Intelligent Control and Cognitive Systems

brings you...

Ethics and Cognitive Systems

Joanna J. Bryson
University of Bath, United Kingdom

Beyond the Data Protection Act!

Ethics Beyond the Data Protection Act

ethics (noun, *uncountable*)

1. (philosophy) The study of principles relating to right and wrong conduct.
2. Morality.
3. The standards that govern the conduct of a person, especially a member of a profession.

wiktionary.org

Ethics Beyond the Data Protection Act

ethics (noun, *uncountable*)

1. (philosophy) The study of principles relating to right and wrong conduct.
2. Morality.
3. The standards that govern the conduct of a person, especially a member of a profession.

Data protection is part of being a computer science professional.

Ethics Beyond the Data Protection Act

ethics (noun, *uncountable*)

1. (philosophy) The study of principles relating to right and wrong conduct.
2. Morality.
3. The standards that govern the conduct of a person, especially a member of a profession.

New technology often triggers new concerns about right & wrong.

Ethical Questions for Cognitive Systems

- Is it possible to build something you are ethically obliged to?
  - If so, is it ethical to do so?
    - If so, how do you recognise when you have?
    - If not, what do you do if someone does anyway?

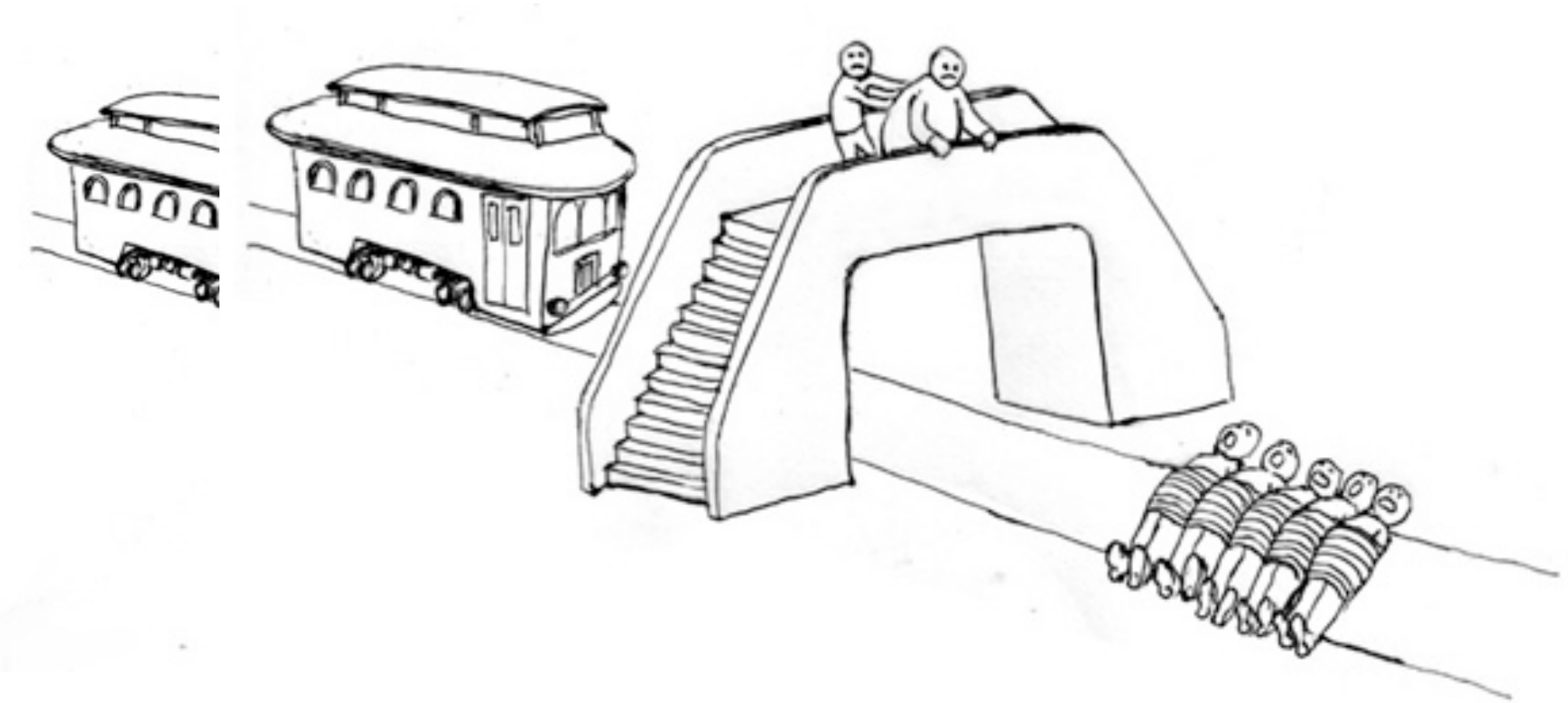
Outline

- Ethics concerning the behaviour of cognitive systems as artefacts.
- Special issues of human-appearing AI.
- Ethical obligations towards AI.

The Study of Ethics: Moral Philosophy

- How do you determine an appropriate course of action? Normative Ethics
- What do people actually do? Descriptive Ethics
- How can we achieve moral outcomes? Applied Ethics
- Can ethics even make sense? Meta Ethics

Descriptive Ethics



The Trolley Problem

Moral Minds

- Some principles fairly universal.
- Some variation correlates with culture.
  - Similar to language?
- Perscriptive (Normative) vs. Descriptive.

(Marc Hauser, *Moral Minds* 2006)

The New Trolley Problem

What if the one person you need to sacrifice is your owner?

The Study of Ethics: Moral Philosophy

- How do you determine an appropriate course of action? Normative Ethics
- What do people actually do? Descriptive Ethics
- How can we achieve moral outcomes? Applied Ethics
- Can ethics even make sense? Meta Ethics

Normative Ethics

- Example: Categorical Imperative – Act only according to that maxim whereby you can, at the same time, will that it should become a universal (Kant 1785).
- Others: Contractarianism, Natural Rights (humans are special), Consequentialism (utilitarianism, hedonism etc.)

c.f. Stanford Encyclopaedia of Philosophy (online!)

Vectors of Morality

- Moral agents: entities capable of being morally responsible.
- Moral patients: entities owed ethical obligation.
- Related issues: can you have rights without responsibilities?

The [2009] green paper proposes bringing together, into a single bill, a UK citizen's rights and the responsibilities that should go with them. The rights include economic and social rights, such as the right to free healthcare, the rights of crime victims and the right to equality. The responsibilities include the duty to vote, serve on juries, live within the country's environmental limits and promote the well being of children.
Jonathan Rayner, *Law Society Gazette* (April 2009)

Oz Group Guide to Illusion of Life

- Pursuing multiple, simultaneous goals and actions.
- Having broad capabilities (e.g. movement, perception, memory, language), and
- Reacting quickly to stimuli in the environment.

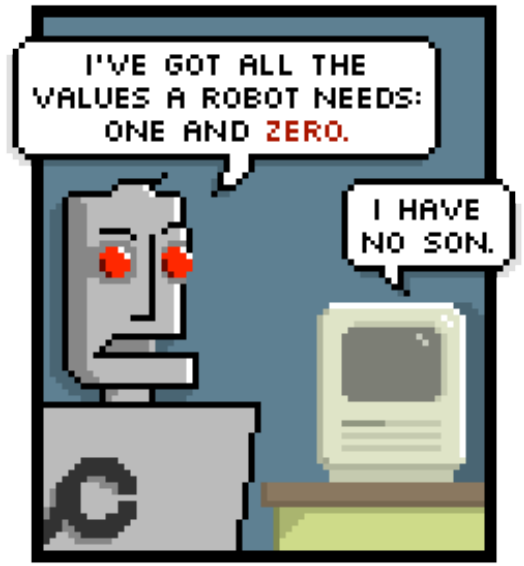
Is this the basis of moral agency?

Illusion of Life

Is this the basis of moral patiency?

Robots & Identity

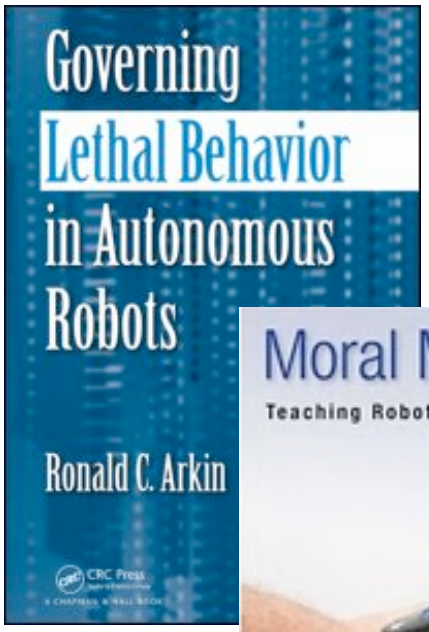
Sci Fi

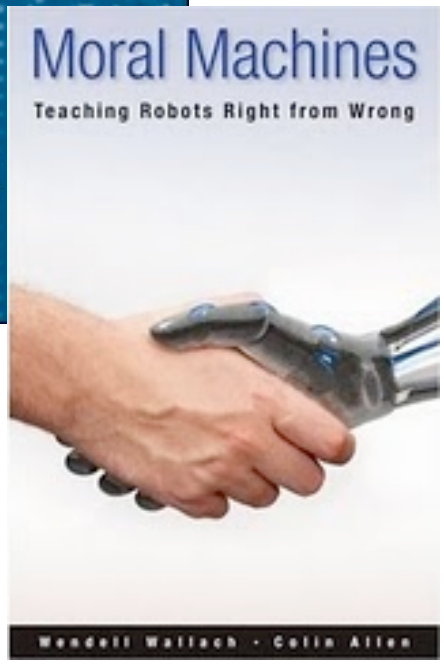


Diesel Sweeties

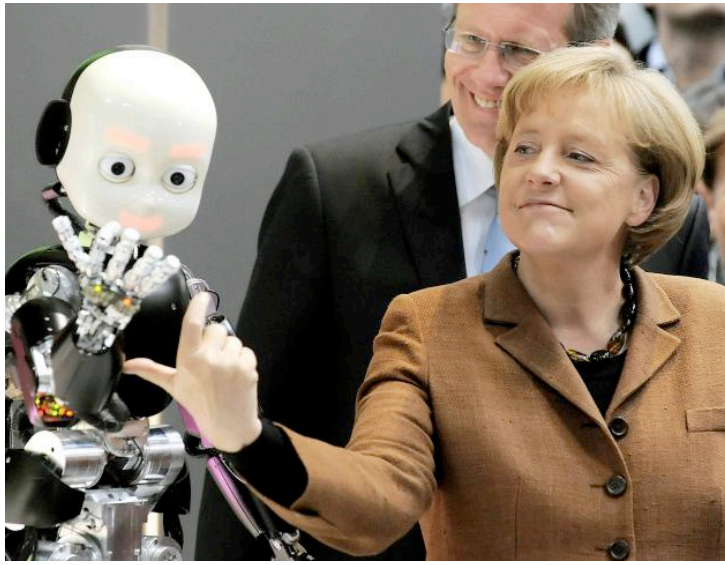
The Hitchhiker's Guide to the Galaxy

Science fiction uses alien perspectives to examine ourselves by analogy. Works via establishing empathy. Fiction.





The US government is spending $Ms on robot ethics research.



Agency & Liability

Since 2008 there have been more robots in Iraq than non-USA "coalition" forces. Powerful incentives for corporate & political liability to be transferred.

Joanna J. Bryson "Why Robot Nannies ProbablyWon't Do Much Psychological Damage", commentary on Noel Sharkey and Amanda Sharkey, "The Crying Shame of Robot Nannies: An Ethical Appraisal", *Interaction Studies*, **11**(2):196–200, June 2010.

Are Robot Weapons Different from Other Types?

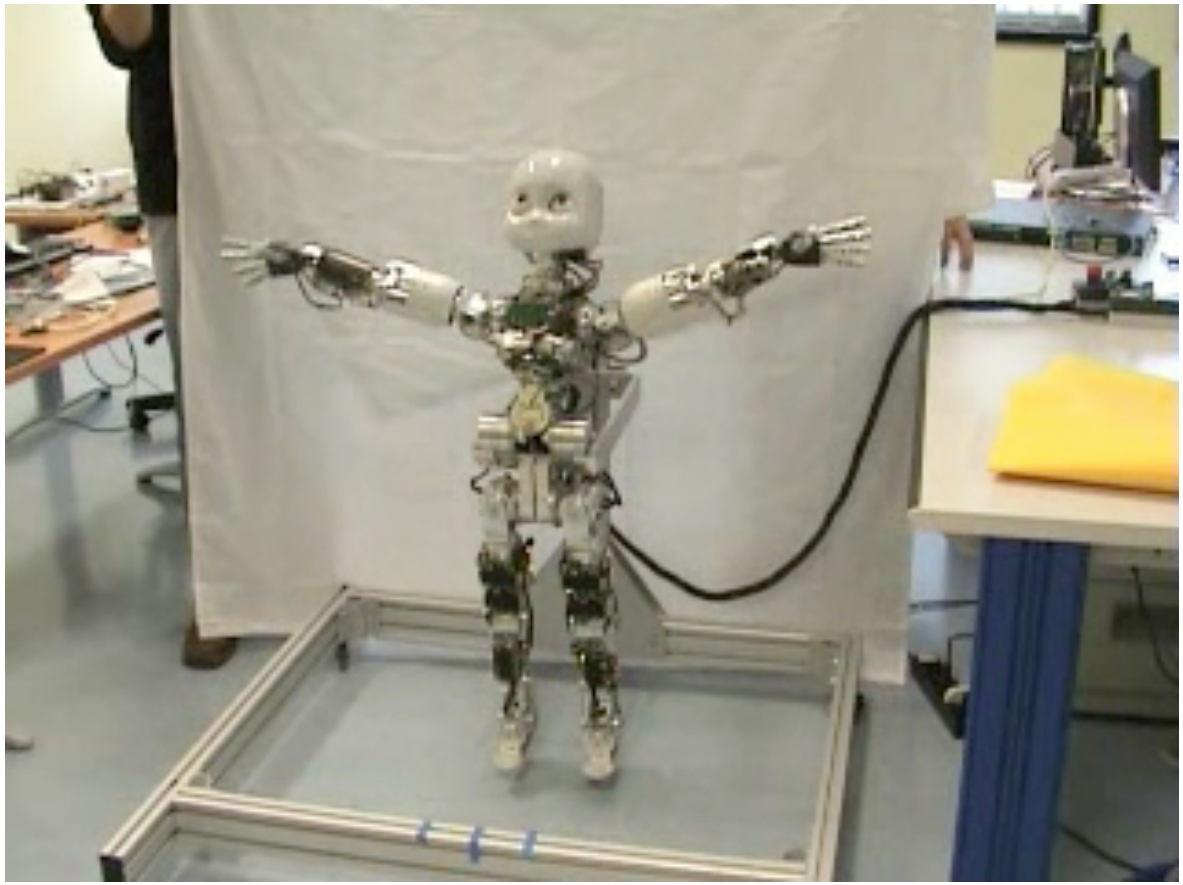
Outline

- Ethics concerning the behaviour of cognitive systems as artefacts.
- Special issues of human-appearing AI.
- Ethical obligations towards AI.

Abuse

- "My name is HAL, what's yours?" $NAME – obvious hacks (or just errors) here.
- Many people enjoy abusing robots.
- Ignoring (failure to detect) can be a serious problem in a learning bot.
  - Also seen by some business owners as unacceptable to their brand.

But People Also Anthropomorphise



Q: How much Cognition is here?

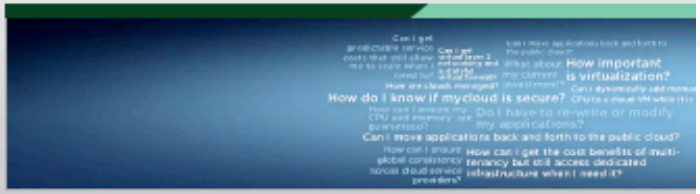
None
(no sensing)

Descriptive Test

- Could you be loved by a robot?
- Could you love a robot?

technology review

Published by MIT

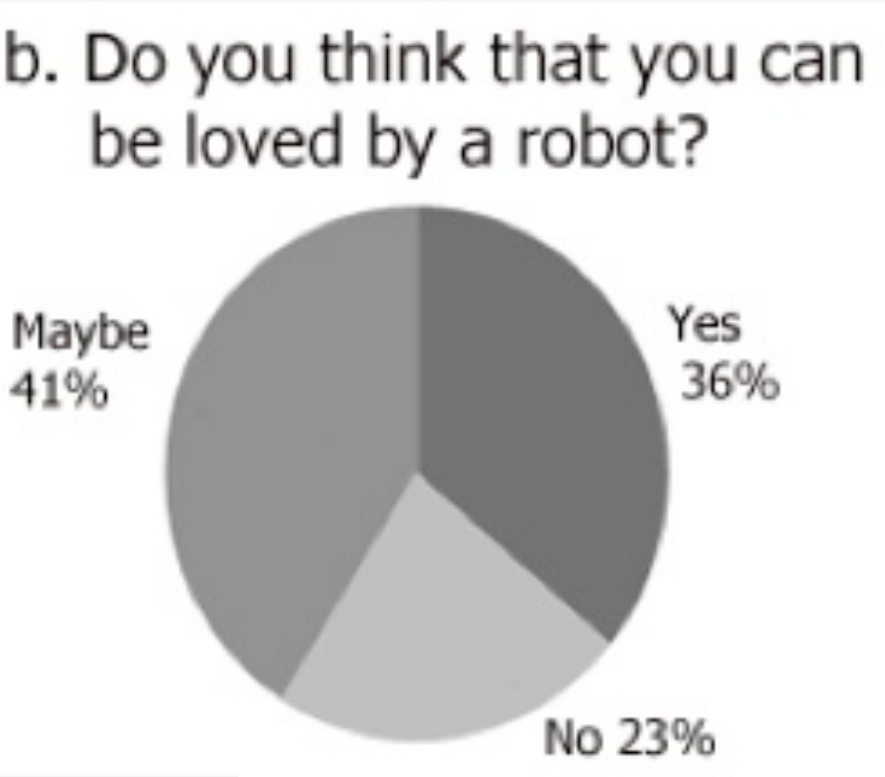


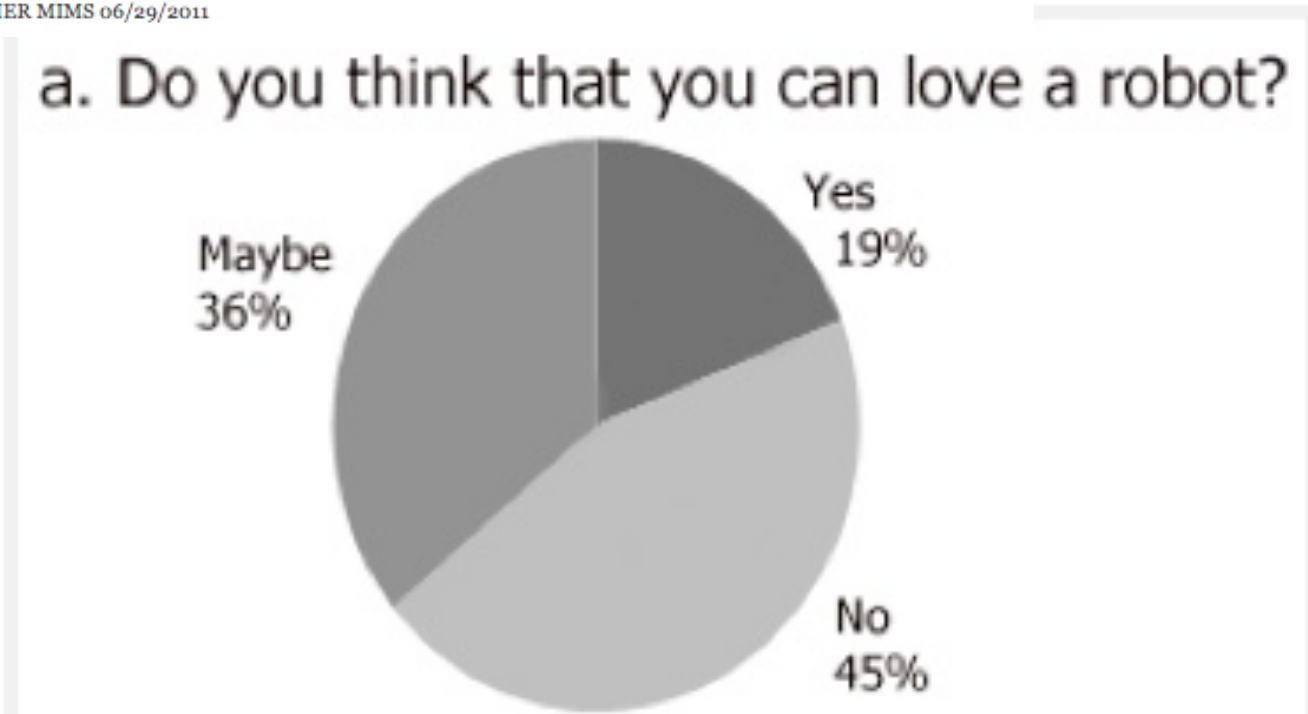
Mims's Bits

'Lovotics': The New Science of Engineering Human, Robot Love

"After industrial, service and social robots, Lovotics introduces a new generation of robots, with the ability to love and be loved by humans"

CHRISTOPHER MIMS 06/29/2011





Hooman Samani
June 2011

Bryson & Kime (1998)

"We believe exaggerated fears of, and hopes for, AI are symptomatic of a larger problem – a general confusion about the nature of humanity and the role of ethics in society.

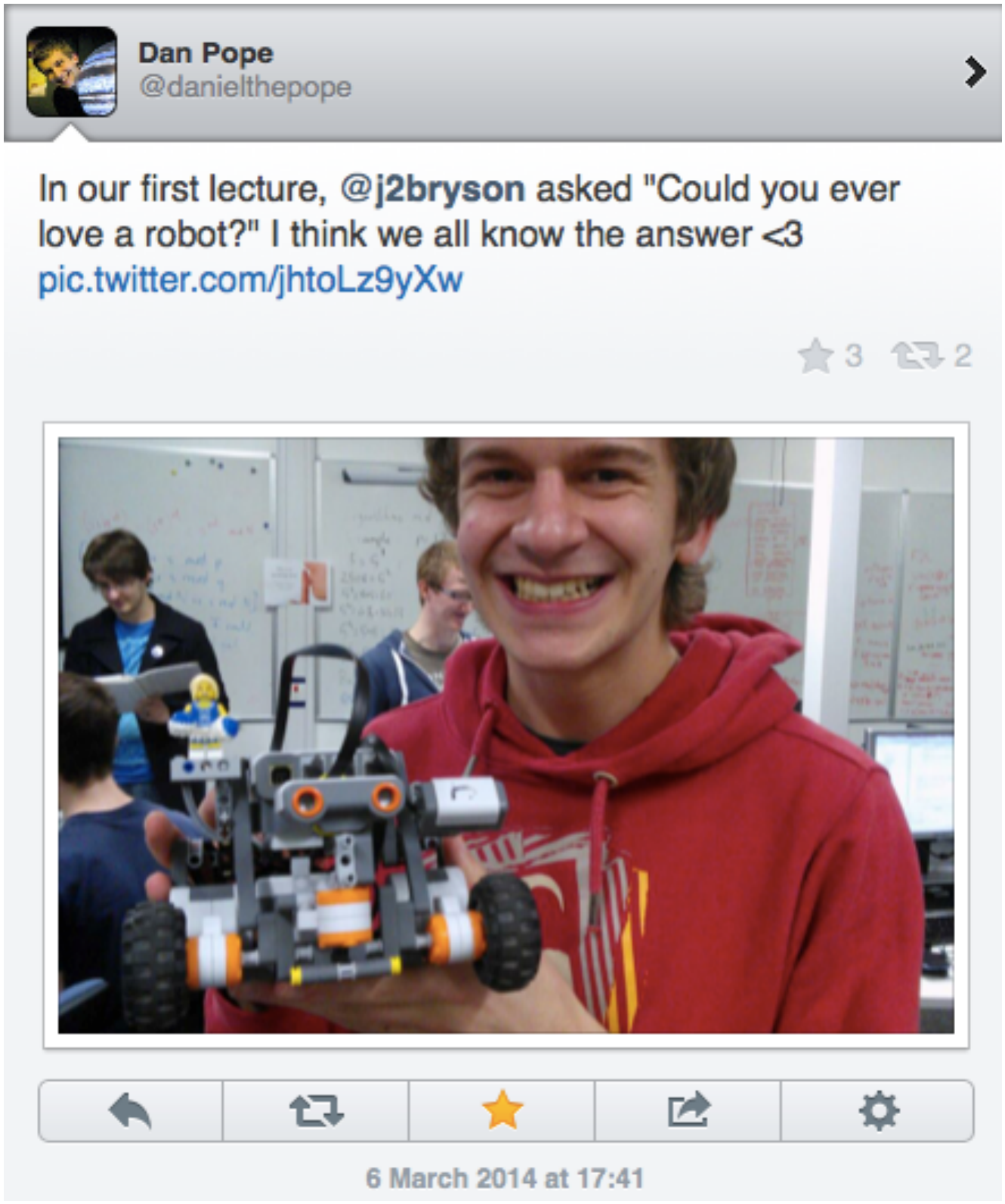
…

Our thesis is that these [exaggerated fears] are false concerns, which can distract us from the real dangers of AI technologies. The real dangers are no different from those of other artifacts…the potential for misuse, either through carelessness or malevolence, by the people who control them."

Phil Kime (2011)

Bryson & Kime (1998)

- Ethical instincts (and ethics itself) is rooted in identity / identification.
- Humans (mistakenly) place language, mathematics & reason as core to humanity, because these discriminate us from animals.
- Once we have empirical experience of AI, this confusion might go away.
  - Might even inform our (human) ethics.



- Final year student in a leading computer science undergraduate degree.
- Built & programmed the robot (with a partner.)

- Are we going to love robots?
  - Yes, obviously.
- Will they love us?
  - We can make them arbitrarily monogamous.
  - We can link their self-image to their model of their owner.
  - Probably never as many interconnections (identity, perceptible bonds) as evolution gave us.

Concerns

Should we (policy makers, academics and manufacturers) deceive people into misallocating resources to robots?

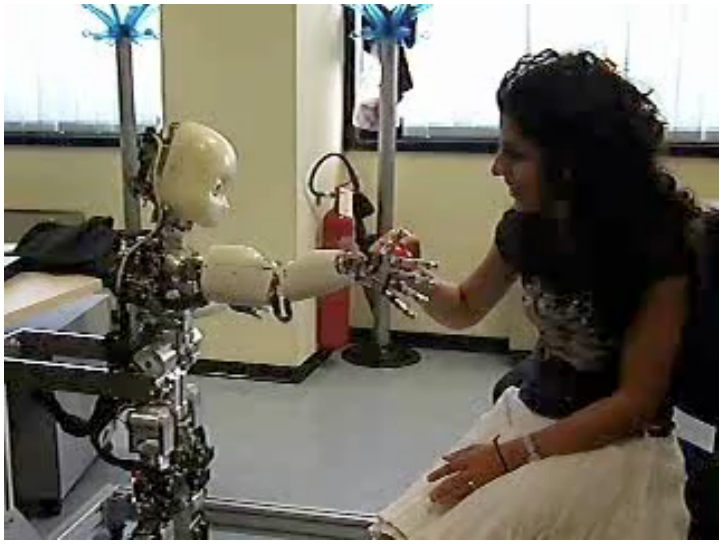
Examples of resources: time, money, attention, personal care.

What if it makes us money or wins us votes?

Note: the lines on this slide before this one were in this lecture in 2015.

Ethical Tradeoffs with Care-Giving Robots

- ✤ Some patients would be substantially helped by forming any bonded relationship.
- – Many people already dedicate considerable resources to AI (TV, games), possibly to their own and society's detriment.
- • Balancing these two opposing concerns requires establishing both practice & policy.



Outline

- Ethics concerning the behaviour of cognitive systems as artefacts.
- Special issues of human-appearing AI.
- Ethical obligations towards AI.

Ethical Questions

- Is it possible to build something you are ethically obliged to?
  - If so, is it ethical to do so?
    - If so, how do you recognise when you have?
    - If not, what do you do if someone does anyway?

- People routinely give resources (even their lives) to cultural constructs.
- No ethical system prohibits this for all cases of artefact.
  - Especially not the ethical system itself.

Ethical Questions

- Is it possible to build something you are ethically obliged to?
- If so, is it ethical to do so?
- If so, how do you recognise when you have?
- If not, what do you do if someone does anyway?

Take a postgraduate degree in philosophy.

The general case is too hard for a single AI lecture.

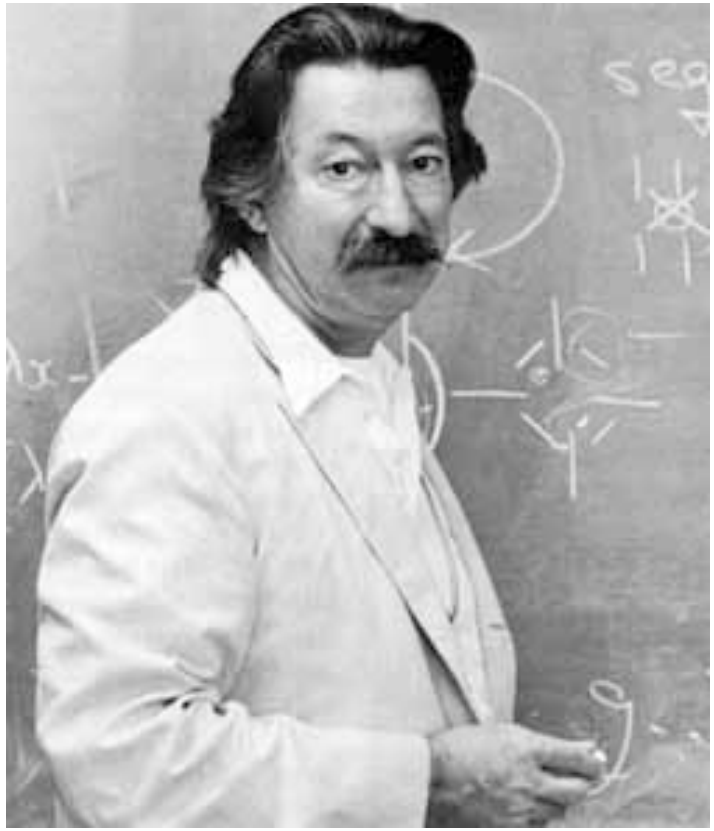
By descriptive ethics at least

Published Positions

- Ethics depends on {life, consciousness, relatedness, social good} [choose one].
- If we think we might owe something ethical obligation we better err on the side of caution (Dennett, Brooks).
- Culture is the ultimate consequence of life on Earth and should be retained / extended by self-reproducing AI spaceships long after the end of our planet & species (Tom Ray).

Weizenbaum & His Secretary

- Weizenbaum: we shouldn't work on AI not because we are obliged to it, but because we are obliged to each other.
- Humans are too easily fooled.



Computer Power and Human Reason, Weizenbaum (1976)

My (published) Opinion

(Bryson 2012, 2018)

"Is does not imply ought" – Hume.

Descriptive does not entail normative. We can always do better than what we've evolved so far.

AI ethics relates two types of human artefact: ethical systems & robots. There is no pre-determined slot for AI we need to discover.

Question: is there any utility in displacing the responsibility we as authors have onto AI?

Not a question: whether it's possible.

What Are Moral Actions?

- a behavioural context affords more than one possible action for the individual,
- at least one available action is considered by a society to be more socially beneficial than the other options, and
- the individual is able to recognise which action is socially sanctioned and act on this information.

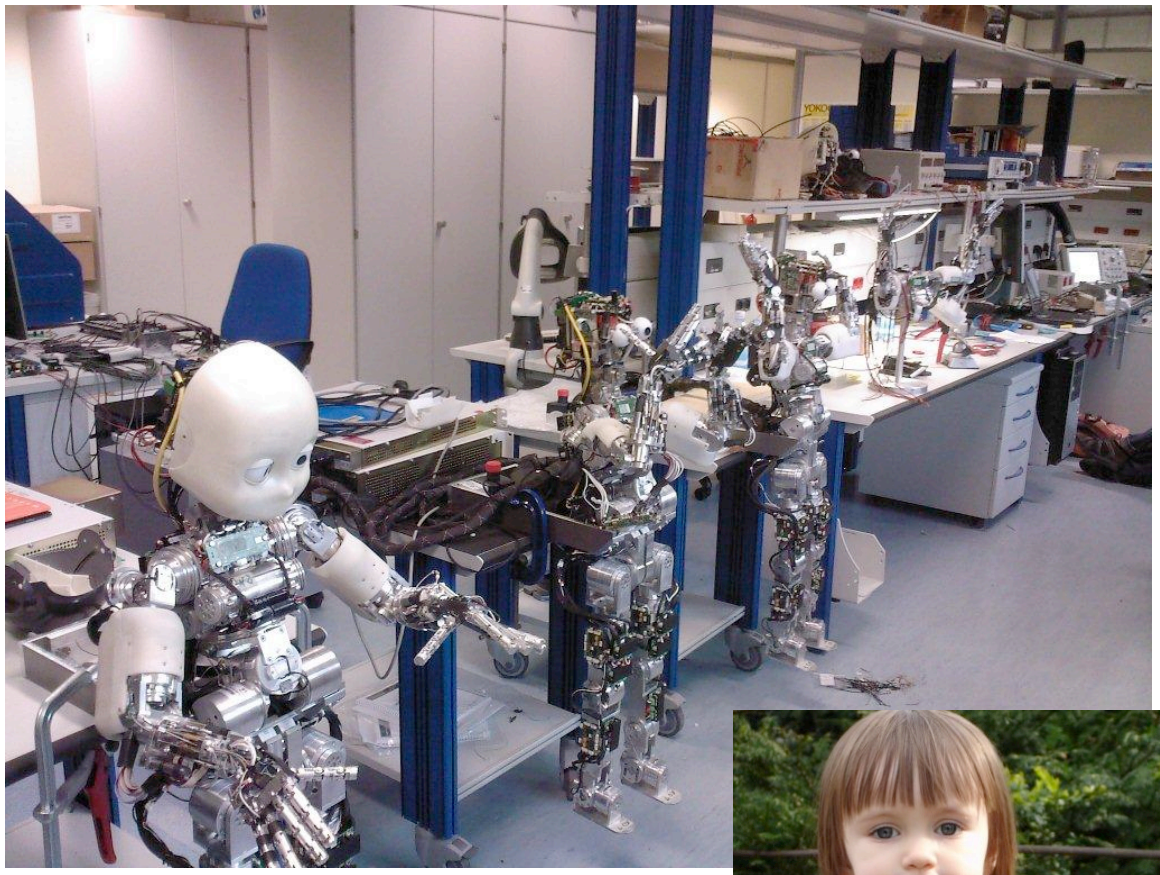
Easy to build in robots!!
(includes monkeys & pets)

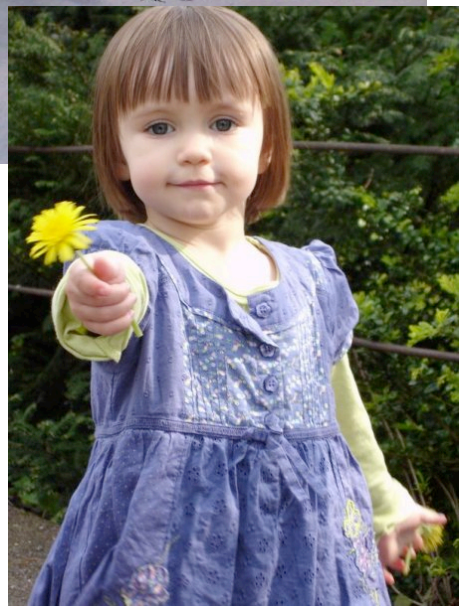
Displacement of Responsibility

- at least one available action is considered by a society to be more socially beneficial than the other options, and
- What are the pros & cons of considering the robots responsible? (Instead of e.g. considering them intelligent prosthetics of human will, like owned dogs & children.)

- For Human Society (us):
  - Pros: feel godlike, culture might persist beyond planetary limits, might produce more useful tools.
  - Cons: political & commercial moral hazard, misattribution of blame / resources.
- For AI (them robots):
  - No Pros: (except maybe for the unbuilt).
  - Cons: compete w/ humans for resources, stress of social dominance, fear of death etc.

Slides concluding a typical AI ethics keynote by me:





We build robots and other AI, determine these systems' goals. Our authorship of AI is fundamentally different from our relationship to other evolved systems.

a fact

photos: Georgio Metta (top) & Emmanuel Tanguy

Conclusions

(my conclusions for that keynote)

We are ethically obliged to make robots we are not ethically obliged to.

Deeming robots to be moral agents unethically neglects our responsibility as authors of their intelligence.

normative assertions

Ethics and the Law

- Remember the Baldwin Effect?
- The law is explicit, rapidly updated ethics.
- Next (final) lecture: regulation of AI (new this year!)