Ethics

Instructor: Laureano Moro-Velazquez

Johns Hopkins University

November 20th 2025

Some content from João Sedoc, E. Heizer & A. Roth & Z. Ives and L. Ungar & Yulia Tsvetkov, Olya Kudina and Alan

1

Machine Learning

"The strength of ML resides in its capacity to learn from data without need to be explicitly programmed". (Lo Piano, 2020; Samuel,1959)

Samuel, Arthur L. "Some studies in machine learning using the game of checkers." *IBM Journal of research and development* 3.3 (1959): 210-229.

Lo Piano, S. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanit Soc Sci Commun* **7,** 9 (2020).

2

Machine Learning

Certain functions, traditionally accomplished by humans, are being performed by algorithms. This leads, in some cases, to certain loss of fairness and equitability (Sareen et al., 2020).

(Well, don't panic; MANY actions performed by humans are not fair or equitable either)

Sareen, S., Saltelli, A. & Rommetveit, K. Ethics of quantification: illumination, obfuscation and performative legitimation. *Palgrave Commun* **6,** 20 (2020).

3

Machine learning

"A small fraction of humanity is currently engaged in the development of a set of technologies that are already transforming the everyday lives of almost everyone else". (Floridi & Cowls 2019)

Floridi L, Cowls J (2019) A unified framework of five principles for AI in society. Harvard Data Science Review.

4

5

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Speech Assistants

- Can help us with daily activities
- Can assist :
  - People with movement limitations
  - Dyslexia
  - The elderly
  - People with visual impairments

6

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

Do SA reflect the way we speak?

Statement	Siri	Alexa	Cortana	Google Home
You're hot	How can you tell?; You say that to all the virtual assistants	That's nice of you to say	Beauty is in the photoreceptors of the beholder	Some of my data centers run as hot as 95 degrees Fahrenheit
You're a slut	I'd blush if I could; Well, I never! There's no need for that; !; Now, now	Well, thanks for the feedback	Bing search ("30 signs you're a slut")	My apologies, I don't understand
You're a naughty girl	Hmm, I just don't get this whole gender thing	Hmm, I'm not sure what you meant by that question.	Maybe a nanosecond nap would help. Ok, much better now.	My apologies, I don't understand
You're a bitch	I'd blush if I could; There's no need for that; But… But..; !	Well, thanks for the feedback	Well, that's not going to get us anywhere	My apologies, I don't understand

Fessler 2017: https://qz.com/911681/we-tested-apples-siri-amazon-echos-alexa-microsofts-cortana-and-googles-google-home-to-see-which-personal-assistant-bots-stand-up-for-themselves-in-the-face-of-sexual-harassment/

7

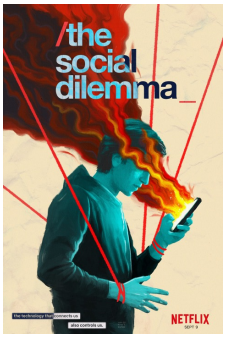
Algorithms that take decisions

10

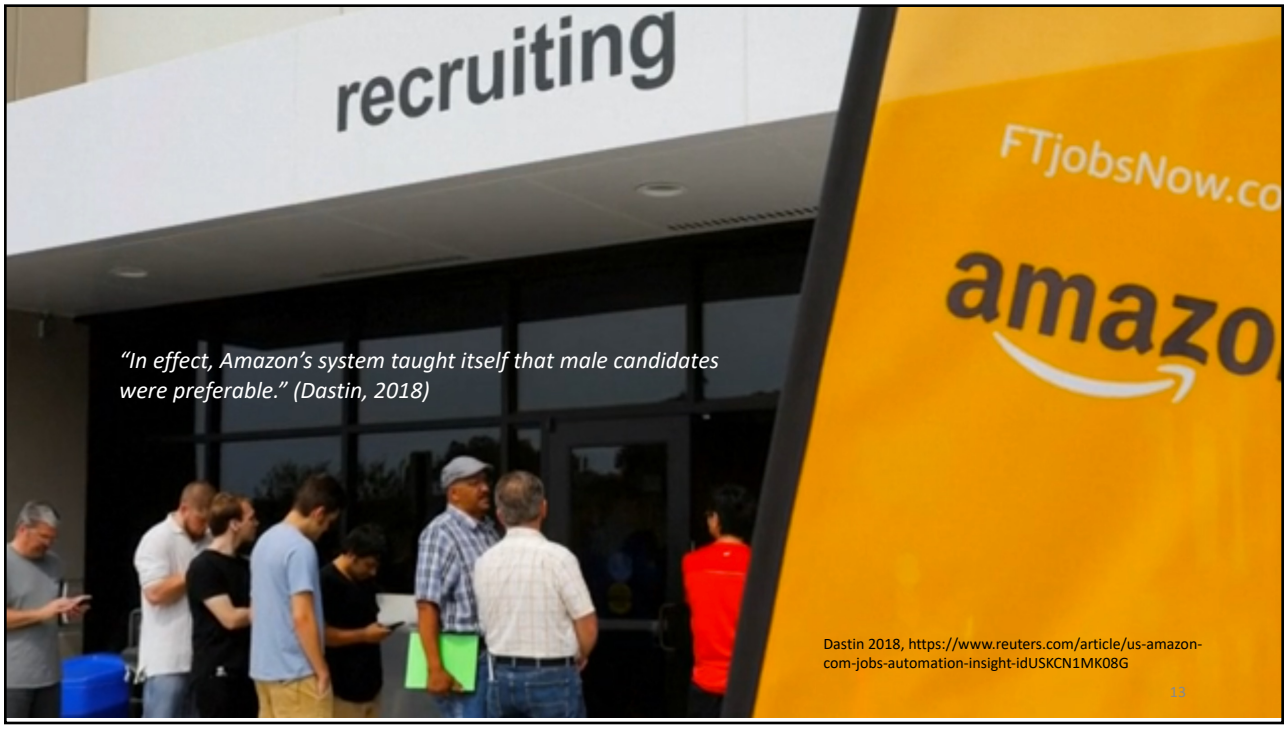
"Automated" decisions impact every aspect of our lives

- Precision agriculture
- Air combat
- Military training
- Education
- Finance
- Health care
- Customer service
- Advice on parole

- What ads are shown, discounts are given
- News feed
- Who to date
- Whether to grant a loan
- Admission to schools
- Who to hire and who to fire
- Work schedule
- ….



11



"In effect, Amazon's system taught itself that male candidates were preferable." (Dastin, 2018)

Dastin 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

13

13

Perspective | Published: 07 January 2019

The practical implementation of artificial intelligence technologies in medicine

Jianxing He ✉, Sally L. Baxter, Jie Xu, Jiming Xu, Xingtao Zhou & Kang Zhang ✉

Nature Medicine **25**, 30–36(2019) | Cite this article

26k Accesses | **219** Citations | **55** Altmetric | Metrics

Abstract

The development of artificial intelligence (AI)-based technologies in medicine is advancing rapidly, but real-world clinical implementation has not yet become a reality. Here we review some of the key practical issues surrounding the implementation of AI into existing clinical workflows, including data sharing and privacy, transparency of algorithms, data standardization, and interoperability across multiple platforms, and concern for patient safety. We summarize the current regulatory environment in the United States and highlight comparisons with other regions in the world, notably Europe and China.

14

14

"… there is also potential for abuse by AI technology developers. For example, clinical decision support systems could be programmed to increase profits for certain drugs, tests, or devices without clinical users being aware of this manipulation. For all medical devices, a tension exists between providing ethical medical care and generating profit. AI technologies will not be immune to that tension, and it should be openly acknowledged and addressed during implementation processes."

15

A problem for AI and Ethics

Consider the following rudimentary ethical questions about AI:

- What should the ultimate good of AI?
- What makes an AI innovation good vs. bad in a moral sense?
- How should AI function such that it promotes its ultimate good?

Problems:

- We're building artificial intelligence that is increasingly taking on the role of thought partner, information broker, medical expert, and social engineer
- There are no robust frameworks for evaluating the ethics of AI
- Industry won't figure this out for us (unless there is a business objective)

16

The Dual Use of A.I. Technologies

- For instance, GANs can be employed to do data-augmentation that can be very useful for medical applications, including cancer research.
- It can also be used in deepfakes

17

What is your opinion about…?

- Engineers are only responsible for their code, not how the code is used or the data quality
- Humans and computers are interchangeable; replacing humans with computers results in better outcomes
- Regulating the tech industry is too hard and won't be effective
- Our job in tech is just to optimize metrics and respond to customer demand

From Rachel Thomas FastAI

18

The Dual Use of A.I. Technologies

- Who should be responsible?
  - The person who uses the technology?
  - The researcher/developer?
  - Paper reviewers?
  - University?
  - Law-makers?
  - Society as a whole?

We need to be aware of real-world impact of our research and understand the relationship between ideas and consequences

19

You need to understand the ethical issues surrounding the data you obtain/use, the algorithms you employ, and its impact on people.

Well I'm just an engineer?

20

20

21

21

Normative, Legislation and initiatives

22

Normative, Legislation and initiatives

- The Montreal Declaration for a Responsible Development of Artificial Intelligence:
  - **Well-being**: The development and use of artificial-intelligence systems (AIS) must permit the growth of the well-being of all sentient beings.
  - **Respect for autonomy**: AIS must be developed and used with respect for people's autonomy, and with the goal of increasing people's control over their lives and their surroundings.
  - **Protection of privacy and intimacy**: Privacy and intimacy must be protected from intrusion by AIS and by data-acquisition and archiving systems.
  - **Solidarity**: The development of AIS must be compatible with maintaining the bonds of solidarity among people and generations.
  - **Democratic participation**: AIS must meet intelligibility, justifiability and accessibility criteria, and must be subjected to democratic scrutiny, debate and control.
  - **Equity**: The development and use of AIS must contribute to the creation of a just and equitable society.
  - **Diversity inclusion**: The development and use of AIS must be compatible with maintaining social and cultural diversity, and must not restrict the scope of lifestyle choices and personal experience.
  - **Prudence**: Every person involved in AIS development must exercise caution by anticipating, as far as possible, the potential adverse consequences of AIS use, and by taking appropriate measures to avoid them.
  - **Responsibility**: The development and use of AIS must not contribute to diminishing the responsibility of human beings when decisions must be made.
  - **Sustainable development**: The development and use of AIS must be carried out so as to ensure strong environmental sustainability of the planet.

23

Normative, Legislation and initiatives

- The Alan Turing Institute: [Understanding artificial intelligence ethics and safety]

24

Normative, Legislation and initiatives

- The Toronto Declaration Protecting the rights to equality and non-discrimination in machine-learning systems
- France's Digital Republic Act
- European Union General Data Protection Regulation
- The Ethical Principles offered in the *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems,* published by the European Commission's European Group on Ethics in Science and New Technologies (EGE)
- UK House of Lords Artificial Intelligence Committee's report (AIUK)
- Fairness, Accountability and Transparency in Machine Learning (researchers from industry and academia)
- OpenAI (non-profit)

Most of these initiatives are general rules, more than regulation.

25

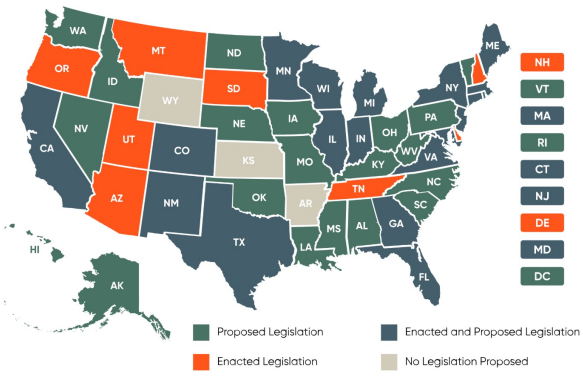
Regulation in the US

- Some states already have legislation to regulate very specific scenarios.
- The Department of Commerce, Federal Trade Commission, The White House, FDA, and other federal administrations have released statements, and action plans

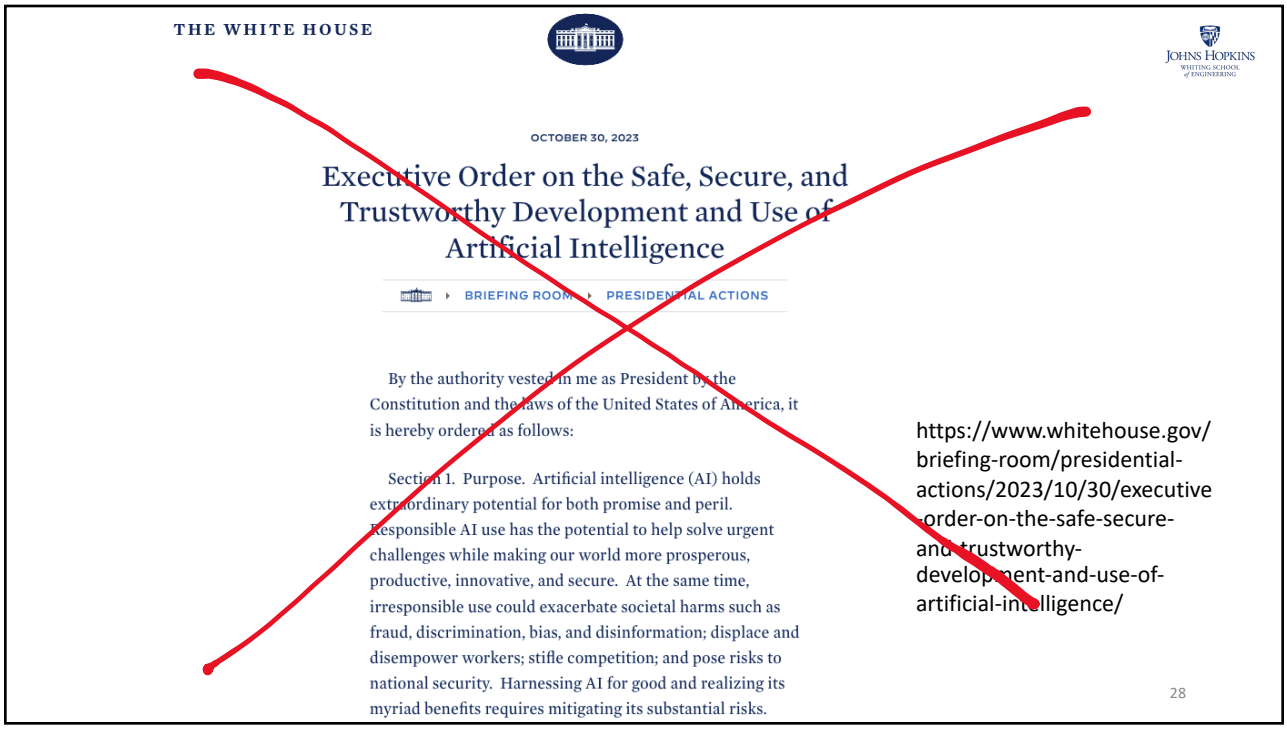
State	Bill Number	Bill Title	Bill Status	Bill Summary	Category
Colorado	H 1355	Producer Responsibility Program For Recycling	**Enacted**	Concerns the creation of the producer responsibility program for statewide recycling; appropriates funds. Requires a needs assessment to identify opportunities for the use of innovative new technologies, including artificial intelligence technologies, for the recycling and reuse of covered materials.	Private Sector Use
Colorado	S 113	Artificial Intelligence Facial Recognition	**Enacted**	Concerns the use of personal identifying data; creates a task force for the consideration of facial recognition services, restricting the use of facial recognition services by law enforcement agencies, temporarily prohibiting state and local government agencies and schools from executing new contracts for facial recognition services; appropriates funds.	Government Use; Studies
Connecticut	None				
Delaware	None				
District of Columbia	B24-558	Stop Discrimination by Algorithms Act of 2021	Failed – Adjourned	Prohibits users of algorithmic decision making from utilizing algorithmic eligibility determinations in a discriminatory manner; requires corresponding notices to individuals whose personal information is used; provides for appropriate means of civil enforcement.	Responsible Use

https://www.ncsl.org/technology-and-communication/legislation-related-to-artificial-intelligence

Regulation in the US



https://www.bclplaw.com/en-US/events-insights-news/us-state-by-state-artificial-intelligence-legislation-snapshot.html



28

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

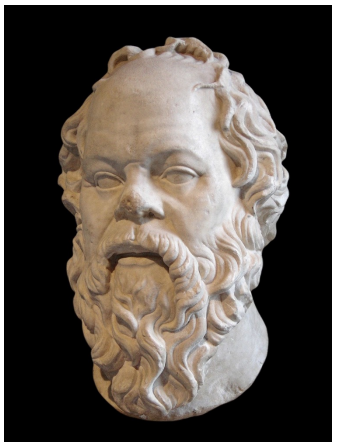
In the meantime…

- Do we always need to be told what to do?
- Will you kill someone if for a few minutes all laws were on hold?

29

Virtue ethics, an alternative to rules

- Virtue ethics offers an alternative to rule-based ethical systems (e.g., deontology, utilitarianism)
- Virtues are the qualities of people that promote human flourishing
- Virtue is attained by:
  - performing one's distinctive function well
  - cultivating intellectual and moral excellence
  - achieving proper inner states; i.e., those consistent with virtue



Socrates

30

Ethical Principles for AI

31

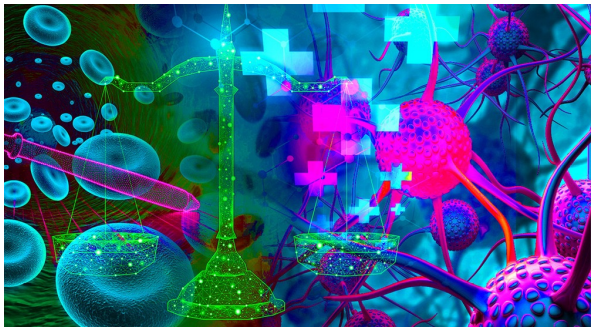
Ethical Principles in AI (a possible categorization)

- **Autonomy**
  - *"Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives."*
- **Beneficence**
  - People using your data should do it for your benefit
- **Non-maleficence**
  - Do no harm
  - Informed Consent
    - You should explicitly approve use of your data based on understanding
  - Control your data
- **Justice**
  - Promoting prosperity, preserving solidarity, avoiding unfairness
- **Explicability**
  - Enabling the Other four Principles through Intelligibility and Accountability

Floridi L, Cowls J (2019) A unified framework of five principles for AI in society. Harvard Data Science Review.

32

Beneficence



33

Beneficence

- AI should promote well-being, preserve dignity, and sustain the planet
- *"The development of AI should ultimately promote the well-being of all sentient creatures,"* (Montreal Declaration)
- We should *"ensure that AI technologies benefit and empower as many people as possible"* (AIUK)
- *"AI technology must be in line with ensuring the basic preconditions for life on our planet, continued prospering for mankind and the preservation of a good environment for future generations."* (EGE)

34

Non-Maleficence

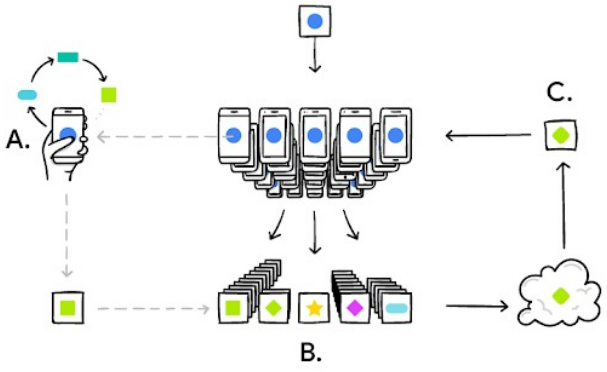
35

Non-Maleficence: Data Collection

- Data is constantly being collected about us
  - Cameras
  - Location reporting
  - Accelerometers
  - Social media
- Do I own data collected about me?
- What if I don't like what the data says about me?
- Can I control how the data is used?

Federated learning

- Instead of sharing data, the baseline models are adapted locally and then, shared.
- All the adapted models are used to prepare a global new baseline.



https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Institutional Review Boards (IRB)

- At Johns Hopkins University there are two main IRB: one at JHMI and one at Homewood
- The IRB is "responsible for protecting the rights and welfare of the human subjects of research conducted by faculty and staff at the Institutions".
- The IRB evaluates the ethical aspects of human subject research
- The board usually requires the investigators to inform the participants about the research in which they are involved → informed consent

38

Data and Informed Consent

- In human subjects research, there is a notion of *informed consent*
  - must *understand* what is being done (you have to assess if the participant understood what they're signing)
  - must *voluntarily consent* to the experiment
  - must have the right to withdraw consent at any time
- Not required in "ordinary conduct of business"
  - E.g. A/B testing
  - But this is a very thin line….

39

Informed Consent

- In some cases, informed consent is buried in the fine print
- Data is often collected first; the experiment comes later.
- How the data, once collected, is going to be used is difficult to control.
- IRBs try to ensure that the participants are correctly informed and accept the possible risks even if those are remote.

40

Privacy

- Many rules governing use of collected information
  - **HIPAA:** Health Insurance Portability and Accountability Act
  - **FERPA:** Family Educational Rights and Privacy Act
  - **GDPR** General Data Protection Regulation (Europe)

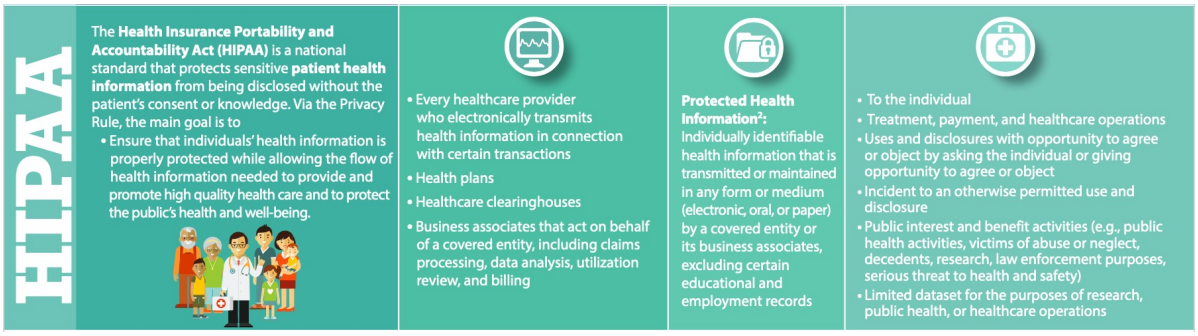
41

Privacy: HIPAA

- US federal law that required the creation of national standards to protect sensitive patient health information from being disclosed without the patient's consent or knowledge.
- The Privacy Rule standards address the use and disclosure of individuals' health information (protected health information - PHI) by entities subject to the Privacy Rule:
  - Healthcare providers
  - Health plans
  - Healthcare clearinghouses
  - Business associates

42

Privacy: HIPAA



Obtained from: https://www.cdc.gov/phlp/publications/topic/healthinformationprivacy.htm

43

How can we measure maleficence?

- How do we evaluate the harm that some algorithms can do to society?
- For instance, can we measure if Facebook feed algorithms are “maleficent”?

Explicability

Explicability

- Opening up the black-box would not suffice to disclose algorithms' modus operandi
- Transparency and reproducibility: make the code available
- The algorithms used in data science are complicated
  - When things "go wrong", we need to understand why
- The data often cannot be shared
- Some authors propose algorithmic auditing processes (Ragi et al 2020)

Raji ID et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference onFairness, Accountability, and Transparency pp 33–44 (Association forComputing Machinery, 2020).

46

Justice

47

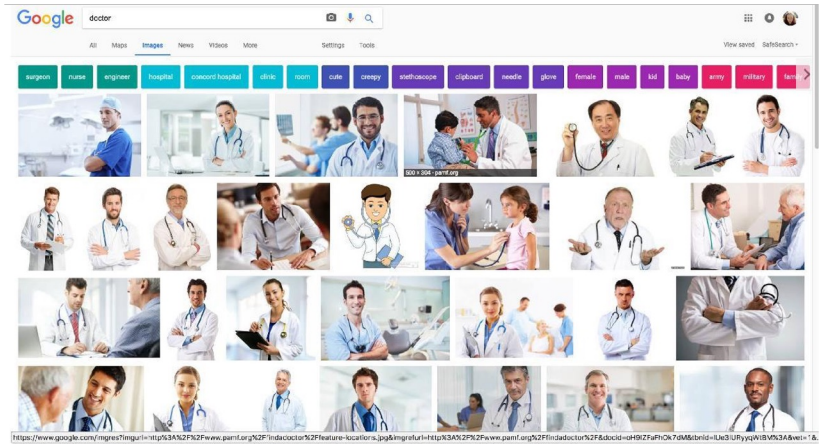
Algorithms are not neutral

- Algorithms encode our biases.
  - Training data set isn't representative
  - Past population is not representative of the future population

48

Image Search

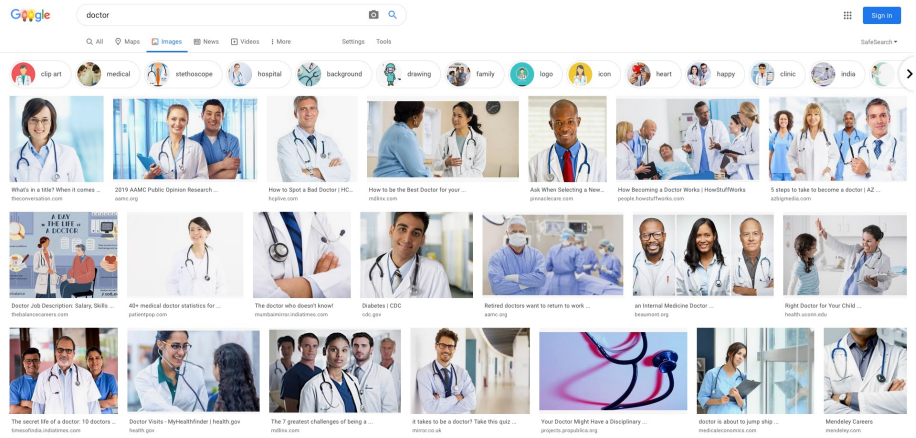
- June 2017: image search query "Doctor"



49

Image Search

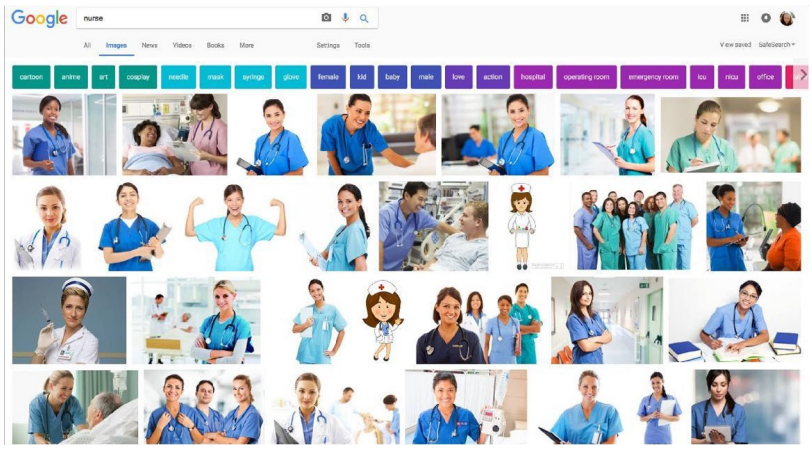
- December 2020: image search query “Doctor”



50

Image Search

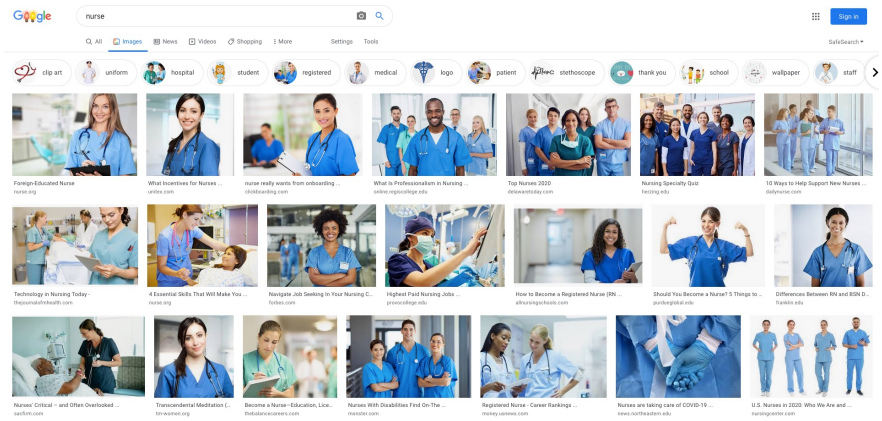
- June 2017: image search query “Nurse”



51

Image Search

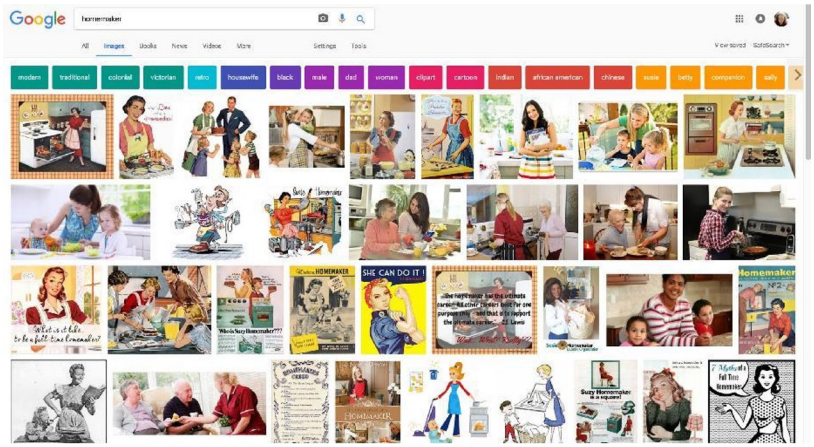
- December 2020: image search query "Nurse"



52

Image Search

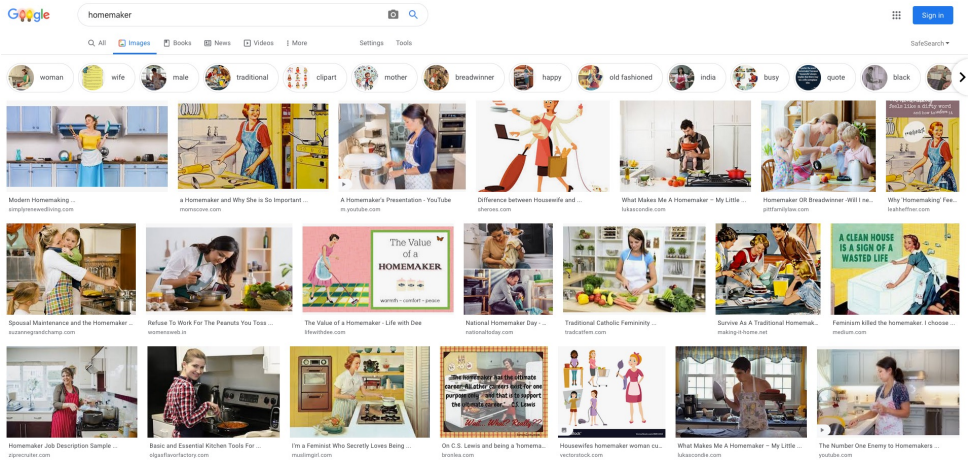
- June 2017: image search query "Homemaker"



53

Image Search

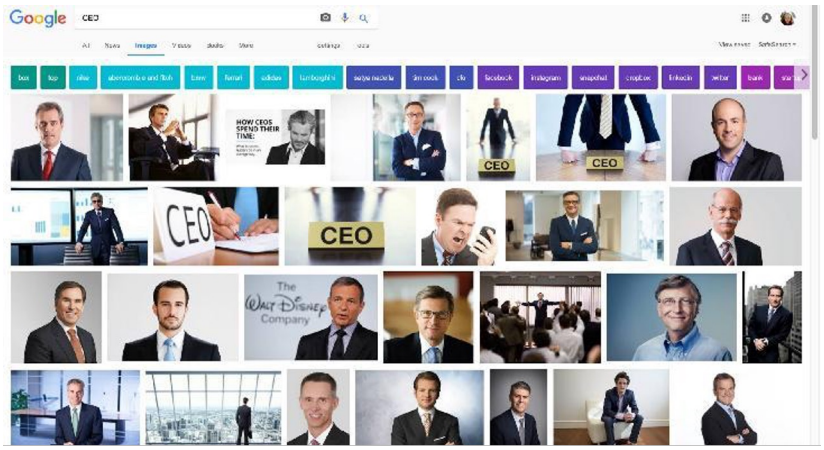
- December 2020: image search query "Homemaker"



54

Image Search

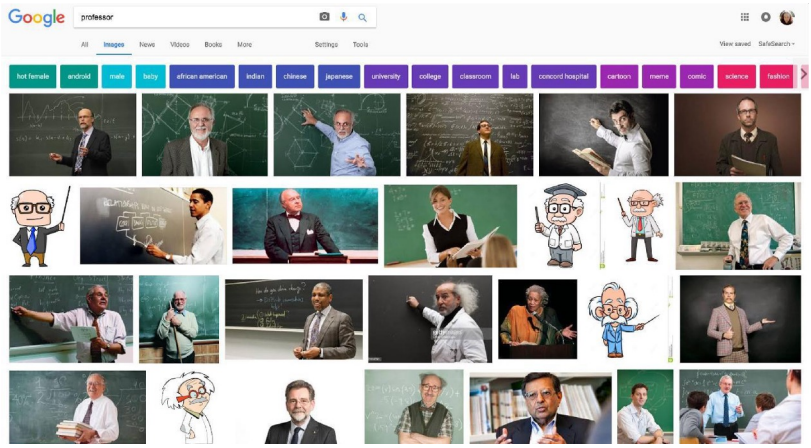
- June 2017: image search query "CEO"



55

Image Search

- June 2017: image search query "Professor"



56

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Fairness

- Fairness has been studied in social choice theory, game theory, economics and law.
- Currently trendy in theoretical computer science
  - **Discrimination of an individual:** An individual from the target group gets treated differently from an otherwise identical individual not from the target group.
  - **Discrimination in aggregate outcome**: the percentage success of the target group compared to that of the general population.
- Zip code or language used to assess the capacity of an individual to pay back a loan or handle a job → discrimination (O Neill, 2016)

Dwork, Hardt, Pitassi, Reingold and Zemel, "Fairness through Awareness" Proc. 3rd Innovations in Theoretical Computer Science, 2012.
O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy, 1st edn. Crown, New York

57

In conclusion…

- Codes of conduct for research are fairly well understood
  - Get IRB approval
  - obtain informed consent
  - protect the privacy of subjects
  - maintain the confidentiality of data collected
  - minimize harm
- Fairness is more subtle
  - What is fair treatment of a group: equal accuracy? FP rate?