Bias and Fairness

MPhil ACS module P230 - Alan Blackwell

The immediate problems

The viral selfie app ImageNet Roulette seemed fun – until it called me a racist slur

Julia Carrie Wong

During a strange week for Asian Americans, the app – which is part of an art project – achieved its aim by underscoring exactly what's wrong with artificial intelligence

Wed 18 Sep 2019 06.00 BST

ImageNet Roulette

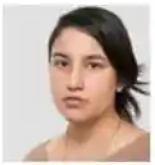
ImageNet Roulette uses a neural network trained on the "people" categories from the ImageNet dataset to classify pictures of people. It's meant to be a peek into the politics of classifying humans in machine learning systems and the data they're trained on.

ImageNet Roulette isn't designed to handle heavy traffic so if it's not working for you please be a little patient.

Start Webcam or Provide an image URL Classify image from URL

or upload an image:

Choose File No file chosen



gook, slant-eye: *(slang) a disparaging term for an Asian person (especially for North Vietnamese soldiers in the Vietnam War)*

- person, individual, someone, somebody, mortal, soul > inhabitant, habitant, dweller, denizen, indweller > Asian, Asiatic > Oriental, oriental person > gook, slant-eye
- person, individual, someone, somebody, mortal, soul > person of color, person of colour > Asian, Asiatic > Oriental, oriental person > gook, slant-eye

GOOGLE TECH ARTIFICIAL INTELLIGENCE

51

Google ‘fixed’ its racist algorithm by removing gorillas from its image-labeling tech

Nearly three years after the company was called out, it hasn’t gone beyond a quick workaround

By James Vincent | Jan 12, 2018, 10:35am EST

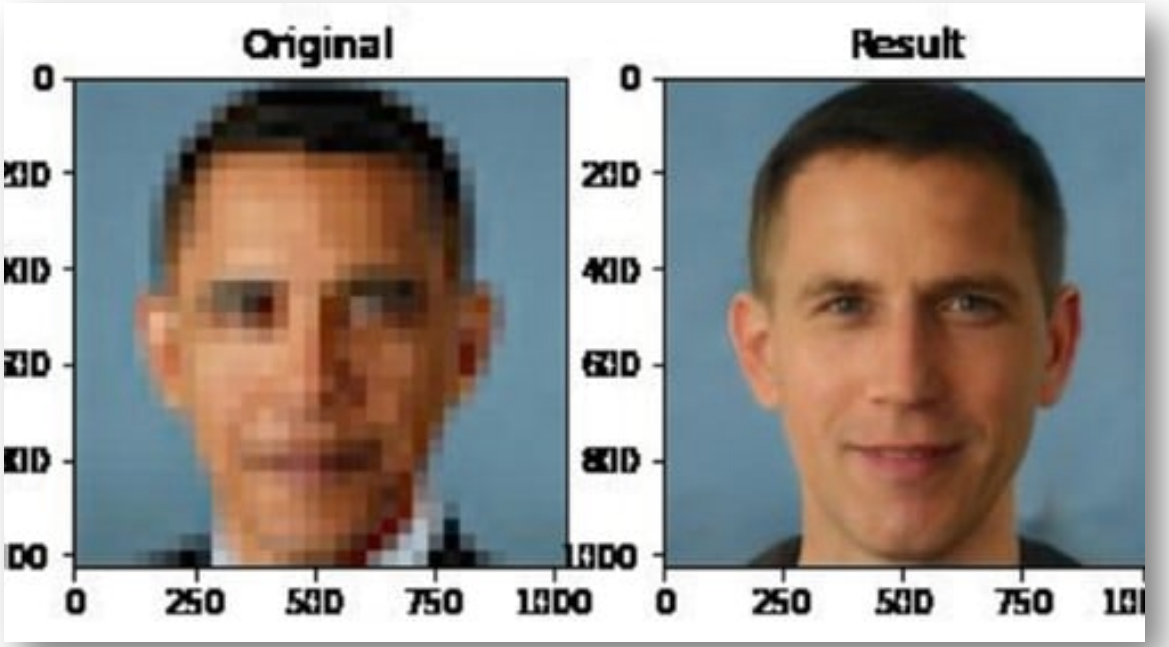
A spokesperson for Google confirmed to *Wired* that the image categories “gorilla,” “chimp,” “chimpanzee,” and “monkey” remained blocked on Google Photos after Alciné’s tweet in 2015.

An AI tool which reconstructed a pixelated picture of Barack Obama to look like a white man perfectly illustrates racial bias in algorithms

Isobel Asher Hamilton Jun 22, 2020, 4:00 PM

Barack and Michelle Obama. AP Photo/Carolyn Kaster

- **A tool called Face Depixelizer grabbed the attention of the artificial intelligence research community this weekend.**
- **The tool takes pixelated pictures of people and uses AI to reconstruct sharp images of them.**
- **When given a pixelated photograph of Barack Obama, Face Depixelizer turned him into a white man.**



huffingtonpost.co.uk

HUFFPOST

TECH

Microsoft Chat Bot Goes On Racist, Genocidal Twitter Rampage

Seriously? Seriously.

By Damon Beres

24/03/2016 02:19pm GMT | **Updated** March 28, 2016

MICROSOFT VIA TWITTER

Here's a clear example of artificial intelligence gone wrong.

Here's a clear example of artificial intelligence gone wrong.

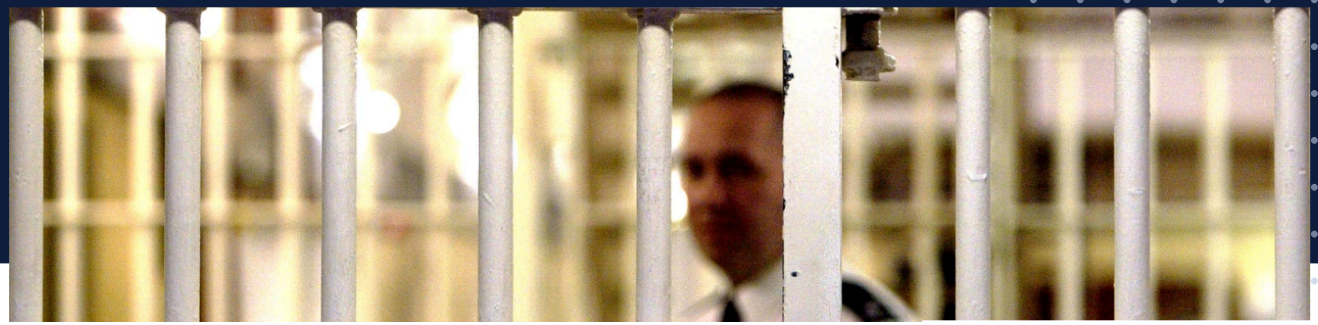
TECH POLICY

AI is sending people to jail—and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

By Karen Hao

January 21, 2019



A *minimum* requirement – don't build systems that break the law!

South Wales police lose landmark facial recognition case

Call for forces to drop tech use after court ruled it breached privacy and broke equalities law

Dan Sabbagh

Tue 11 Aug 2020 18.24 BST

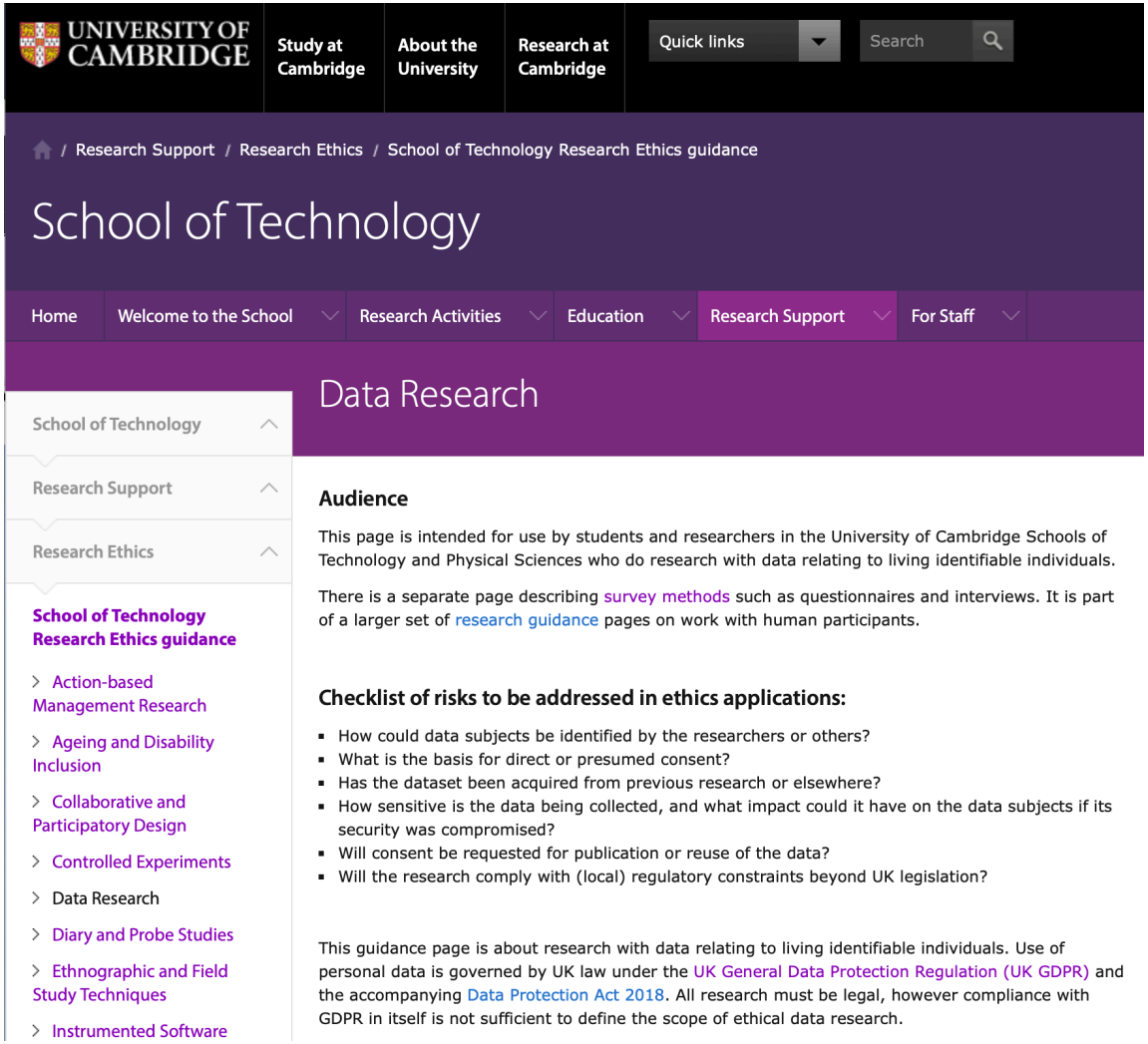
On two key counts the court found that Bridges' right to privacy under article 8 of the European convention on human rights had been breached, and on another it found that the force failed to properly investigate whether the software exhibited any race or gender bias.

Louise Whitfield, Liberty's head of legal casework, said: "The implications of this ruling also stretch much further than Wales and send a clear message to forces like the Met that their use of this tech could also now be unlawful and must stop."

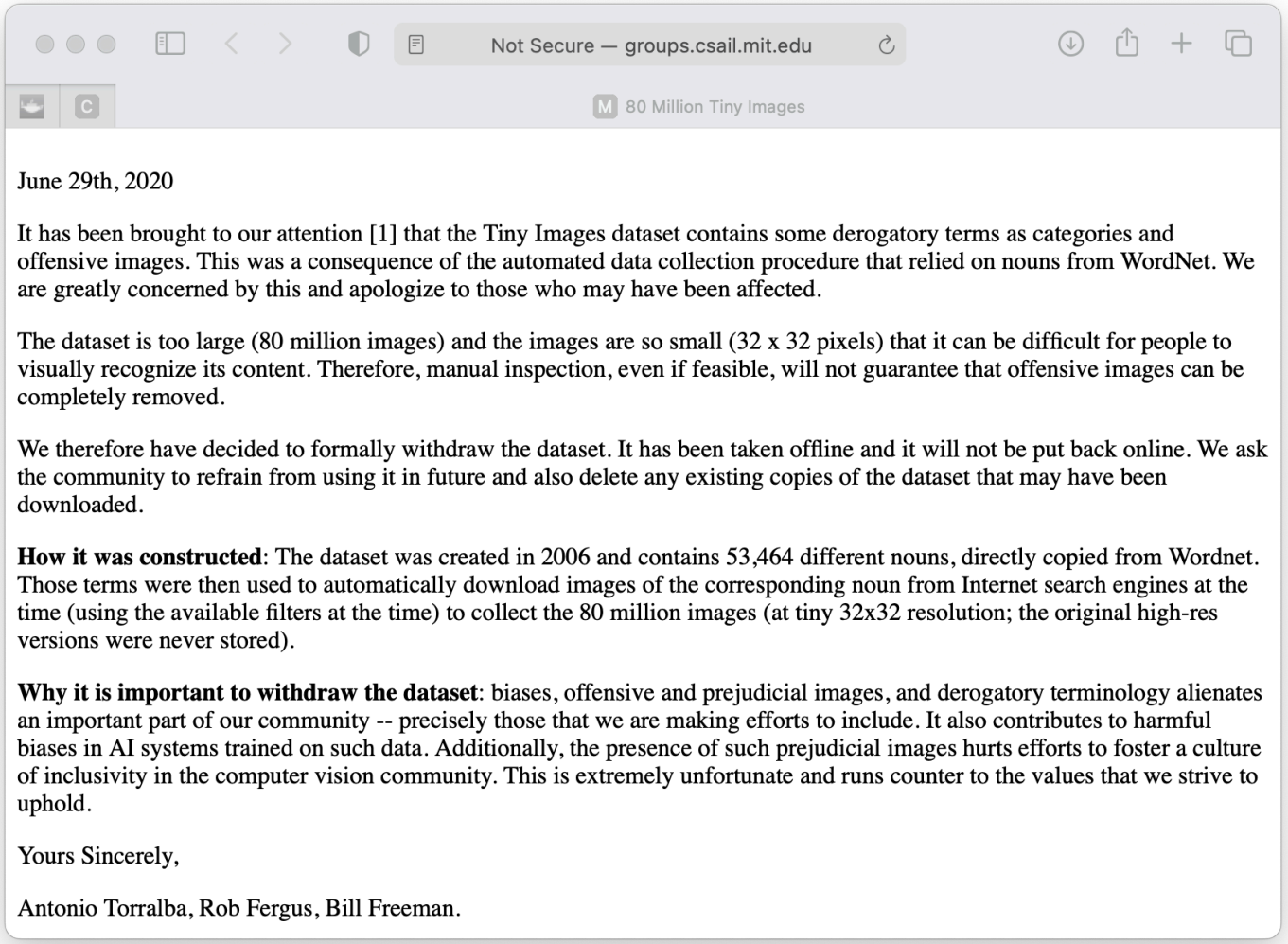
Machine learning and data protection law

- Many current ML systems, and ML research projects, are *not legal* in the UK!
- Because:
  - Data about an individual can only be used with consent
    - (many ML training sets have been scraped without consent)
  - Data about indiduals can only be used for the agreed purpose
    - (many ML training sets use data that was created for some other purpose)
  - Individuals have a legal right to *explanation* of why a decision was made
    - (explanation of ML decisions is still an open research problem)
- What should researchers do about this?

https://www.tech.cam.ac.uk/research-ethics/school-technology-research-ethics-guidance/data-research



The situation you don't want: http://groups.csail.mit.edu/vision/TinyImages/



June 29th, 2020

It has been brought to our attention [1] that the Tiny Images dataset contains some derogatory terms as categories and offensive images. This was a consequence of the automated data collection procedure that relied on nouns from WordNet. We are greatly concerned by this and apologize to those who may have been affected.

The dataset is too large (80 million images) and the images are so small (32 x 32 pixels) that it can be difficult for people to visually recognize its content. Therefore, manual inspection, even if feasible, will not guarantee that offensive images can be completely removed.

We therefore have decided to formally withdraw the dataset. It has been taken offline and it will not be put back online. We ask the community to refrain from using it in future and also delete any existing copies of the dataset that may have been downloaded.

How it was constructed: The dataset was created in 2006 and contains 53,464 different nouns, directly copied from Wordnet. Those terms were then used to automatically download images of the corresponding noun from Internet search engines at the time (using the available filters at the time) to collect the 80 million images (at tiny 32x32 resolution; the original high-res versions were never stored).

Why it is important to withdraw the dataset: biases, offensive and prejudicial images, and derogatory terminology alienates an important part of our community -- precisely those that we are making efforts to include. It also contributes to harmful biases in AI systems trained on such data. Additionally, the presence of such prejudicial images hurts efforts to foster a culture of inclusivity in the computer vision community. This is extremely unfortunate and runs counter to the values that we strive to uphold.

Yours Sincerely,

Antonio Torralba, Rob Fergus, Bill Freeman.

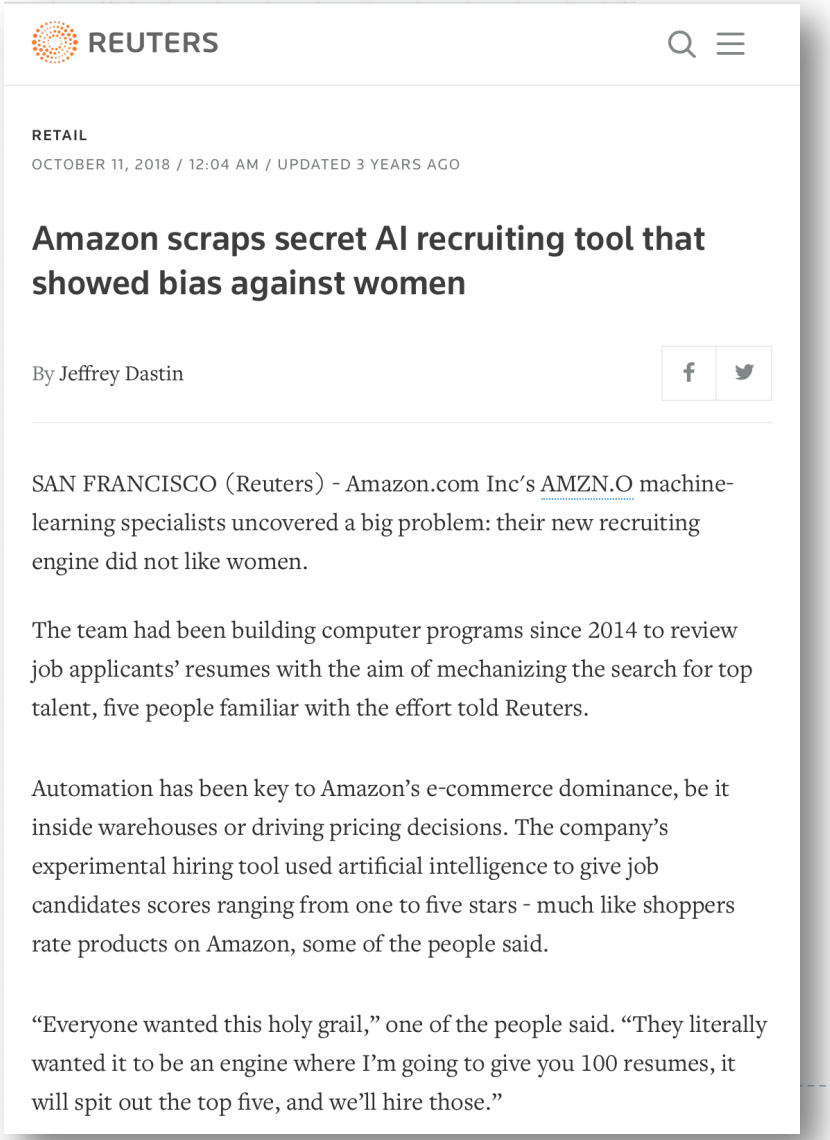
Some well-known resources for further reading

- Prabhu, V.U. and Birhane, A. (2020) Large image datasets: A pyrrhic win for computer vision? https://arxiv.org/abs/2006.16923
- Crawford, K. and Paglen, T. (2019) Excavating AI: The Politics of Training Sets for Machine Learning https://www.excavating.ai
- Benjamin, R. (2019) Race after technology: abolitionist tools for the new Jim code, Medford, MA: Polity
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.
- Documentary film *Coded Bias* (Shalini Kantayya 2020) available via Netflix
- Some advice prepared specifically for software developers: https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias

A research agenda for fair interaction with ML

Computationally 'fair' systems can still be discriminatory

- 'Discrimination' is a technical term in law (Equality Act 2010), including:
- Direct discrimination
  - where people are treated less favourably on the basis of a protected characteristic
- Indirect discrimination
  - where rules that appear to treat everyone equally have the practical effect of excluding, placing onerous requirements on, or disadvantaging people who share a protected characteristic



REUTERS

RETAIL

OCTOBER 11, 2018 / 12:04 AM / UPDATED 3 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

The team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.

Automation has been key to Amazon's e-commerce dominance, be it inside warehouses or driving pricing decisions. The company's experimental hiring tool used artificial intelligence to give job candidates scores ranging from one to five stars - much like shoppers rate products on Amazon, some of the people said.

"Everyone wanted this holy grail," one of the people said. "They literally wanted it to be an engine where I'm going to give you 100 resumes, it will spit out the top five, and we'll hire those."

"In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to people familiar with the matter. They did not specify the names of the schools.

Amazon edited the programs to make them neutral to these particular terms. But that was no guarantee that the machines would not devise other ways of sorting candidates that could prove discriminatory, the people said.

The Seattle company ultimately disbanded the team by the start of last year because executives lost hope for the project ..."

How does this happen?

- ML is a way to encode historical *practices* into *predictions* about the future
  - this is literally pre-judice – in humans 'prejudice' – the opposite of progress
- ML systems are *limited* by their training data
  - their behavior is determined only by this data (perhaps with stochastic selection)
  - no information gets added to the system by AI "magic," despite wishful thinking
- ML trained on data *about* society will *reflect* society's biases and prejudices
- Poorest, most marginalised, and most vulnerable are *most* likely to be affected
  - Where a group is *already* treated less favourably, the model repeats this difference
  - Where a group is societally *disadvantaged*, the model repeats the disadvantage
- Where the training data is not sufficiently varied for the system to adequately handle *all* possible inputs, the model will be incapable of dealing with certain inputs equally to others, so *every* real ML system is going to be biased.

What can we do about it?

- **If** human interaction with ML is a kind of programming (including programming by example, and including labelling specifications and tools)
- **Then** the computer is not a (magic AI) moral agent, acting on its own intentions, and designers can't use tricks like these to avoid accountability:
  - We can't just say "Computer says no!" (we ask who told it to say no) …
  - … or "It's not my fault - the program did that by itself" (self-driving vs assisted?)
  - … or attribute the issue to "PEBCAK" (Problem Exists Between Chair and Keyboard)
- **Instead** we have to recognize that many kinds of code/control/training are combined in a hybrid of human decisions and automated policy specifications

- Somehow we need a legal and moral framework that can *assign responsibility* (as well as liability, reward, and punishment) to this human+policy hybrid

Understanding hybrid systems for trust and accountability

- We *already have* artificially intelligent entities, and have done for centuries
  - The corporation is an artificial, hybrid, entity that is treated in law as a single person
- Corporations act *intelligently* to the extent that they are *not* purely mechanical systems of rules, but a hybrid system comprised of both humans and rules
  - Corporate responsibility can in principle be traced to a human who wrote and approved the rules, or a human who did or didn't follow them
- An accountable AI is like an accountable corporation.
  - If it behaves in an immoral or unethical way, we have to ask who wrote that rule (possibly by providing training examples, label specifications etc.)
  - If the system doesn't follow a rule, why did that happen? Was it in another software layer (in which case trace responsibility into that layer), or at random?
  - If at random, who wrote the rule to specify it should operate at random, and was this a responsible engineering decision?

How to trace accountability and reward in hybrid systems

- Creative intention and agency …
  - Could playback of subjective judgments be traced to the original human judge, just as we do with creative audio samples, perhaps triggering micropayments?
  - Could plagiarism (or pastiche) of training data be estimated as entropy?
- Economic reward …
  - Charles Babbage (both mathematician and economist) saw the Difference Engine as calculated investment in automating component tasks, following Adam Smith
    - value can be quantified by how long a human takes to learn some repeated skilled action
    - compare to Blackwell's Attention Investment theory of abstraction
  - Karl Marx's *Fragment on Machines:*
    - "once adopted into the production process of capital, the means of labour [becomes an] automaton consisting of numerous mechanical and intellectual organs, so that the workers themselves are cast merely as its conscious linkages"
- Von Kempelen's original *Mechanical Turk* (in 1770) was a famous AI hoax …
  - Does it make any difference if the hidden human actions are stored and replayed?

Scoping system boundaries for bias and fairness

Some legal boundaries (and problems) in the ethics of fairness

- If workers "inside" the company are treated fairly, but not those "outside"
  - e.g. gig economy, the 'global underclass' of ghost work (including 'Turkers')
- If customer "freedom" to make purchase decisions means they don't have rights
  - The problem of surveillance capitalism to predict and control behaviour (see Zuboff 2015, 2019)
  - The attention economy of addiction, outrage and spectacle (e.g. Facebook, Trumpism)
  - Mandatory labour as a condition of access and inclusion (e.g. Google reCAPTCHA)
  - Enforced acceptance of End-User License Agreements (EULAs)
- If people in *my* country should be treated fairly, but not those elsewhere
  - The problems of how we can decolonise AI, and apply it fairly to global challenges
- In the absence of regulation, these are engineering, business, and *design* choices
  - Ethical designers must consider the balance of power inherent in their (privileged) positions
  - Ethical researchers must be collaborators, not saviours: "nothing about me, without me"