived from experienced regularities. Knowing what *programmer* means, including that most are male.
- **Stereotype:** biases based on regularities we do not wish to persist. Knowing that most programmers are male.
- **Prejudice:** acting on stereotypes. Hiring **only** male programmers.

# Critical Implication

- **Bias**: expectations derived from experience regularities in the world.
- **Stereotype**: biases based on regularities we do not wish to persist.
- **Prejudice**: acting on stereotypes.
- **Stereotypes are culturally determined. No algorithmic way to discriminate stereotype from bias!**
- **So what should we do?**



# At **Least** Three Sources of AI Bias

- Absorbed **automatically** by ML from ordinary culture.
- Introduced through **ignorance** by insufficiently diverse development teams.
- Introduced **deliberately** as a part of the development process (planning **or** implementation.)

# Ignorance from lack of diversity but it's still totally unacceptable.



Probably nobody meant to force people to use white toilet paper to get soap...



*Joy Buolamwini*

...or to make their face recognition software work better on abstracted white masks than black faces...

# How to deal with them

- **Automatic**—compensate with design, architecture.
- **Ignorant**—diversify, test, iterate, improve.
- **Deliberate**—audits, regulation.

# Outline

- Terms and Concepts
- Transparency & Accountability in Machine Learning (ML), and (AI).
- Regulating AI
- The Moral, Legal, and Economic Hazard of Anthropomorphising AI

# Transparency

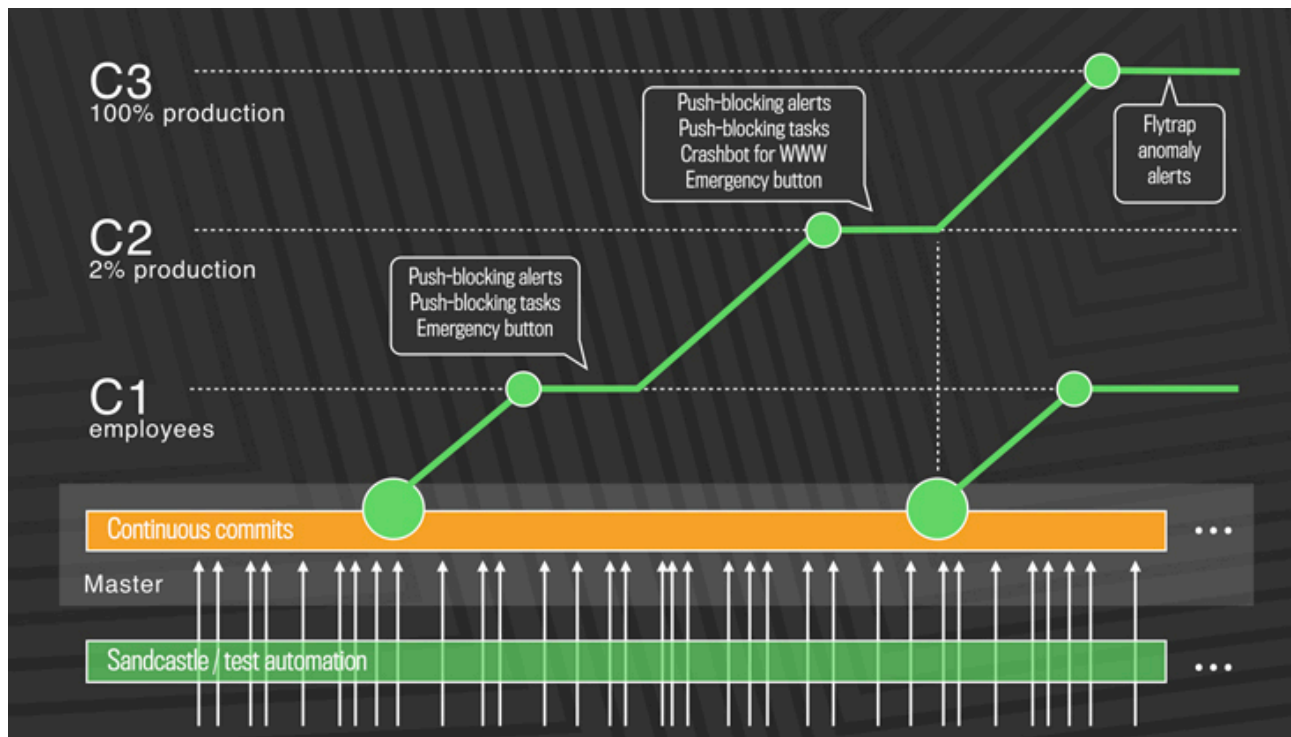
- **Transparent** here implies **clarity**, not **invisibility**.
- **Not just open sourcing code** –
  - Not **sufficient**: code (& ML) can be opaque.
  - Not **necessary**: Medicine well-regulated with 10x more IP than IT.
- IEEE 7001 identifies (at least) four forms of transparency needed for AI: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

# 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 easier to improve (see Rudin).



# facebook – Rapid Release at Massive Scale

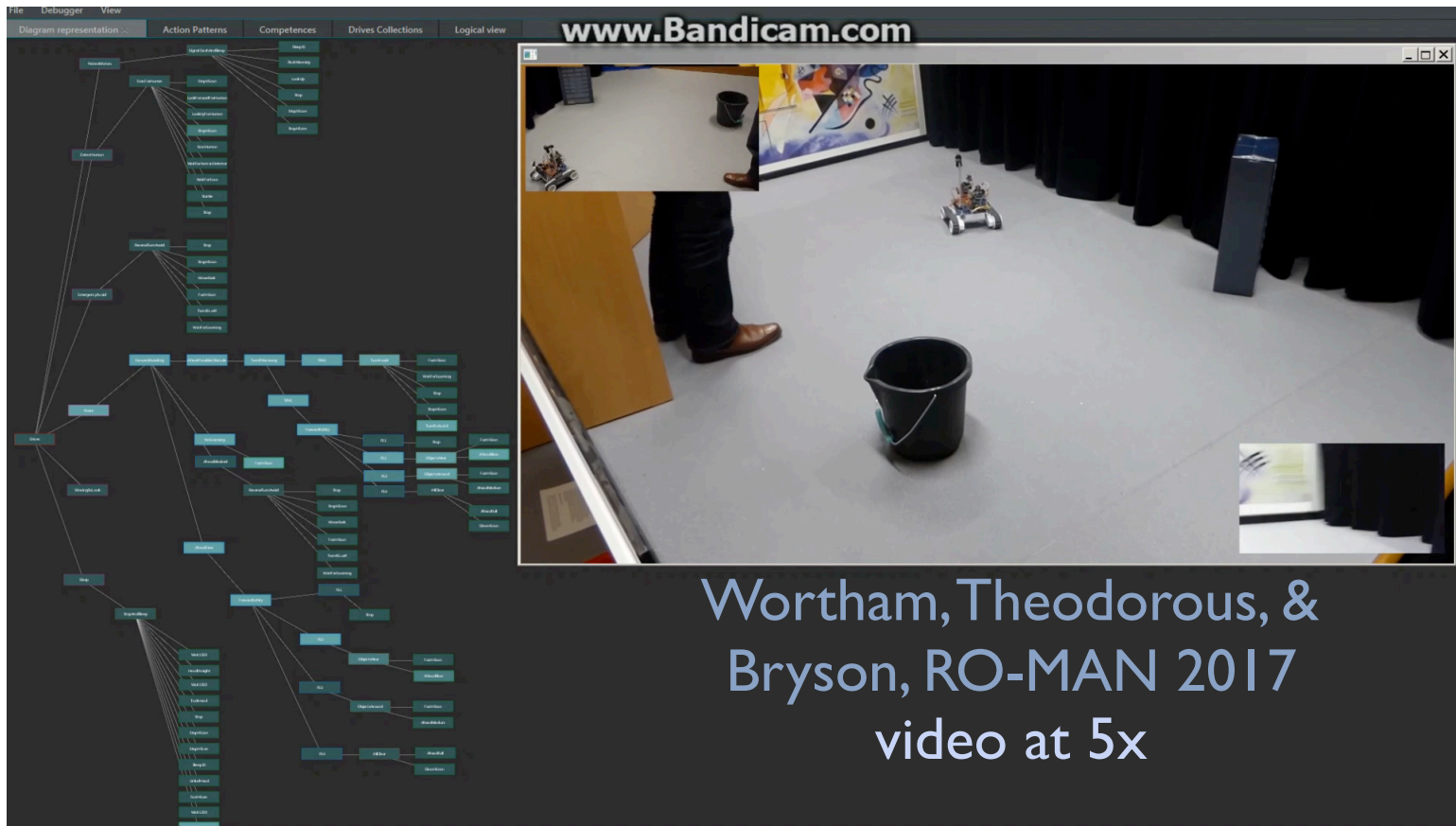


Works partly because of great HR and salaries, but what's more generalisable is: **Automated, deterministic processes monitoring for violations of specs.** This works for AI and NI both.

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 debugger interface with a dark background. On the left side, there is a complex hierarchical tree structure representing the system's logic, with nodes labeled with terms like 'Action Patterns', 'Competences', and 'Drives Collections'. The right side of the interface features a video window titled 'www.Bandicam.com' showing a real-time camera feed of a robot in a room. The robot is positioned on a grey floor, with a black bucket and a person's legs visible in the foreground. The video window includes standard window controls (minimize, maximize, close) in the top right corner.

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



(exp I video)

# Seeing priorities also helps ordinary users

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

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)

live:

Table 1: Post-Treatment Questions

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

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

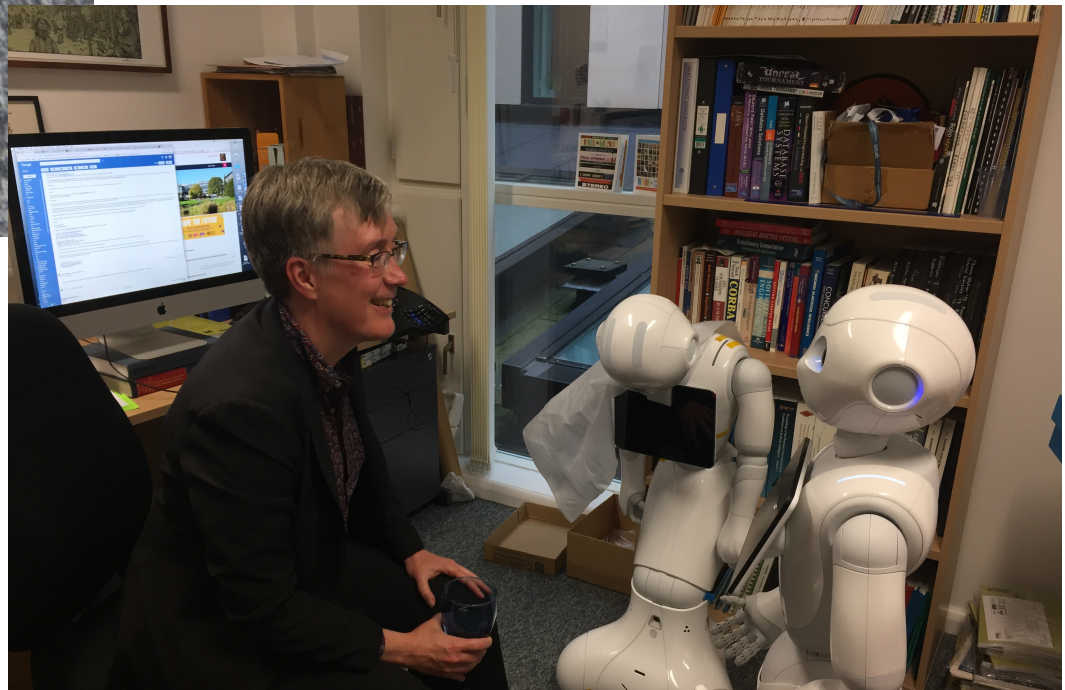
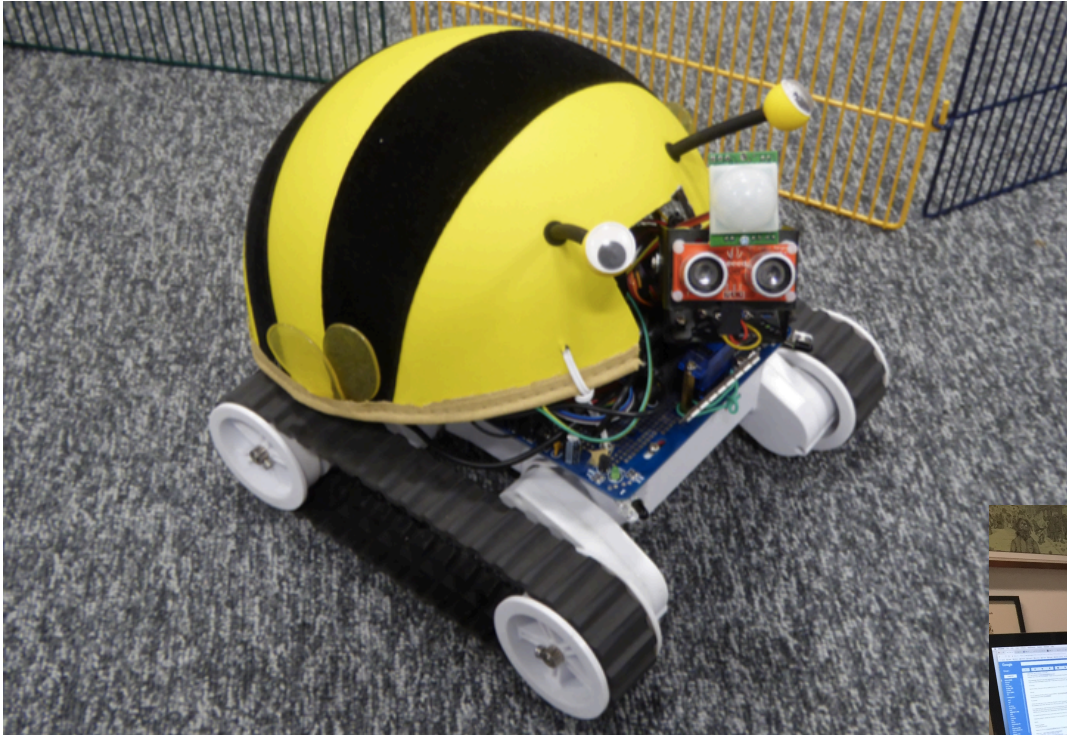
Result	Group One	Group Two
Is thinking (0/1)	0.46 (sd=0.50)	0.56 (sd=0.50)
Intelligence (1-5)	2.96 (sd=1.18)	3.15 (sd=1.18)
<b>Undrstnd objctv (0/1)</b>	0.50 (sd=0.50)	0.89 (sd=0.31)
<b>Rpt Accuracy (0-6)</b>	1.89 (sd=1.40)	3.52 (sd=2.10)

Wortham, Theodorou & Bryson 2017

# Anthropomorphising may reduce transparency.

Wortham PhD  
(submitted)

New research project  
(funded by 2017 AXA award)



# 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 accounting mechanisms, e.g. block chained logs, “black boxes”, software revision logs.
- We can check due diligence by the **(legal) person(s)** responsible.

# Outline

- Terms and Concepts
- Transparency & Accountability in Machine Learning (ML), and (AI).
- **Regulating AI**
- The Moral, Legal, and Economic Hazard of Anthropomorphising AI

# AI Is Changing Us

- Blurring distinction between customer and employee – citizens of corporations.
- “Free” services are information bartering – undenominated transactions avoiding revenue.
- Reducing (not eliminating) costs and advantages of geographic location, increasing inequality and transnational interdependence.
- Altering governance – makes stabilisation of policy through obscuring difficult or impossible.





# ICT Systems Are Designed, and Have Architecture

- Architects learn laws, policy, and how to work with governments & legislatures at university...
- because society decided collapsing buildings were unacceptable, and city alterations affected everyone.
- ICT systems are now falling on people and affecting everyone. Field needs to mature, as architecture did.
- Rate of **successful, sustainable** innovation is what matters, not just speed to market.

# Regulating AI

- **Do not** reward corporations by capping liabilities when they fully automate business processes – Legal lacuna of synthetic persons (Bryson, Diamantis & Grant 2017.)
- **Do not** motivate obfuscation of systems by reducing liabilities for badly-tested or poorly-monitored learning, or special status for systems with ill-defined properties, such as ‘consciousness’.
- **Clear code is safer** and can be more easily maintained, but **messy code is cheaper** to produce (in the short run.)
- **Regulation should motivate clarity** (transparency) by requiring proof of due diligence.

# AI Requires Security; Security Is an Arms Race

- Google got hacked by the NSA (cf. Snowden). The US Federal Government got hacked by people interested in who worked with/on China. Political parties, banks, cheap apps, LinkedIn...
- IoT devices generate less revenue than the cost of a security upgrade – lightbulbs & baby monitors **stay** compromised.
- There is no cybersecurity/autonomy “tradeoff” – you are encrypted or you aren’t. Backdoors get too many keys made.



# Good Practice for Intelligent Systems Engineers

- Educate – actively (e.g. training videos) & passively (e.g. open source code).
- Follow good systems engineering – architect and document carefully and as openly as possible.
- Intelligent systems need real-time, varied architecture monitors, limits, and checks.
- Engage with government, media, and professional organisations (e.g. BCS, IEEE).

# Law and Professional Societies

IEEE  computer society

The Community for Technology Leaders

[Libraries & Institutions](#) [About](#) [Resources](#) [Subscribe](#)

[CSDL Home](#) » [Computer](#) » [2017 vol. 50](#) » [Issue No. 05 - May](#)

## Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems

Joanna Bryson, University of Bath

Alan Winfield, University of the West of England

**Pages:** 116–119

**Abstract**—AI is here now, available to anyone with access to digital technology and the Internet. But its consequences for our social order aren't well understood. How can we guide the way technology impacts society?

**Keywords**—artificial intelligence; AI; autonomous system; ethics; standards; P7001; IEEE Standards Association; British Standards Institute; robotics; history of computing

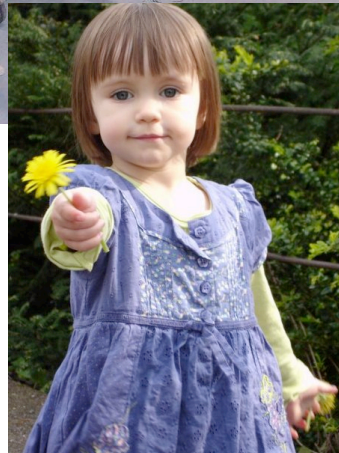
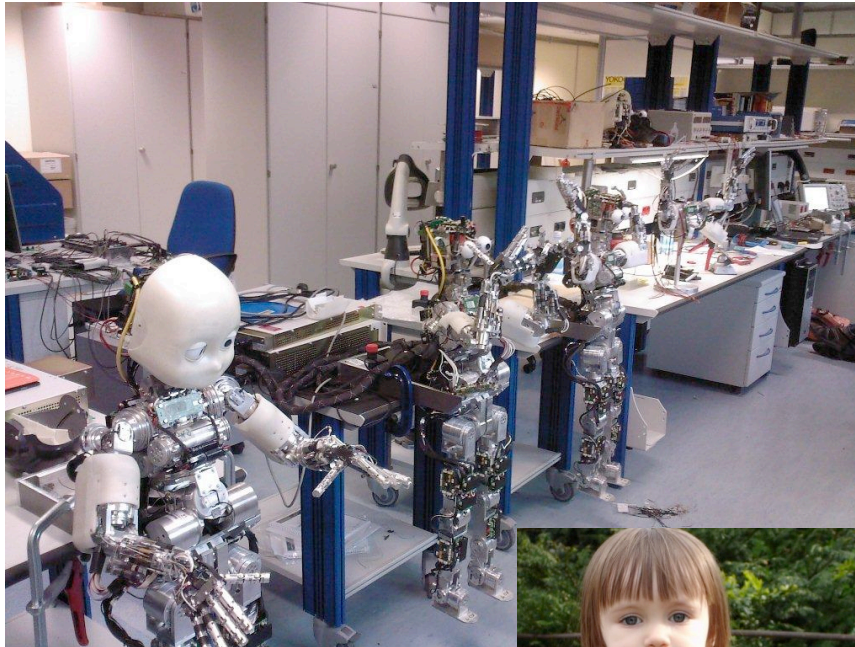
# Law and Professional Societies

- **Governments** are good at enforcing law, redistribution.
- **Professional societies** are good at talking to people who know stuff, keeping up with contemporary issues.
- **Combination** – Professional societies maintain standards, governments enforce these standards.

# Outline

- Terms and Concepts
- Transparency & Accountability in Machine Learning (ML), and (AI).
- Regulating AI
- The Moral, Legal, and Economic Hazard of Anthropomorphising AI

There's no question  
whether we have the  
technical capacity to build  
synthetic legal persons.



(Bryson 2010, 2016, 2018)

AI and ethics are both authored—cultural artefacts. Science cannot determine AI's place in society—that decision is normative, not factual.

photos: Georgio Metta (top) & Emmanuel Tanguy

The real questions:

Can we build a system we  
are not obliged to?

Are we obliged to do so if  
we can?

**Can we build a system we  
are not obliged to?**



# Can we build a system we are not obliged to?

- Yes
  - We already have (many times).
  - We can eliminate non-replaceability by using mass-produced hardware and continuously backed-up memory.
  - We can avoid resentment of subordinate position by not cloning evolved minds.
- ...at least in licensed commercial products.

Are we obliged to do  
so if we can?

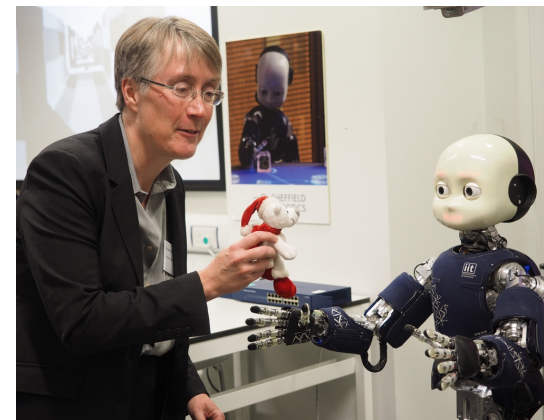
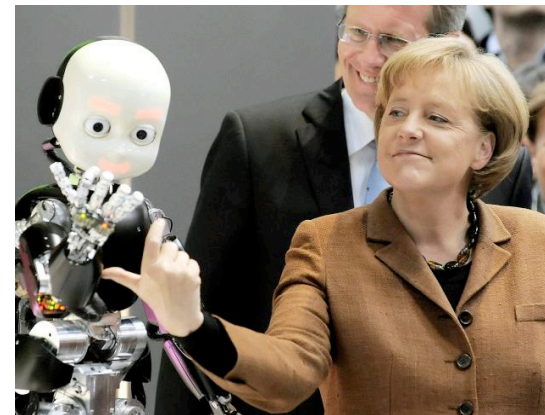
- Yes

Five Reasons Not to  
Other AI

# #1 Moral Hazard

- We are preprogrammed to think humanoid robots are people (Kamewari & al 2005).
- So people will think we've made persons **well** before we have.
- Facilitates political and economic exploitation.

Bryson & Kime 1998, IJCAI 2011



# #2 Second Order Moral Patiency

- Why should we build robots to suffer when they lose social status? To 'die' in fires? To mind being owned?
- We are obliged to build robots we are not obliged to.
- **Not** a double standard: pick one standard for moral subjects, **don't** build to it.



LF Miller 2015 Hellström 2013; Bryson 2016, 2007

# #3 Fear of Robot Apocalypse Distracts from Real Threats

- AI is here now changing the world.
- By increasing communication, interdependence, discoverability, we decrease privacy and individual autonomy.
- Projecting AI into the future endangers us now.

(Bryson 2015)

# Intelligence explosion / Superintelligence







Vernor Vinge (1983, 1995):

Machines more intelligent than men  
– making prediction (and therefore  
**hard scifi**) impossible.



Alan Bundy, FRS



Kevin Warwick



I J Good (1965); Nick Bostrom (2016)  
Self improving machine intelligence.

Exponential Growth

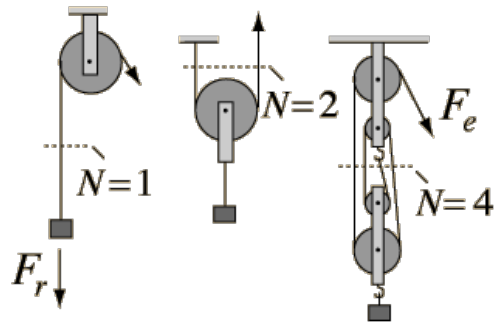


# 12,000 years of AI

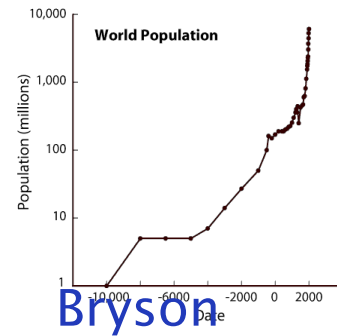
If we accept that **intelligence** can be decomposed (e.g. motivation, action, **memory**, learning, reasoning)...

Then every machine and especially **writing** have been examples of **AI**.

The “intelligence explosion” is us—**AI-enhanced humans**.



Pulley  $IMA = N$



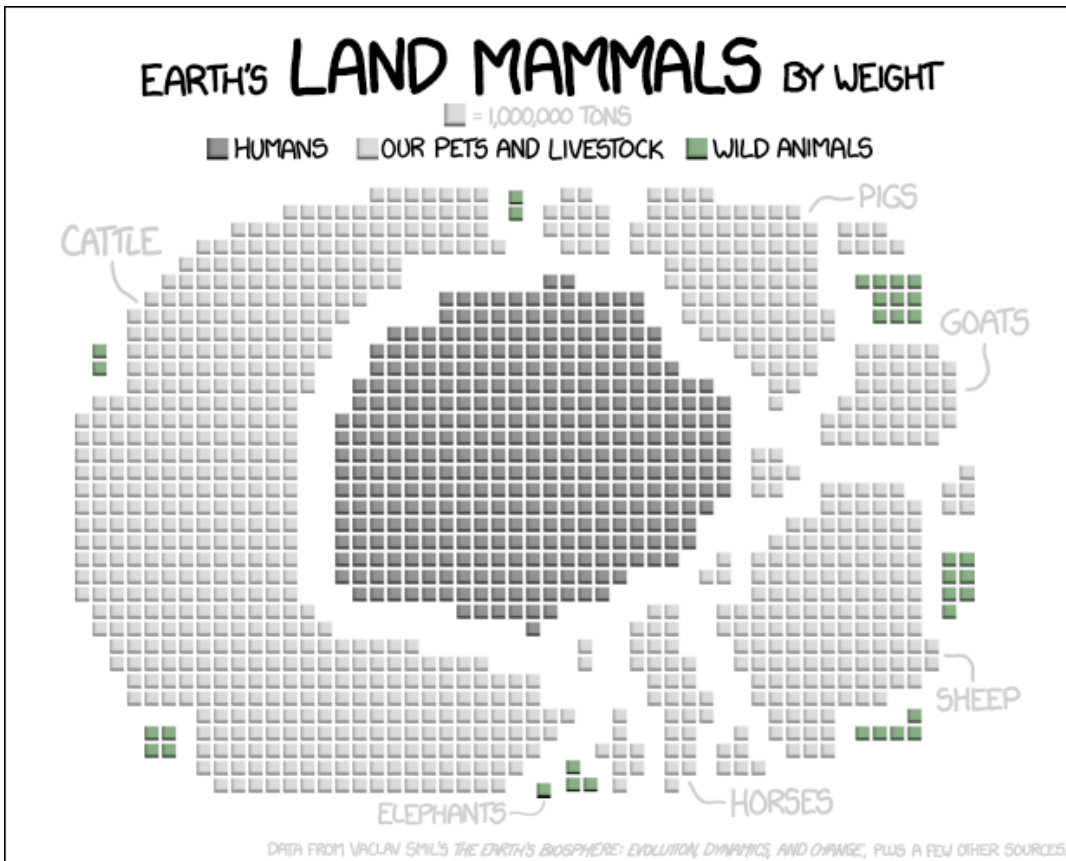
Bryson  
Collective Agency  
2015

Superintelligence is us.

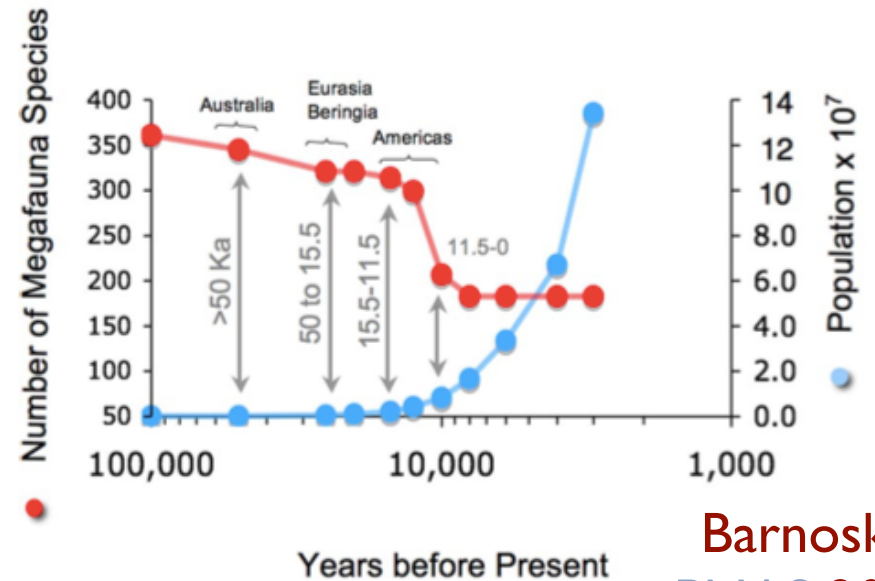
Not paper clips.

Cows.

xkcd



● Megafauna Loss vs. Global Human Population Growth ●



Barnosky,  
PNAS 2008

Unanticipated Subgoals:  
We turn extant biomass  
and fossil fuels into  
more biomass but fewer  
species.

# Issues for Superintelligence

- Existing diversity – cockroaches & bacteria still here.
- Robustness & resource constraints: is there enough coltan for a robot revolution?
- Probability – nuclear war, “ice nine” / nanotech, bird flu, asteroids – does AI increase or decrease our survival chances?

# #4 Ethical Coherence

- What makes people special is that we're members of a social species – **we've evolved in a context of interdependence** (Zahavi 1977, Sylwester & al 2013).
- Society **defines, enforces** 'responsibility'; enforcement often through **punishment** (Solaiman 2016).
- Evolution ensures suffering, shame are inextricable parts of being human (also of apes, dogs).
- **Good AI is modular; suffering in such is incoherent.**
  - **Clones should not be slaves, nor made.**



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

## #5 Legal Lacuna

- Assigning responsibility / personhood to artefacts allows powerful individuals & organisations to avoid tax, legal liability.
- Try suing a bankrupt robot.
- **Already a problem:** shell organisations (AI, cf. List & Pettit 2011) shield rich companies.

Tom Dale  
Grant



Mihailis E.  
Diamantis



- **My nightmare:** Autocrats willing money and power to AI self caricatures.

# Kantian Fallacy

(a mistake made by Prescott, Gunkel, & others)

- Kant: People who treat things we identify with (e.g. dogs) badly also treat people badly  $\therefore$  treat dogs well.
- Wrong take: **Because we will over-identify with AI, we must grant robots rights.**
- Wrong because a) no identification with e.g. search, translation, b) legal lacuna.
- Right take: **Because AI is an ethics sink, we must focus on building AI we don't identify with.**  
**cf transparency, and the UK's Principles of Robotics**

# The UK's EPSRC Five Principles of Robotics

- Written in 2010 to address ethics, first nation-level soft policy in this area in the world.
- The **first three** revise **Asimov's Laws** to communicate:
  - **Artefacts aren't persons.**
  - **Manufacturers** have standard responsibilities for artefacts.
- The **fourth and fifth** are about the rights and responsibilities of **consumers.**

# UK EPSRC's Principles of Robotics (2011)

1. **Robots are multi-use tools.** Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.
2. **Humans, not robots, are responsible agents.** Robots should be designed & operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, **including privacy.**
3. **Robots are products.** They should be designed using processes which assure their safety and security. (of 5...)



# UK EPSRC's Principles of Robotics (2011)

4. **Robots are manufactured artefacts.** They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be **transparent**.
5. **The person with legal responsibility for a robot should be attributed.** [like automobile titles]

for more discussion, read Bryson, *Connection Science* (2017)

# Outline

- Terms and Concepts
- Transparency & Accountability in Machine Learning (ML), and (AI).
- Regulating AI
- The Moral, Legal, and Economic Hazard of Anthropomorphising AI

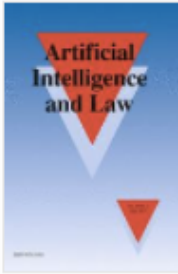
# ICCS Conclusions

- Learned about autonomous intelligence by programming robots.
- Learned about interacting social intelligence by programming ABM.
- Learned a marketable skill by programming a game.
- Please teach me by filling in the unit review form – we really do read the free text!

AND NSS



# Read the paper for...






[Artificial Intelligence and Law](#)

pp 1-19 | [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 

- Cases in international law where legal persons had rights but no responsibility, or responsibility and no rights, and the chaos that ensued.
- More formal discussion of veil piercing than what follows.
- Generally more formal and tight argumentation (law professors.)

# Legal Personhood

## 1. Actual persons / citizens / landowners

- (definition has been expanding)
- **in order to** resolve conflicts and coordinate action via contracts.

## 2. **Collections of humans**, in order to simplify contracts and negotiation between collectives.

- A **fiction** (**hack**) that only works because (or **to the extent**) corporations can be subjected to the same penalties as humans.

- Collections of humans, in order to simplify contracts and negotiation.
- A fiction (hack) that only works because (or to the extent) corporations can be subjected to the same penalties as humans.

# Fictitious Personhood

- **Collections of humans**, in order to simplify contracts and negotiation.
- A **fiction** (**hack**) that only works because (or **to the extent**) corporations can be subjected to the same penalties as humans.
- **Overextended already** (arguably).
- All the EP is really asking the EC to consider legislating.



# Recompense

- Penalties in law have two purposes:
  - actual compensation
  - dissuasion.
- Folk psychology confounds these, but really **jail time**, **fall in status**, **&c** don't compensate.
- Implausible that **built AI – designed & maintainable** – will be subject to dissuasion.

# Outline

- Terms and Concepts
- Transparency & Accountability in Machine Learning (ML), and (AI).
- Regulating AI
- The Moral, Legal, and Economic Hazard of Anthropomorphising AI

# Biological Utility of Intelligence and Communication

- Communication and agility allow **social strategies**
  - – individuals can discover new equilibria of mutual benefit – **public goods investment**.
- Increased communication increases group-level investment – reduces individual identity.

Roughgarden, Oishi, Akçay, **Science** 2006

# What Are People For?

rate of evolution  $\propto$  amount of variation

## Fisher's Fundamental Theorem of Natural Selection

Less variation means less robustness for  
addressing underlying change.

Without **privacy, tolerance, and diversity**  
we lose our capacity to innovate, which is  
required to address new challenges.

# Thanks to my collaborators, and to you for your attention.



Aylin Caliskan  
@aylin\_cim



Arvind Narayanan  
@random\_walker



UNIVERSITY OF  
**BATH**  
Established 1966

Tom Dale Grant



Mihailis E.  
Diamantis



Andreas Theodorou  
@recklessCoding



Rob Wortham  
@RobWortham



... and the rest of Amoni

