Imperial College London

Ethics of Machine Learning and Data Science - Part I

Week 1: Foundations

Dr Chris Anagnostopoulos

Imperial College London

Table of Contents

Do no harm

Imperial College London

AI is wonderful

Figure 1: https://deepmind.com/research/case-studies/alphago-the-story-so-far

Imperial College London

AI is wonderful

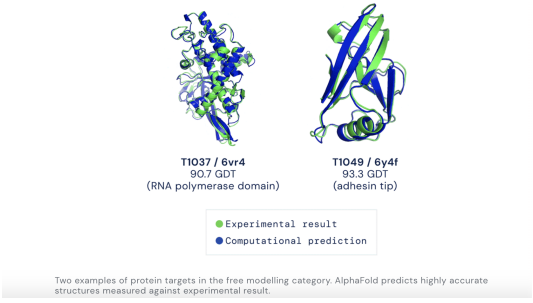
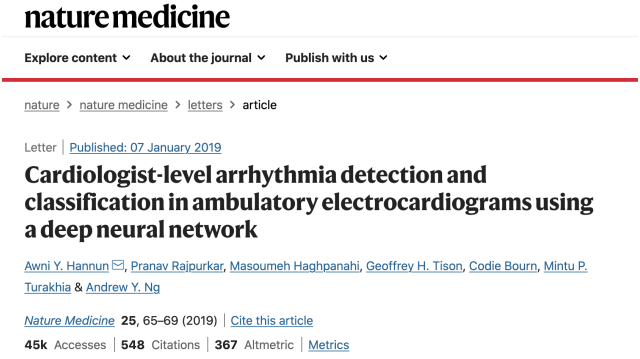


Figure 2: https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

Imperial College London

AI is wonderful



nature medicine

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature medicine › letters › article

Letter | Published: 07 January 2019

Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network

Awni Y. Hannun ✉, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia & Andrew Y. Ng

Nature Medicine **25**, 65–69 (2019) | Cite this article

45k Accesses | **548** Citations | **367** Altmetric | Metrics

Figure 3: https://www.nature.com/articles/s41591-018-0268-3

Imperial College London

AI can cause harm

Figure 4: Left: https://www.bbc.co.uk/news/technology-56799749, Right: https://www.bbc.co.uk/news/technology-54175359

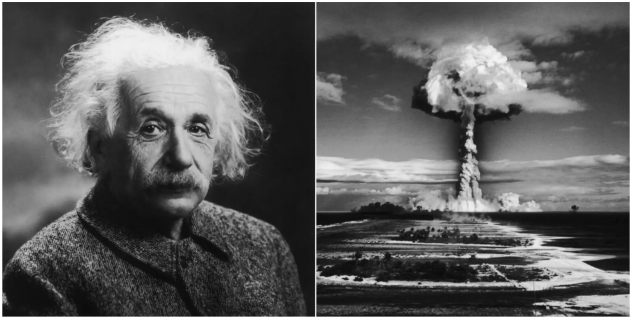
Imperial College London

AI can cause harm

Figure 5: Left: https://www.bbc.co.uk/news/technology-58462511, Right: https://www.bbc.co.uk/news/technology-50865437

Imperial College London

Technological adoption relies on public trust



Albert Einstein around 1939 and nuclear explosion in French Polynesia in October 1971. MPI/Getty Images/Michel BARET/Gamma-Rapho via Getty Images

Figure 6: Einstein was initially in favour of the creation of an atomic bomb, but later became vehemently opposed to the proliferation of nuclear arms.

Imperial College London

Technological adoption relies on public trust

Figure 7: Left: The Chernobyl nuclear reactor (photo by Associated Press), Right: the Fukushima nuclear accident. https://www.bbc.co.uk/news/world-asia-56252695

Imperial College London

Technological adoption relies on public trust

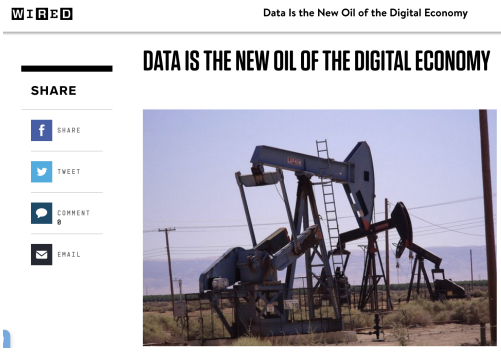


Figure 8: The onset of Big Data circa 2014 has ushered a new era for Data Science.

- DS and AI are becoming ubiquitous and are gradually transforming the majority of economic, social and political interactions.
- They are too complex to understand without specialised training.
- With understanding, comes responsibility.

Imperial College London

Technological adoption relies on public trust

The collapsed 35W bridge in Minneapolis seen on August 2, 2007.
Morry Gash/AP

Figure 9: Thirteen people were killed and 145 injured during the Interstate 35W bridge collapse in 2007

Many other professions carry a similar burden of responsibility, abide by strict codes of ethics and are legally liable for the outcomes of their work:

- Doctors
- Pharmaceutical companies
- Lawyers
- Safety-critical systems engineers
- ...

Imperial College London

The Hippocratic oath

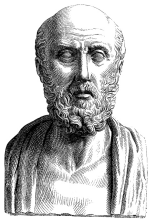


Figure 10: Hippocrates. Unidentified engraver, Public domain, via Wikimedia Commons

First, do no harm

Doctors around the world swear a version of the Hippocratic oath, originating in 400 BCE, and introducing the principles of medical confidentiality and non-maleficence.

Corporate oaths

Google's original internal code of conduct started with the phrase "Don't be evil". This was later rephrased into "You *can* make money without being evil."

The Hippocratic oath

Figure 11: Cathy O'Neill in *Weapons of Math Destruction* (O'neil, 2016) was among the first to make a call for a hippocratic oath for data scientists.

Imperial College London

Doing the right thing is neither obvious nor easy

- Lack of context or poor understanding of the group at-risk.
- Cognitive biases or bad habits.
- Inappropriate incentive structures.
- Inherent trade-offs (moral dilemmas).
- Defeatism.
- Unanticipated consequences.

What is the right thing to do?

We will not attempt to define this here, as this is not a course in moral philosophy. But I hope it will inspire you to take one, or read a book about it(Naudts, 2019).

Imperial College London

Hope for the best, prepare for the worst

Prevalent mindset

1. Success stories / impact
2. Access to data
3. Do good, not evil
4. Someone else will fix it
5. Unsolvable dilemma
6. If used properly, it is safe
7. Technology will fix everything
8. Only monitor model performance
9. Keep it between the engineers

Target mindset

1. Near-misses, anticipate harm
2. Permission to use data
3. Do the best you can, minimise risk
4. You hold the bar
5. Explicit trade-offs
6. How could this be abused?
7. Pragmatism and humility
8. Monitor everything we care about
9. Engage with broader society

Imperial College London

Summary

- With power comes responsibility
- You are best placed to drive positive change
- Anticipate, minimise, communicate
- Human-centred design
- You won't be able to fix everything

Next, we start to become more specific:

- What are we talking about really?
- What progress has been made?
- What principles should guide the design of ethical AI?