The five principles

Imperial College London

Principles

We saw earlier how medical ethics reduces the infinitely complex space of ethical medical care down to just four principles (Beauchamp, Childress, et al., 2001):

- Non-maleficence
- Beneficence
- Equity
- Autonomy

Imperial College London

Is AI ethics fundamentally different?

1. Different countries take different attitudes on the relative importance of patient autonomy versus public health, e.g., regarding compulsory vaccination.
2. Can a patient really give "informed consent" on complex medical matters? Do they understand the tradeoffs well?

1. Digital contact tracing posed the same dilemma between privacy (philosophically a special case of autonomy)and public health.
2. When a loan applicant consents to a digital credit check, do they really understand what that means?

Imperial College London

Is AI ethics fundamentally different?

Floridi and Cowls, 2019 argue that AI poses only a single "new" ethical problem, that of *explainability*, because decisions are made by an algorithm, not a human.

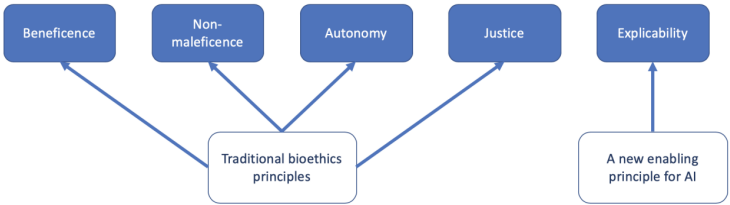

Figure 18: Figure from "A Unified Framework of Five Principles for AI in Society"

Let us spend a few minutes thinking about this.

Imperial College London

Is AI ethics fundamentally different?

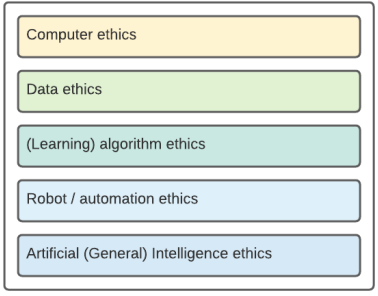

Figure 19: Intersections of ethical frameworks

Imperial College London

Is AI ethics fundamentally different (sci-fi interlude)?

Should AIs themselves have rights?

- Do we know enough about consciousness to rule out that sufficiently advanced AIs might suffer?
- As robots become more anthropomorphic, might abusing them lead to toxic psychological habits?

Photo Illustration: Photos: Steve Gorton and Tim Ridley, Alexander Heimann/Getty Images

Figure 20: It might be OK for your cat to take a ride on your roomba but is it OK for you to kick this small household robot if it fails to clean a part of your living room? These questions form part of *robot ethics*.

Is AI ethics fundamentally different (sci-fi interlude)?

The possibility of AI "escaping control" is a favorite with popular science and fiction:

- *Possible minds: 25 ways of looking at AI*, edited by J. Brockman
- *Human Compatible*, S. Russell
- *Life 3.0*, M. Tegmarck
- *Superintelligence*, N. Bostrom

Artificial General Intelligence

The term AGI is reserved for AI systems that are able to solve arbitrary cognitive problems, as opposed to narrow AI which might be able to have super-human performance in, say, chess, but would not know what to do if you asked it to cook dinner. Whether or not this is achievable, and whether, if achievable, it would likely coincide with the emergence of some kind of consciousness, are open questions.

Imperial College London

From science fiction to real life

AGI might be far away (or not!) but it does invite us to think about:

- Control: how to keep human oversight on autonomous decision-making systems?
- Value alignment: how can we ensure AIs are optimising for the right thing?

Existential risk

An extreme case of lack of value alignment is when a technology poses a risk so high that it could trigger a widespread humanitarian crisis, or even, at the extreme, an extinction level event. Nuclear power falls in this category, but arguably so do:

- Deep fakes and their risk to democracy
- AI for autonomous weapons, including for cyber-warfare

Imperial College London

From science fiction to real life

Autonomous weapons

One clear-cut example of a red line is research in AI for autonomous weapons. Many tech companies and prominent figures have publicly declared a moratorium on this. Moratoria have also been declared on other use cases with subtler risk profile, such as face recognition.

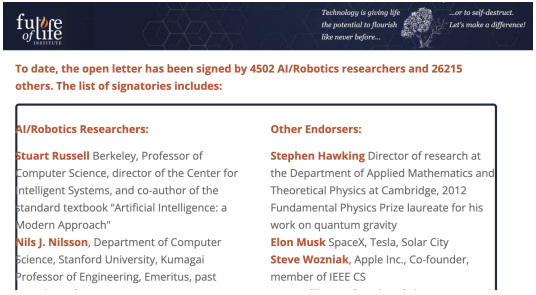

Figure 21: The Future of Life's "Open Letter" has been signed by 4.5K AI researchers and more than 25K other stakeholders.

Imperial College London

Five principles for AI: our version

1. Privacy and autonomy
2. Fairness and non-discrimination
3. Safety and security
4. Explainability and accountability
5. Value alignment and control

1. Autonomy
2. Equity
3. Non-maleficence
4. Explainability
5. Beneficence

Imperial College London

Summary

- AI poses new versions of the same challenges present in other areas of ethics, such as medical ethics, data ethics and technology ethics.
- It also poses some altogether new challenges, like explainability.
- The debate around artificial general intelligence and the risk of it being too powerful to be safe echoes the concern around nuclear power, and is independent of the (also interesting) debate around whether AGI will be conscious
- Although AGI might seem futuristic, there are pragmatic, imminent societal risks posed by powerful AI, such as autonomous weapons and deep fakes.
- To better carve up the technical content into coherent units, we will use our own factorisation of AI ethics into five principles.

Each of the following five weeks is dedicated to one such principle.

Imperial College London

Bibliography I

Beauchamp, Tom L, James F Childress, et al. (2001). *Principles of biomedical ethics*. Oxford University Press, USA.

Floridi, Luciano and Josh Cowls (2019). "A unified framework of five principles for AI in society". In: *Available at SSRN 3831321*.

Mitchell, Shira et al. (2021). "Algorithmic Fairness: Choices, Assumptions, and Definitions". In: *Annual Review of Statistics and Its Application* 8.1, pp. 141–163. DOI: 10.1146/annurev-statistics-042720-125902. URL: https://doi.org/10.1146/annurev-statistics-042720-125902.

Naudts, L. (2019). *Towards a Code of Ethics for Artificial Intelligence by Paula Boddington*. Vol. 2. 2. Springer, pp. 105–106. DOI: 10.21552/delphi/2019/2/11.

O'neil, Cathy (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.