# Ethics

Instructor: Laureano Moro-Velazquez

Johns Hopkins University

December 3rd 2020

Slides: João Sedoc & Laureano Moro-Velazquez
Some content from E. Heizer & A. Roth & Z. Ives and L. Ungar & Yulia Tsvetkov, Olya Kudina and Alan Black

## 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


## Speech Assistants

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

6

# 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

# Some initiatives to this challenge

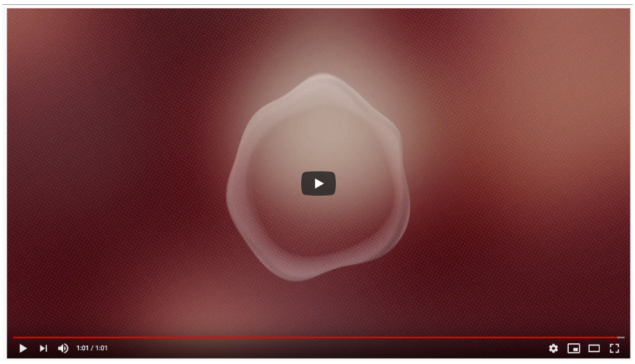
- Zara, the supergirl:
  - Human-like response of sharing emotions
  - Empathy module
  - Handles abusive language
  - Apologizes

Fung, Pascale, et al. "Zara the supergirl: An empathetic personality recognition system." *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. 2016.

8

## Gender-neutral voice assistant

- Q, the genderless voice



https://www.genderlessvoice.com/

9

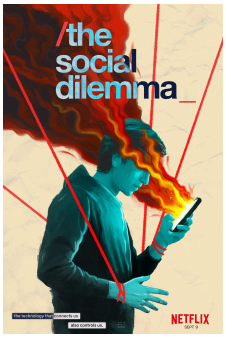
# Algorithms that take decisions

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

DECISION MAKING

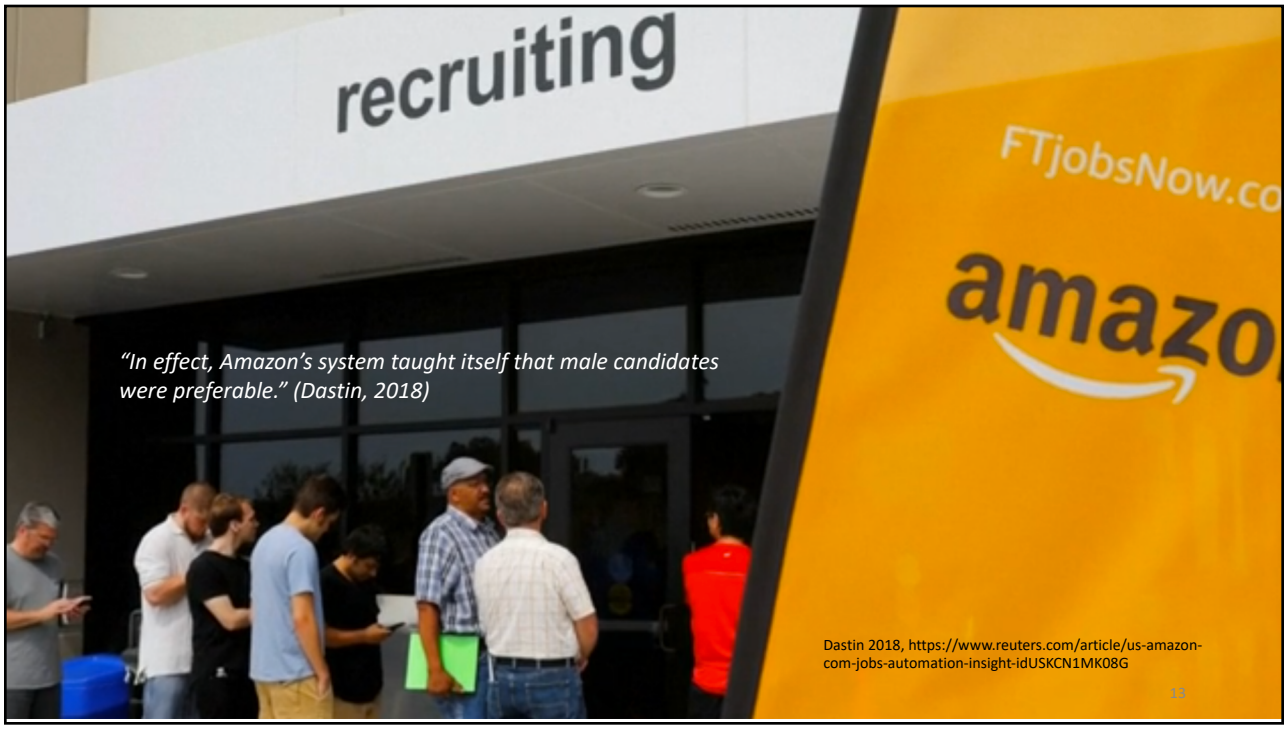
## Want Less-Biased Decisions? Use Algorithms.

by Alex P. Miller

July 26, 2018

Summary Save Share 8 Comment Print $8.95 Buy Copies

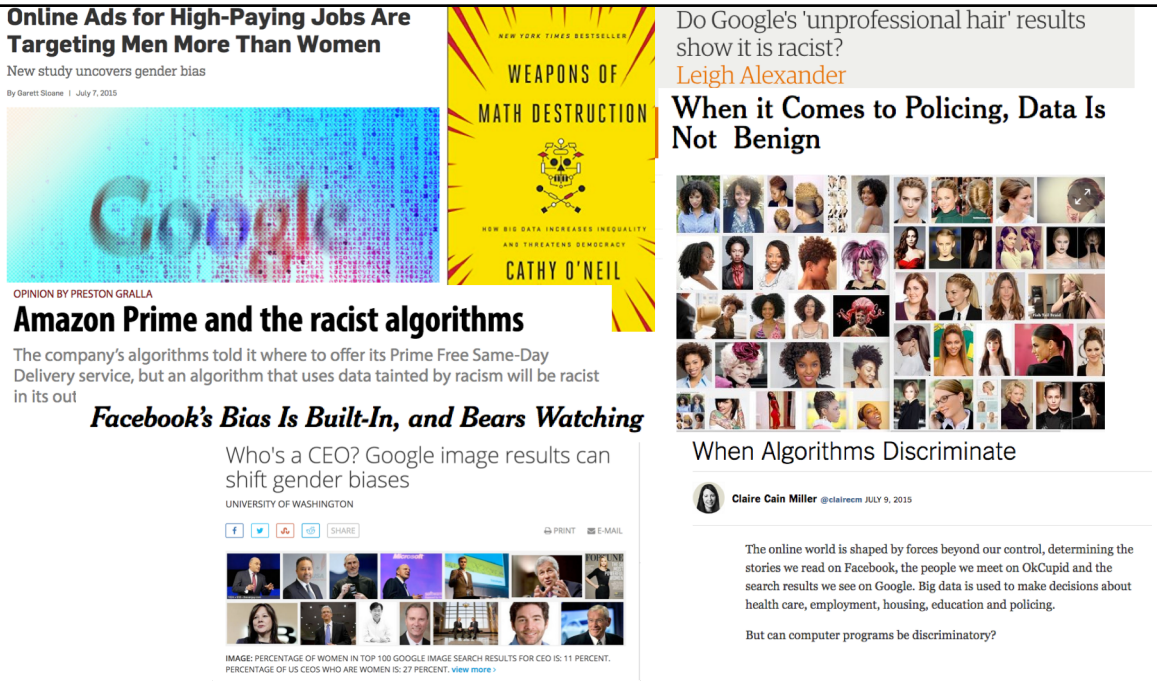
12




*"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




14

Slide content courtesy of Hanna Wallach

# A problem for AI and Ethics

Consider the following rudimentary ethical questions about AI:

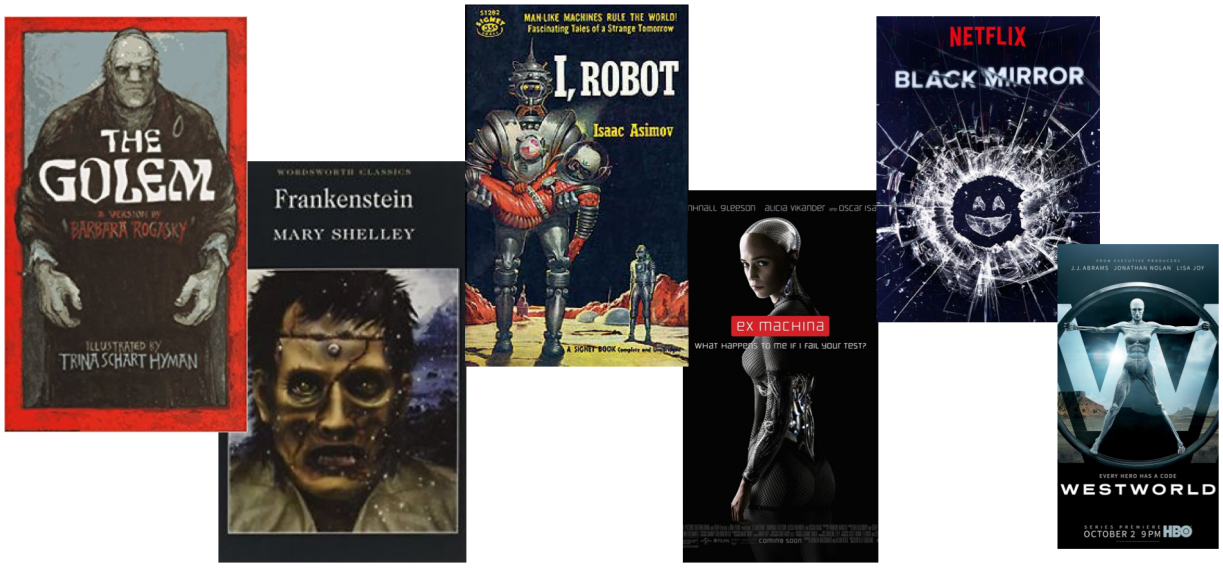
- What should the ultimate good of AI?
- What makes an AI innovation good vs. bad in a moral sense?
- How should AI function such that it promotes its ultimate good?

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)

15

# The Long History of Ethics and AI



# 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

# 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**

**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?

19

20

## Misconceptions about AI that can lead to harm

- 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

21

## Normative, Legislation and initiatives

22

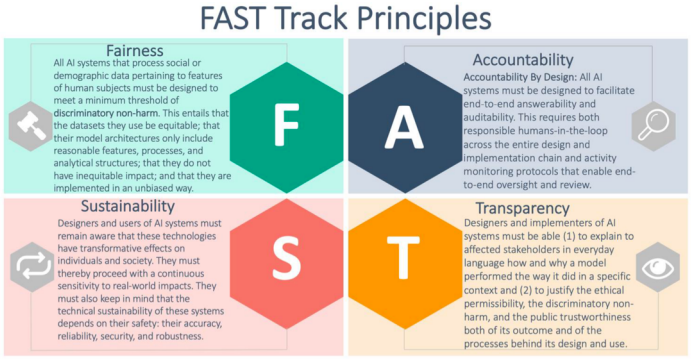
# Normative, Legislation and initiatives

- 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

# What we'll discuss…

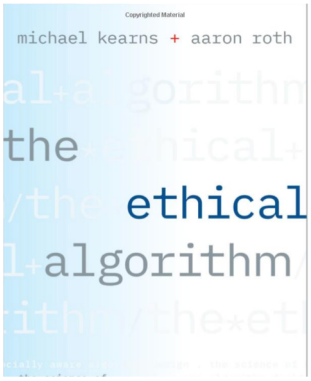
- **What are ethics?**
- **Ethics and AI:**
  - **Ethics surrounding data**
  - Privacy and informed consent
  - Ownership and intellectual property
  - **Ethics surrounding algorithms**
  - Biased algorithms
  - Bad results from good data
    - reproducibility
  - Fairness

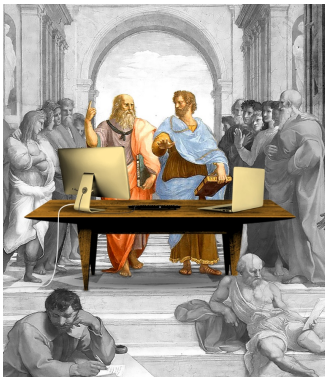
26

# Virtue ethics for humans

- 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
- Virtues may be cultivated based on learning and emulation


27



Algorithmic Virtue ???

Ethical Algorithms

28

## Need to *embed* social values in algorithms

- Requires being precise about the definitions, developing their consequences.
  - Privacy
  - Fairness
  - Accountability
  - Interpretability

30

# Ethical Principles for AI

31

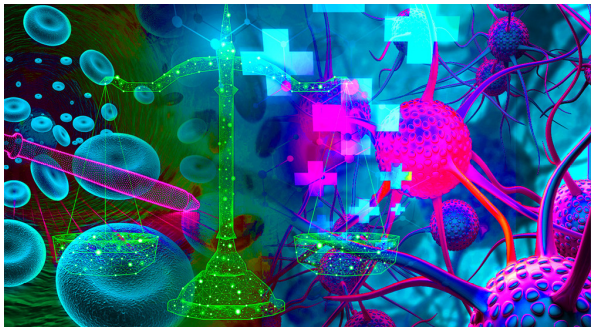
# Ethical Principles in AI (a possible categorization)

- **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

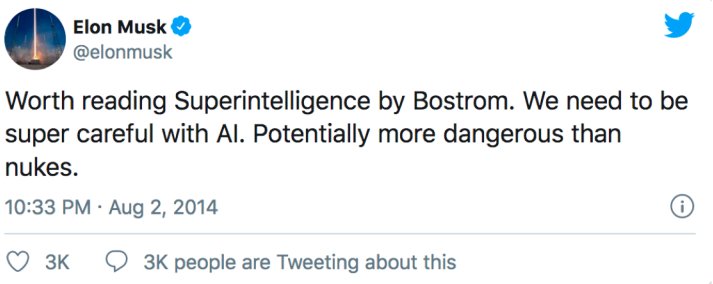


33

## 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

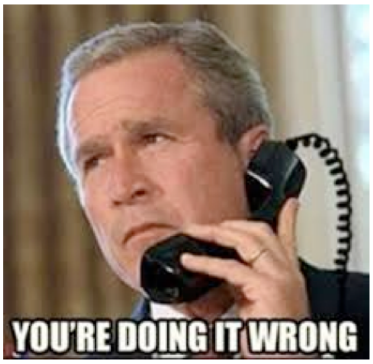



## Non-Maleficence

35

# Non-Maleficence: Data Collection

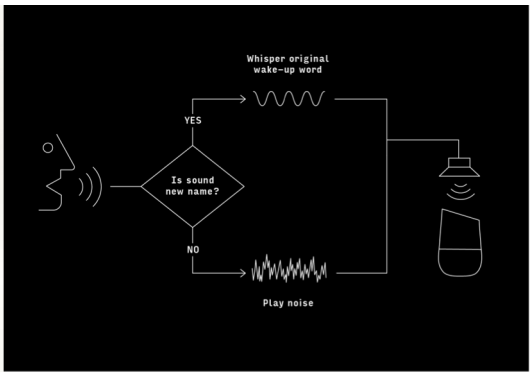
- Data is constantly being collected about us
  - Cameras
  - Location reporting
  - Accelerometers
  - Social media
- Do I own data collected about me?
- What if I don't like what the data says about me?
- Can I control how the data is used?




35

# One initiative: Project Alias

- Some engineers are working on giving the user more control




http://bjoernkarmann.dk/project_alias

37

# 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.

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)
- However, "information leakage" can lead to unexpected disclosures
  - e.g. smart water meters
- "Privacy by trust" versus "privacy by design"

41

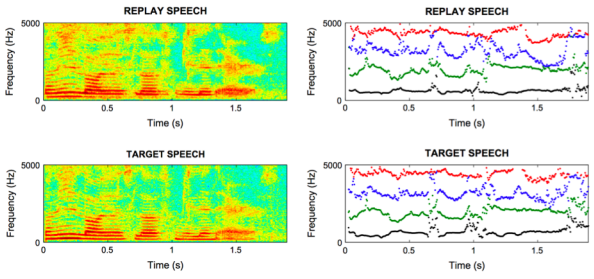
# How can we measure maleficence?

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

42

# Can we always ensure non-maleficence while looking for beneficence?

- Discussion: how ethical are voice conversion challenges?
  - http://www.vc-challenge.org/




- These challenges can help better understanding spoofing but, can these push boundaries and disseminate knowledge about how to do spoofing?

43

# Explicability



## Explicability

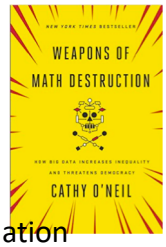
- Opening up the black-box would not suffice to disclose algorithms' modus operandi
- Transparency and reproducibility: make the code available
- The algorithms used in data science are complicated
  - When things "go wrong", we need to understand why
- The data often cannot be shared
- Some authors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

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

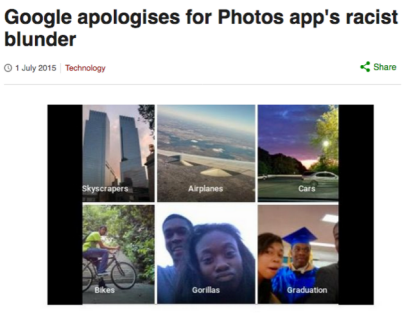
# Justice

48

## Algorithms are not neutral



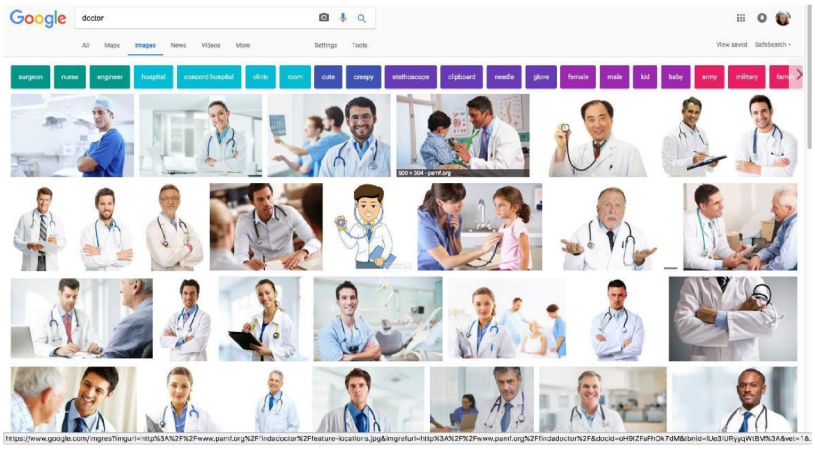

- Algorithms encode our biases.
  - Training data set isn't representative
  - Past population is not representative of the future population




49

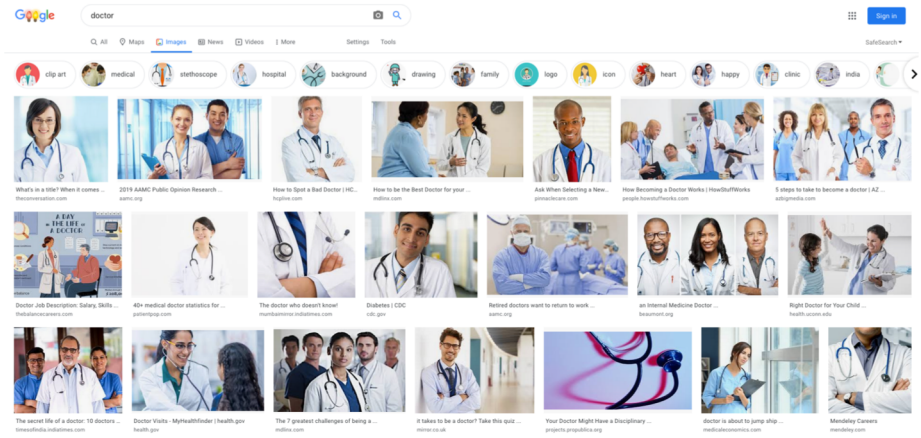
# Image Search

- June 2017: image search query “Doctor”



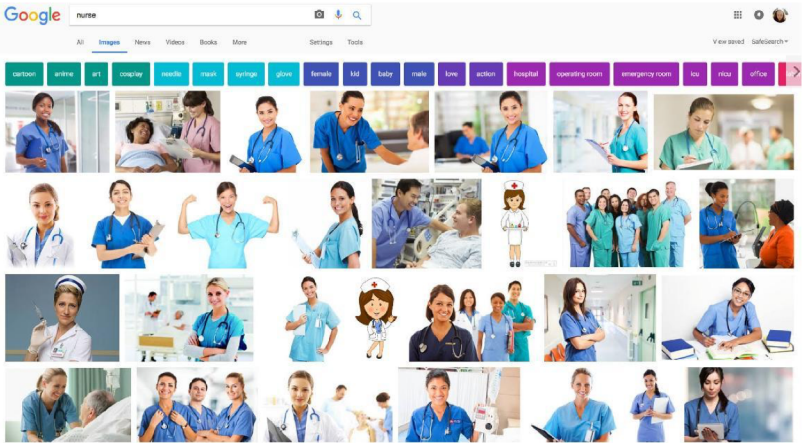

# Image Search

- December 2020: image search query “Doctor”



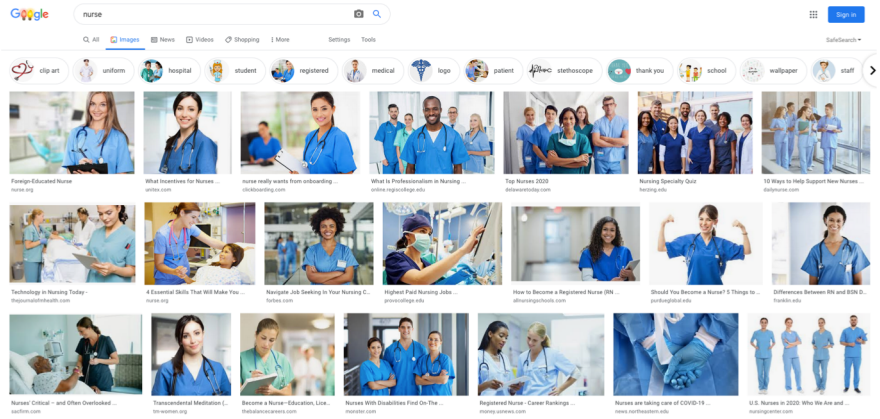
# Image Search

- June 2017: image search query "Nurse"



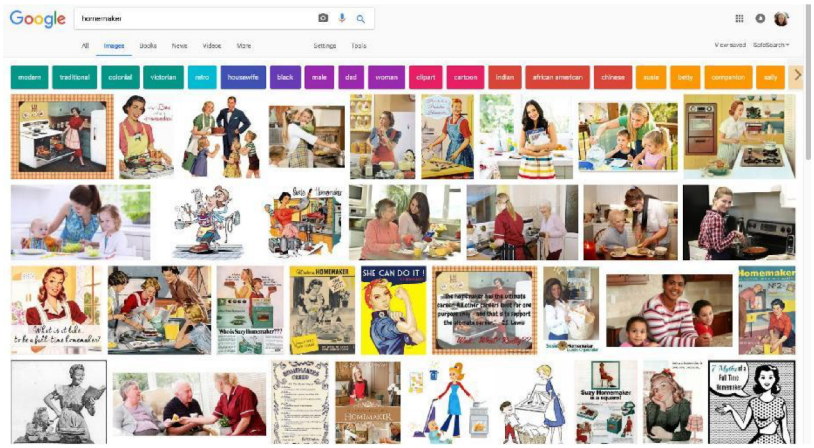

# Image Search

- December 2020: image search query "Nurse"



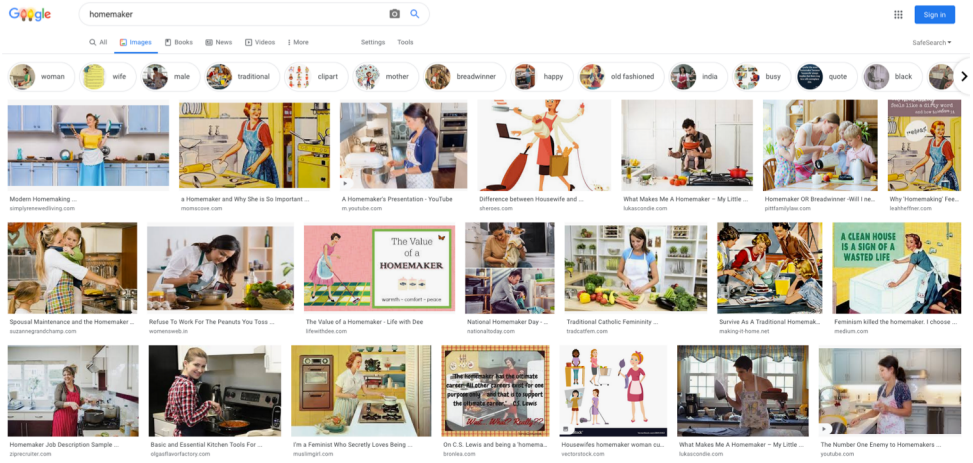
# Image Search

- June 2017: image search query "Homemaker"



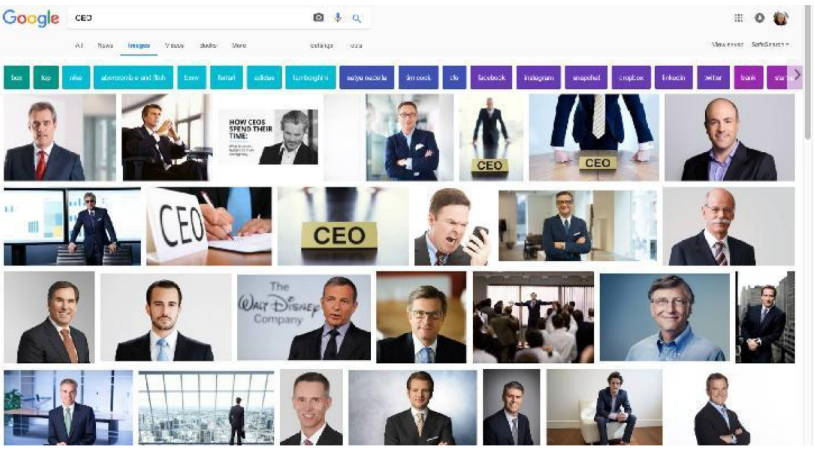

# Image Search

- December 2020: image search query "Homemaker"



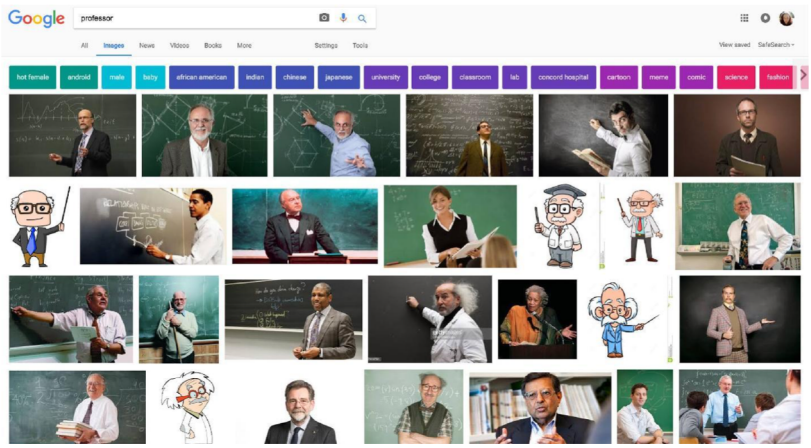
# Image Search

- June 2017: image search query "CEO"




# Image Search

- June 2017: image search query "Professor"

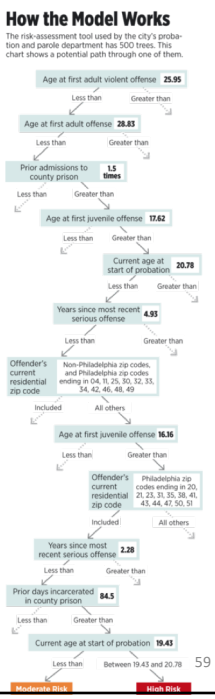


# Fairness

- 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

58


59

# Why is computer advice on parole controversial?

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

60

# AI in recidivism prediction

- COMPAS algorithm: evaluates the risk of violent recidivism as a score.

→Some studies suggest that it has a potential racial bias (Anwing & Larson 2016)

→According to the authors of the report, the recidivism-risk was systematically overestimated for black people:

→The company (Northpointe) refused to disclose the full details of it's proprietary code.

Angwin J, Larson J (2016) Machine bias.https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

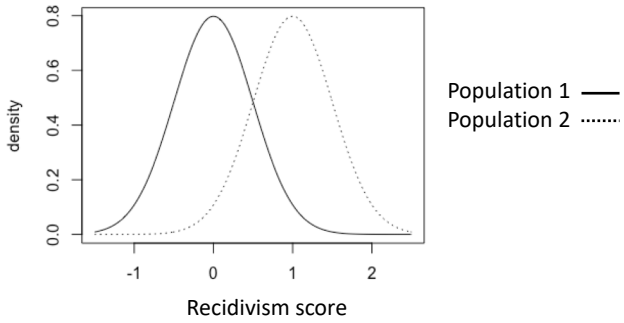
61

# An example: Beneficence against Justice?

Imagine that your AI company has these distributions before training a new model (to be paid by an administration):

- Including the population's information as input to the system reduces the system's error significantly. Consider that you have to protect population from dangerous criminals.
- If the only difference between these two distributions is race, would you use race as the input of your recidivism score predictor?
- What if the difference is not race but the place where you live or your monthly income?

**- Correlation does not mean causation**
**- Being part of an ethnic/social group shouldn't be a burden**




62

# 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?

63

## Some upcoming seminars and events

- CLSP seminar: "Five Sources of Biases and Ethical Issues in NLP and What to Do about Them", Dirk Hovy from Bocconi University, Friday, December 4 at 12:00pm.
- IAA Speakers series: "The Ethical Algorithm" featuring speakers Michael Kearns and Aaron Roth, professors in the Computer and Information Science department at the University of Pennsylvania, presenting virtually on Tuesday, December 15th at 11 a.m.

64

# Some insights about privacy and intellectual property

Extra content

65

# Intellectual Property

- Artistic expression can be **copyrighted**: exclusive legal right to print, publish, perform, film or record and authorize others to do the same.
- **Derivative** work can be created with permission.
- There's also a notion of **citation**, in which we give credit to the owner.
- What about data?
  - Wikipedia, Yelp, Rotten Tomatoes, TripAdvisor

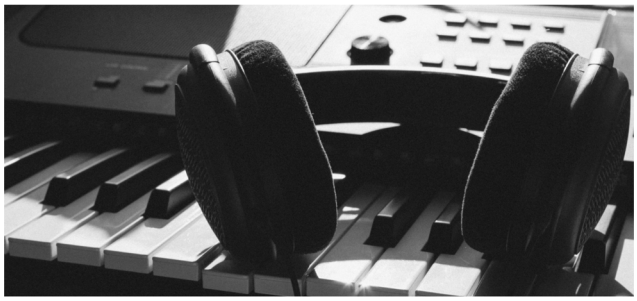
66

# Intellectual Property - Copyrights

Neural Networks for Music Generation

Can we reproduce artists' creativity through AI?

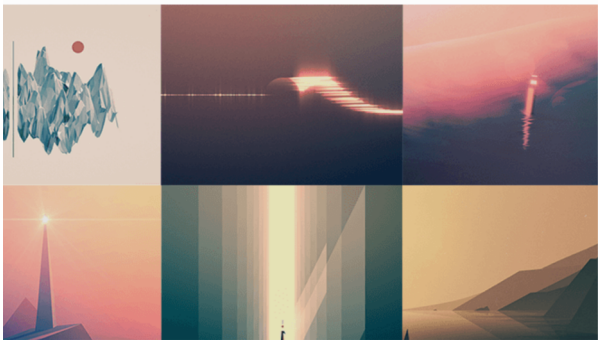
Andy Spezzatti Follow
Jun 24 · 13 min read ★



An exciting application of the recent advance in AI is Artificial Music Generation. Can we reproduce artists' creativity through AI? Can a Deep Learning model be an inspiration or a productivity tool for musicians? Those questions bring us to the definition of creativity and the

**Creative Tools to Generate AI Art**

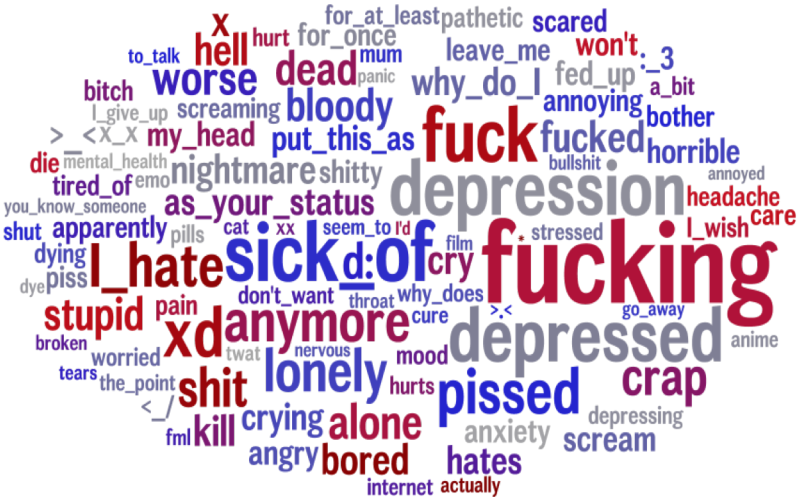
Wondering how to make AI art? Scroll down for our list of tools to generate AI art.



67

What can we learn from your social media posts?

Neurotic

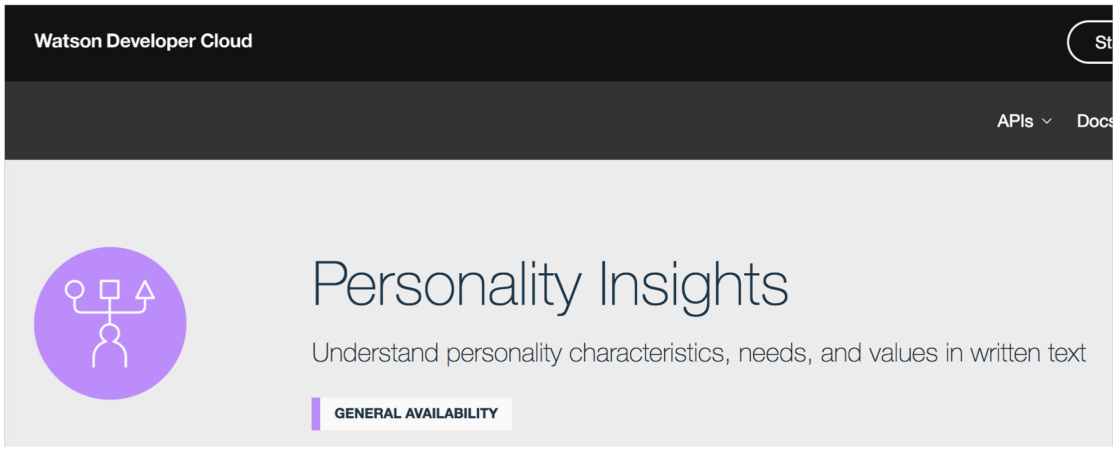





What can we learn from your social media posts?

Well adjusted

Targeted marketing from IBM




70

# Admiral to price car insurance based on Facebook posts

Insurer's algorithm analyses social media usage to identify safe drivers in unprecedented use of customer data

theguardian

71

## Targeted marketing: Cambridge Analytica


72

## Privacy

73

# OKCupid Data Publicly Released

WIRED, Michael Zimmer 5/14/16

- On May 8, a group of Danish researchers publicly released a dataset of nearly 70,000 users of the online dating site OkCupid, including usernames, age, gender, location, what kind of relationship (or sex) they're interested in, personality traits, and answers to thousands of profiling questions used by the site.
- When asked whether the researchers attempted to anonymize the dataset the response was: "… all the data found in the dataset are or were already publicly available, so releasing this dataset merely presents it in a more useful form."

74

# Was the OKCupid data public?

- The methodology used to obtain the data was not fully explained, but involved a scraping bot.
  - Likely from an OkCupid profile researchers created.
- Since OkCupid users have the option to restrict the visibility of their profiles to logged-in users only, it is likely the researchers collected—and subsequently released—profiles that were intended to *not* be publicly viewable.

75

# Anonymity?


76

# Correlating data

- **Netflix Prize Competition:** released a de-identified data set with user ID, date, movie name, and the rating given by the user for that movie.
  - Researchers were able to link users with IMDb's system where the users were identified, and talked about (some of) the movies they watched.
- **Problem: "Sparsity" of data**
  - In Netflix data, no two profiles are more than 50% similar.
  - If a Netflix profile is more than 50% similar to a profile in IMDB, then there is a high probability that the two profiles are of the same person

A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets …," Proc. 29th IEEE Symp. Security and Privacy, 2008.

77

# Differential Privacy

- When do you feel safe releasing personal information, e.g. completing a survey about your tastes in movies?
  - My answers have no impact on the privatized released result?
  - With high probability, an attacker looking at the privatized released result cannot learn any new information about me?
  - **These are not achievable.**
- **Differential privacy** aims to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying its records.
  - **The privatized released result is nearly the same whether or not I submit my information.**

Dwork and Roth, "Algorithmic Foundations of Differential Privacy," Foundations and Trends in Theoretical Computer Science (2014).

78

# Differential Privacy – Speech applications

- Can we extract any information from the speech of a person?
  - From the content of the speech.
  - If they are participanting in a certaing study (related to certain addictions or medical issues, for instance.)
- How can we re-identify someone? → Most of the people cannot do it but big tech-companies can.

79