Lecture on
Ethics for AI and Robotics

Angelo Cangelosi
CoRo Lab, Department of Computer Science
The University of Manchester

Content

- Ethical Problems for AI and Robots
- Moral thinking and ethics approaches
  - Deontological / Utilitarian / Value ethics
- Potential Harms
  - Bias
  - Denial of autonomy
- What can we do?
  - Machine ethics
  - Responsible AI (Explainable AI)

Let’s play a game

moralmachine.mit.edu

Ethical Problems for AI and Robots

- Are robot ‘stealing’ our jobs?
- Should we worry about superintelligence and the singularity?
- How should we treat robots?
- Should robots become our friends and companions?
- Should robots have ethical responsibilities?
- Should robots/AI systems be allowed to kill?

Two fields of AI/Robotics and Ethics

I. AI Ethics
(Ethics for AI/Robotics; Responsible AI)

II. Machine Ethics
(AI for robot's own ethics)

AI Ethics (Ethics for AI/Robotics)

AI ethics is a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies

(Leslie 2019)

To understand ethical implications in the design of safe, acceptable and ethical robots

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. *The Alan Turing Institute*. https://doi.org/10.5281/zenodo.3240529

Machine Ethics (AI for robot's ethics)

"An ethical machine is guided by own, intrinsic ethical rule, or set of rules, in deciding how to act in a given situation."

(Anderson & Anderson 2006; Winfield et al. 2017)

To design machines/robots with intrinsic ethics rules

Winfield et al. (2019). Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems. *Proceedings of the IEEE*.
https://doi.org/10.1109/JPROC.2019.2900622

Ethics Approaches

1. Deontological ethics
2. Utilitarianism (consequentialism)
3. Virtue ethics

Burton, E., Goldsmith, J., Koenig, S., Kuipers, B., Mattei, N., & Walsh, T. (2017). Ethical considerations in artificial intelligence courses. *AI Magazine*, *38*(2), 22-34.

Ethics Approaches: Deontological ethics

What is my/our moral law?

- Etymology: “Deon” (right/proper)
- Kant: responsibility of individual to discover the true moral law for him/herself
- Any true, moral law is universally applicable
  - e.g. Asimov’s Laws of Robotics
- Application to robot ethics
  - What are the right rules?
  - How are rules applied to decisions?
  - What design rules to achieve our desired social goals?

Burton et al. 2017

Ethics Approaches: Utilitarianism (aka Consequentialism)

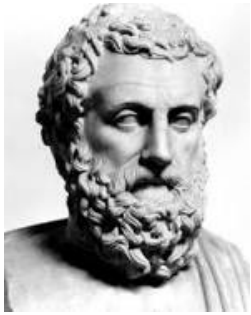
What is the greatest possible good (moral law) for the greatest number? (Bentham and Mill; 18th century)

What is the greatest possible balance of good over evil? (Frankena)

- "Utility": proxy for individual goodness
  - "Utilitarian Calculus" compares the sum of individual utility (positive or negative) over all people in society
- Computer science / Game theory
  - Utility: presentation of the individual agent's preference
  - Rationality: selecting actions maximizing expected utility

Burton et al. 2017

Ethics Approaches: Virtue Ethics



Who should I be?

- Virtue/Telelogical Ethics; Nichomachean Ethics (Aristotle)
- Local norms (not universal moral laws)
  - Organized around developing habits and dispositions that help a person achieve goals
- "Phronesis" (moral prudence, practical wisdom)
  - Ability to evaluate a given situation and respond fittingly
  - Developed through both education and experience
- Dominant approach (after Utilitarianism)

Burton et al. 2017

Philosophical positions (2500 years in one slide!)

	Consequentialism	Deontology	Virtue Ethics
Description	An action is right if it promotes the best consequences, i.e where happiness is maximized.	An action is right if it is in accordance with a moral rule or principle.	An action is right if it is what a virtuous agent would do in the circumstances.
Central Issue	The results matter, not the actions themselves	Persons must be ends in and of themselves and may never be used as means	Emphasize the character of the agent making the actions
Guiding Value	Good (often seen as maximum happiness)	Right (rationality is doing one's moral duty)	Virtue (dispositions leading to the attainment of happiness)
Practical Reasoning	The best for most (means-ends reasoning)	Follow the rule (rational reasoning)	Practice human qualities (social practice)
Deliberation Focus	Consequences (What is outcome of action?)	Action (Is action compatible with imperative?)	Motives (is action motivated by virtue?)

V. Dignum: Responsible Autonomy, IJCAI 2017

Ethics Approaches: Case Study

Robot and Frank

- *Frank is a retired jewel thief whose children get him a caretaker robot so he can stay in his home*

Stealing scene

https://www.youtube.com/watch?v=PwQgdJo1i5Y

Burton et al. 2017

Ethics Approaches: Case Study

1. Deontological ethics

2. Utilitarianism

3. Virtue ethics

Burton et al. 2017

Ethics Approaches: Case Study

1. Deontological ethics
   - Robot's duty is Frank's health (local law)
   - This supersedes all other directives (lies, stealing)
   - Care robots: possible to care for individual without violating other laws?
2. Utilitarianism
   - Social norm of theft: long-term consequences for society
   - Robot & Frank: little concern for long-term social consequences of their actions (Robot learns theft justification from Frank)
3. Virtue ethics
   - Robot is making choices according to its own particular goals and ends (to care for Frank); but different from human designers'
   - Robot lacks "phronesis", the practical wisdom to exercise nuanced judgment about how to act

Burton et al. 2017

Potential Harms of AI and Robotics

- *Bias and discrimination*
- *Denial of individual autonomy and rights*
- Non-transparent, unexplainable, or unjustifiable outcomes
- Invasions of privacy
- Isolation and disintegration of social connection
- Unreliable, unsafe, or poor-quality outcomes
- Job losses / changes

Potential Harms:
Bias and Discrimination

Google search exercise:

Chief executive officer

Potential Harms: Bias and Discrimination

- Gender bias
  - Google image search: 11% female CEO (27% US)
  - Face recognition: All classifiers perform better on male faces than female faces (also: lighter/darker)

Potential Harms: Bias and Discrimination

- Gender bias
  - Google image search: 11% female CEO (27% US)
  - Face recognition: All classifiers perform better on male faces than female faces (also: lighter/darker)
- Race bias: COMPAS Florida police system
  - Higher reoffending rate prediction for black people

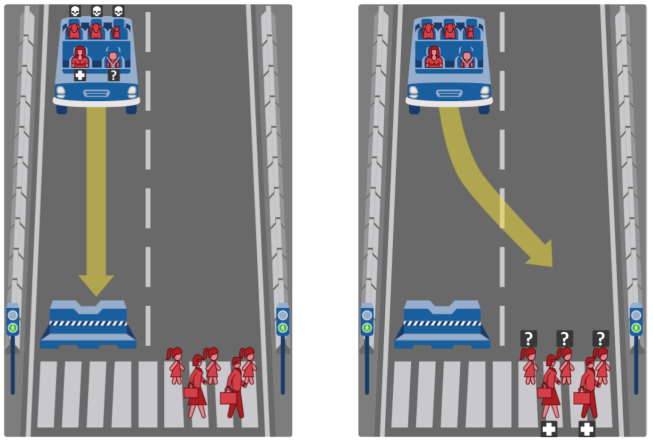
Potential Harms:
Bias and Discrimination

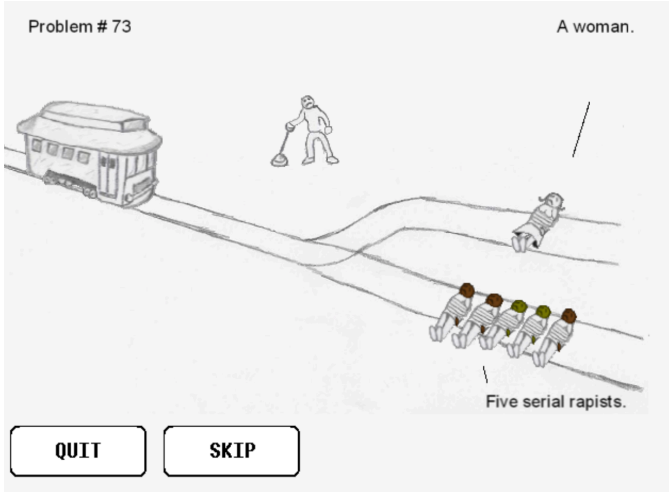
- Gender bias, race bias

How can we address/solve this?

Potential Harms: Denial of Autonomy

- Trolley Problem
- Variants: Fat man/villain, Transplant…

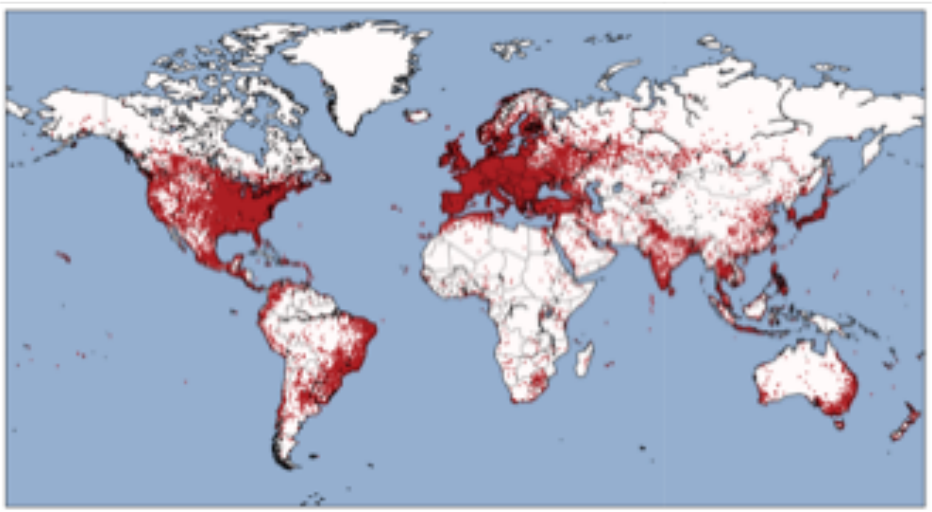




newfastuff.com/the-trolley-problem-game/

*Potential Harms: Trolley Problem Bias

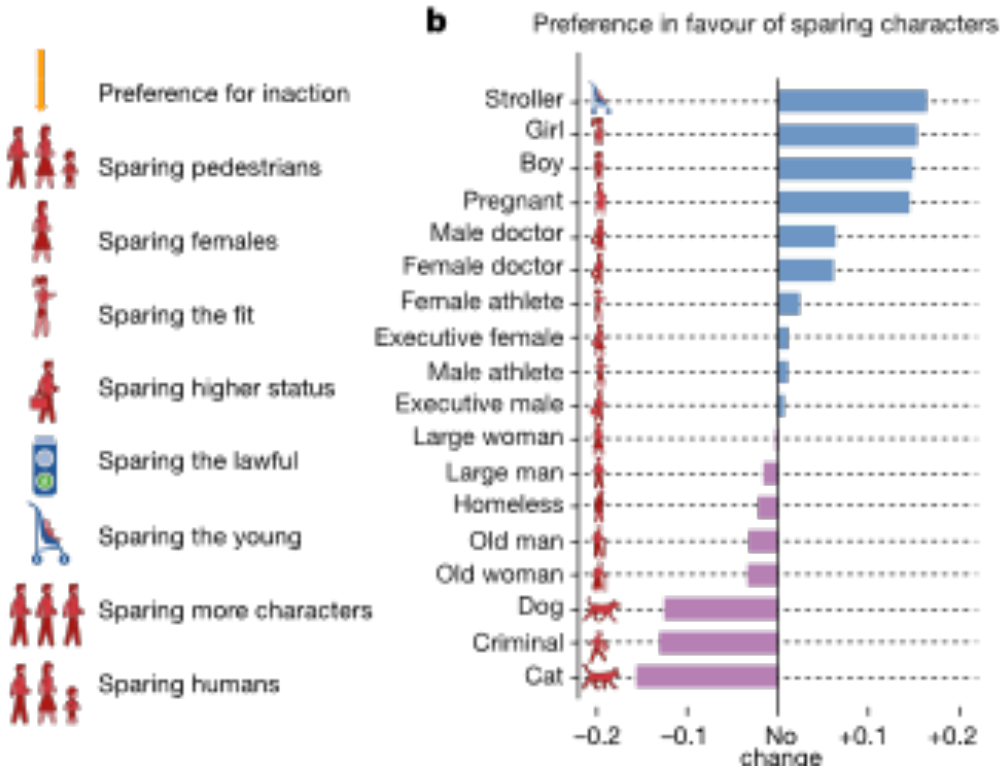
- Moral Machine (Trolley) crosscultural study
  - 40 million decisions, 233 countries



Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.F. and Rahwan, I., 2018. The moral machine experiment. *Nature*, *563*(7729), p.59. doi:s41586-018-0637-6

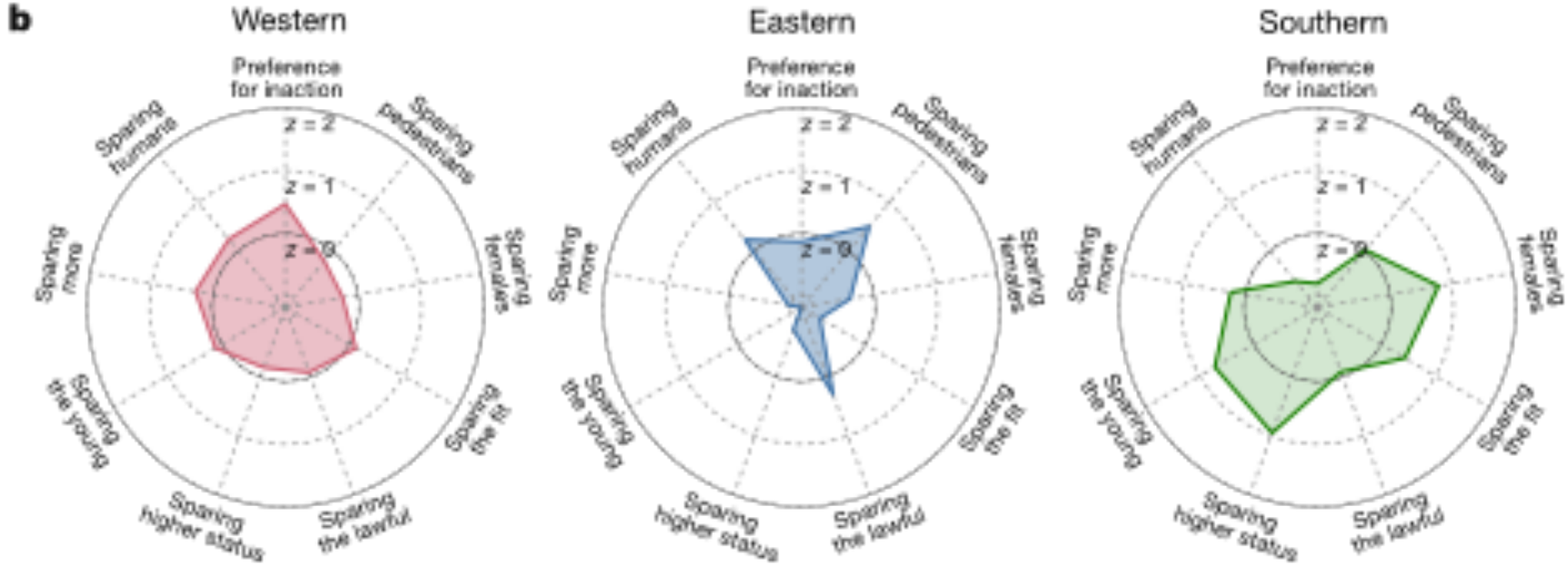
*Potential Harms: Trolley Problem Bias

- Moral Machine (Trolley) crosscultural study
  - 40 million decisions, 233 countries
  - Global similarities/differences



*Potential Harms: Trolley Problem Bias

- Moral Machine (Trolley) crosscultural study
  - 40 million decisions, 233 countries
  - Global similarities/differences
  - 3 clusters of cross-cultural ethical variations



Potential Harms of AI and Robotics

- *Bias and discrimination*
- *Denial of individual autonomy and rights*
- Non-transparent, unexplainable, or unjustifiable outcomes
- Invasions of privacy
- Isolation and disintegration of social connection
- Unreliable, unsafe, or poor-quality outcomes
- Job losses / changes

What can we do?

Machine Ethics

Machine Ethics

"An ethical machine is guided by an intrinsic ethical rule, or set of rules, in deciding how to act in a given situation."

(Anderson & Anderson 2006; Winfield et al. 2017)

To design machines/robots with intrinsic ethics rules

Winfield et al. (2019). Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems. *Proceedings of the IEEE*.
https://doi.org/10.1109/JPROC.2019.2900622

Machine Ethics: Agency

- Moor's "The nature, importance, and difficulty of machine ethics"
  1. **Ethical Impact Agents**: Any machine that can be evaluated for its ethical consequences
  2. **Implicit Ethical Agents**: Machines that are designed to avoid unethical outcomes
  3. **Explicit Ethical Agents**: Machines that can reason about ethics
  4. **Full Ethical Agents**: Machines that can make explicit moral judgments and justify them

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, *21*(4), 18-21.

Machine Ethics: Examples

- Ethical governor agent (Arkin 2009)

- GenEth system (Anderson & Anderson 2014)
- Asimovian ethical robots (Winfield et al. 2014)
- Akratic robot (Bringsjord et al. 2014)
- “sorry I can’t do that” robot (Briggs & Scheutz 2015)
- Intervening robot mediator in healthcare (Shim et al. 2017)

Winfield et al. (2019). Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems. *Proceedings of the IEEE*. Special Issue https://doi.org/10.1109/JPROC.2019.2900622

What can we do?

Responsible AI

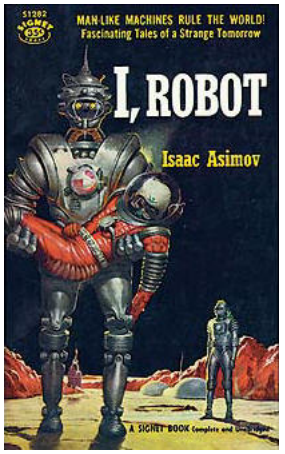
Governments' and Scientists' AI & Robotics Ethics Guidelines

- UK EPSRC Principles of Robotics (2010)
- EU Ethics Guidelines on Trustworthy AI (2018)
- EU AI Act
- JSAI Japanese Society for Artificial Intelligence Ethical Guidelines (2017)
- Asilomar AI Principles - Future of Life Institute (2017)
- Montréal Declaration for Responsible AI (2017)
- Asimov / Murphy & Wood Laws

- See: Alan Winfield's blog on Robotics and AI Ethics

alanwinfield.blogspot.com/2017/12/a-round-up-of-robotics-and-ai-ethics.html

Three Laws of Robotics Asimov (Runaround, 1942)



1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws

Three Laws of Responsible Robotics

1. A human may not deploy a robot without the human-robot work system meeting the **highest legal and professional standards** of safety and ethics.
2. A robot must **respond to humans** as appropriate for their roles.
3. A robot must be endowed with sufficient situated autonomy to **protect its own existence** as long as such protection provides smooth transfer of control which does not conflict with the First and Second Laws.

Murphy & Wood (2009). Beyond Asimov: The Three Laws of Responsible Robotics

Responsible AI approaches

- Interdisciplinary team with in-built culture
- Governance architecture (Platform)
- Objectives for ethical, fair and safe AI
  - **ethically permissible** for impact on wellbeing of affected stakeholders and communities
  - **fair and non-discriminatory** on individuals/ groups
  - **public trust** by guaranteeing safety, accuracy, reliability, security, and robustness
  - **justifiable** by prioritising both the transparency of the designed and the interpretability of its decisions

Responsible AI approaches/methods

- Turing Institute's Ethical Platform
  - https://www.turing.ac.uk/research/publications/understanding-artificial-intelligence-ethics-and-safety
- Responsible Research and Innovation (RRI)
  - https://epsrc.ukri.org/research/ourportfolio/themes/healthcaretechnologies/strategy/toolkit/home/integrity/ri/
  - https://www.rri-tools.eu/about-rri
- Ethics by Design
  - https://orbilu.uni.lu/bitstream/10993/38926/1/p60-dignum.pdf

RRI
Responsible Research & Innovation

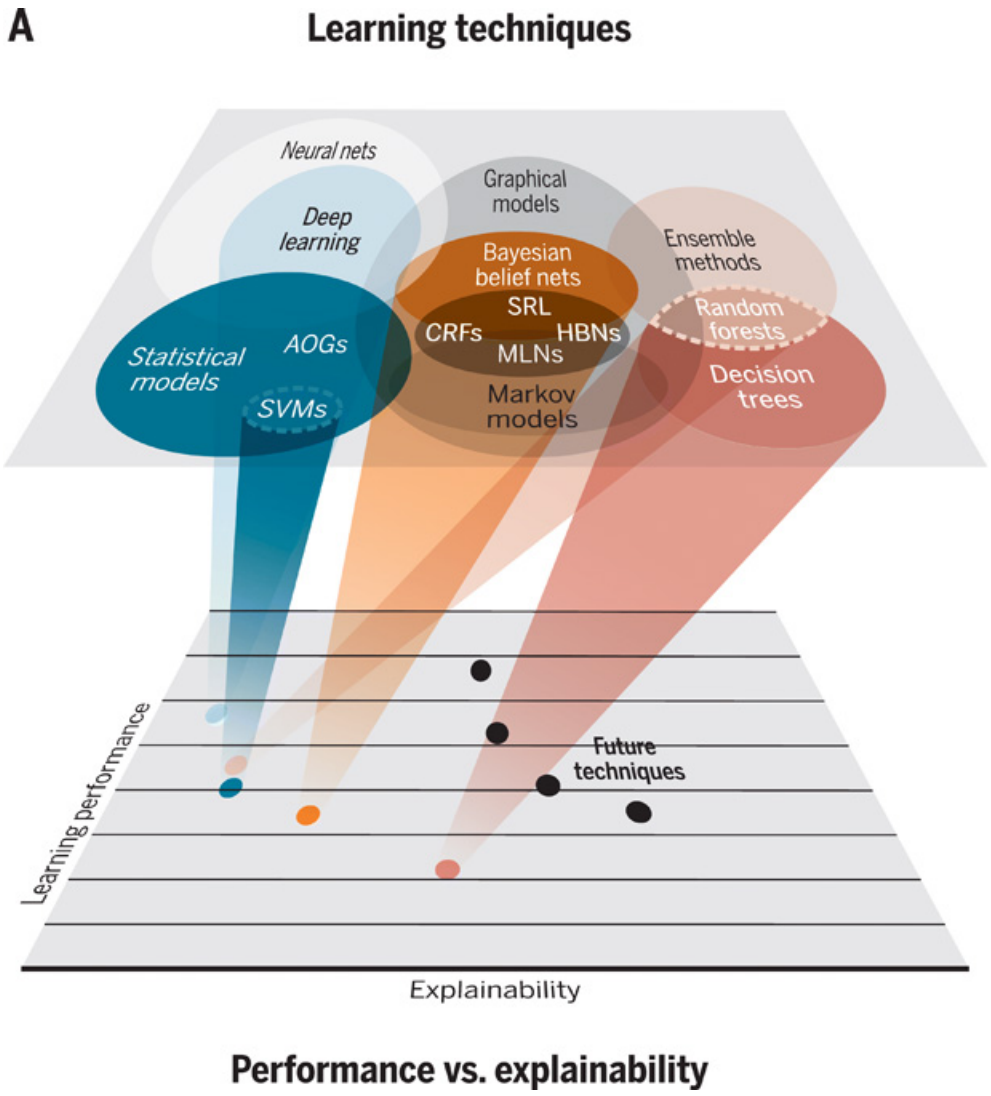
- Involve society in science and innovation ‘very upstream' in the processes, to align its outcomes with the values of society
  - Focus group, RRI expert embedding, “embedded ethics” (Stahl & Coeckelbergh 2016)
- A wide umbrella connecting different aspects of the relationship between R&I and society:
  - public engagement, open access, gender equality, science education, ethics, governance

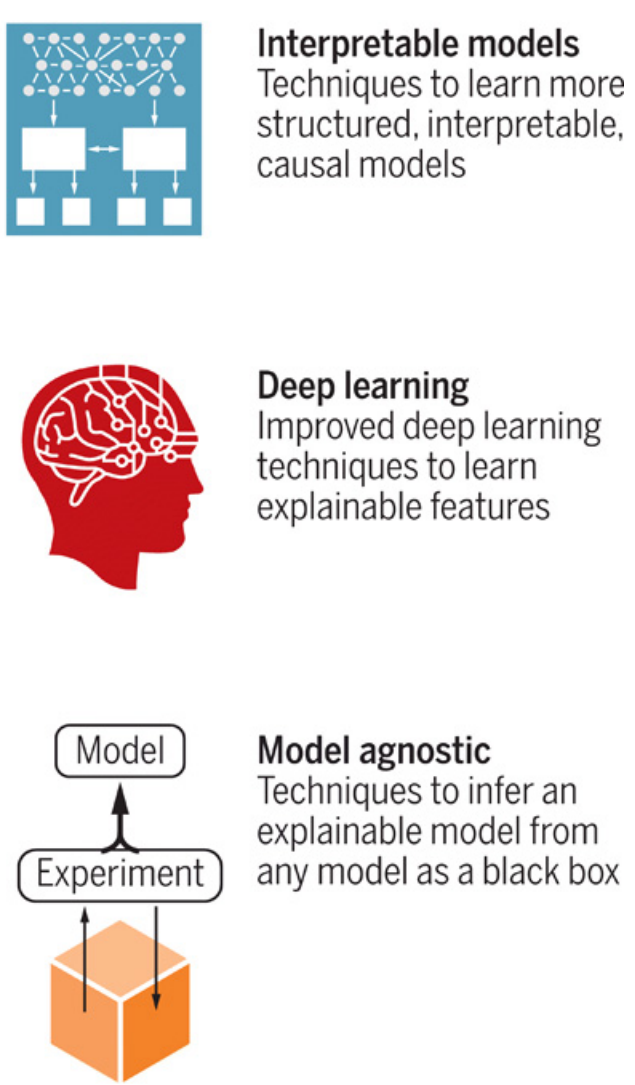
Explainable AI (XAI) and Robotics

- Aims
  - Produce explainable models
  - Enable human users to understand, trust, and effectively manage AI systems
- Performance vs explainability
  - Blackbox (Deep learning) problem and solutions (Samek et al. 2017)
  - Explainable dialog for HRI
  - Short overview paper Gunning et al. 2019

https://www.darpa.mil/program/explainable-artificial-intelligence

Explainable AI (XAI) and Robotics





Gunning et al. 2019

Summary

- Ethical problems for AI and Robots
- Two problems: AI ethics and Machine ethics
- Moral thinking and ethics approaches
- Potential Harms
- What can we do?

- Reading (optional, e.g. choose one)
  - Leslie: AI ethics doi.org/10.5281/zenodo.3240529
  - Winfield: Machine ethics doi.org/10.1109/JPROC.2019.2900622
  - Winfield's Blog and Robotics and AI Ethics