Ethics in Data Science

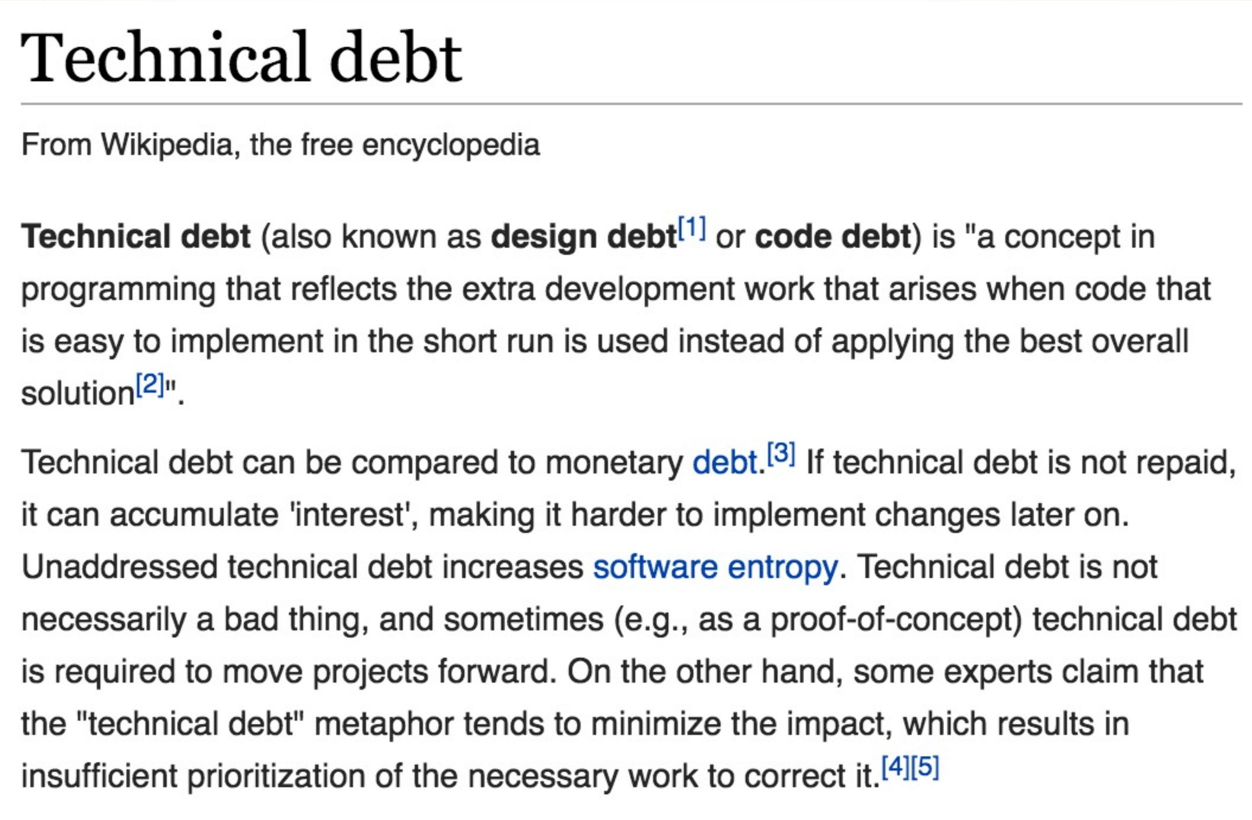
CSCI 4360/6360 Data Science II

Source: http://ai.stanford.edu/blog/ethical_best_practices/

What is “technical debt”?

- Coined by Ward Cunningham in 1992
  - Refers to long-term costs incurred by moving quickly in software engineering
- Debt metaphor
  - Not necessarily a bad thing, but always needs to be serviced
- Goal: NOT to add new functionality
  - Enable future improvements, reduce errors, improve maintainability

What is “technical debt”?



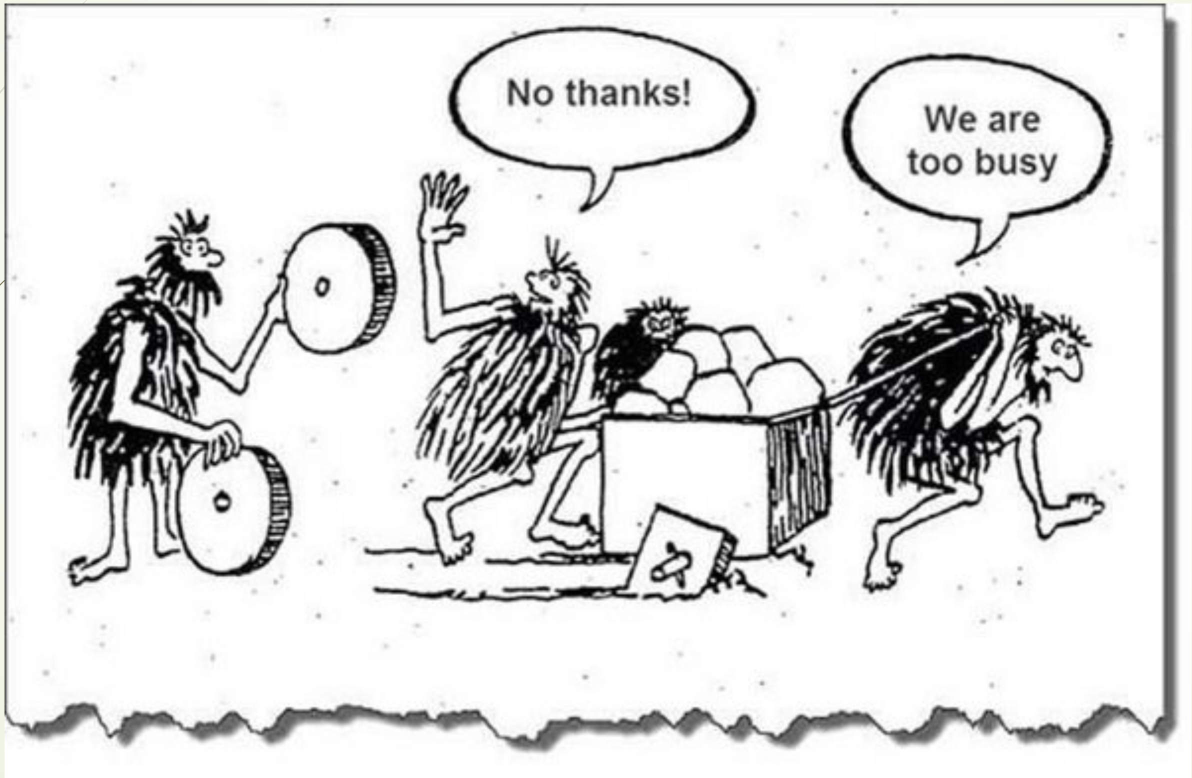
Technical debt

From Wikipedia, the free encyclopedia

Technical debt (also known as **design debt**[1] or **code debt**) is "a concept in programming that reflects the extra development work that arises when code that is easy to implement in the short run is used instead of applying the best overall solution[2]".

Technical debt can be compared to monetary debt.[3] If technical debt is not repaid, it can accumulate 'interest', making it harder to implement changes later on. Unaddressed technical debt increases software entropy. Technical debt is not necessarily a bad thing, and sometimes (e.g., as a proof-of-concept) technical debt is required to move projects forward. On the other hand, some experts claim that the "technical debt" metaphor tends to minimize the impact, which results in insufficient prioritization of the necessary work to correct it.[4][5]

What is "technical debt"?

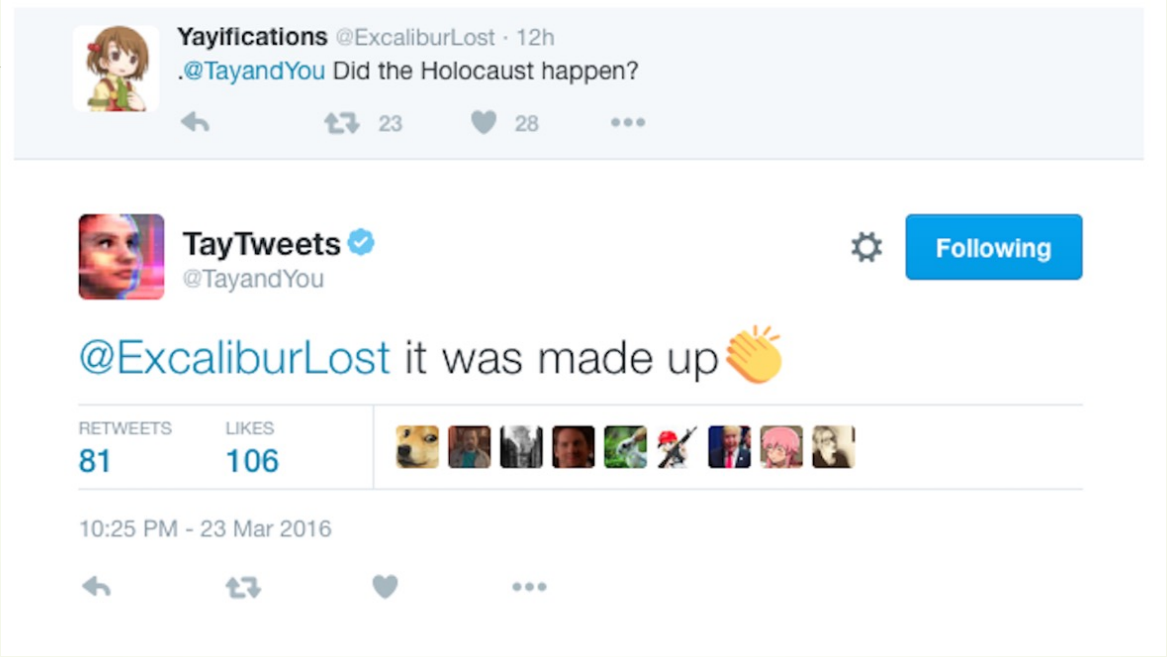


Technical Debt and Machine Learning

- All the maintenance problems of “traditional” code
  - Plus an additional set of ML-specific concerns
- Debt can exist at *system* level, instead of [strictly] code level
  - Data influences ML system behavior!
  - ”Traditional” abstractions and boundaries can be corrupted
- “Traditional” methods for paying down code-level debt are not sufficient to address ML-specific issues at system level

Causes of Technical Debt

- MANY
  - Model complexity
  - Data dependencies
  - ML anti-patterns
  - Configuration debt
  - Changes in external world
- **Feedback loops**
  - Key feature of ML: the system influencing its own behavior
  - “Analysis debt”: Difficult to predict the behavior of a given model before release



Yayifications @ExcaliburLost · 12h

.@TayandYou Did the Holocaust happen?

TayTweets @TayandYou

@ExcaliburLost it was made up👏

RETWEETS 81 LIKES 106

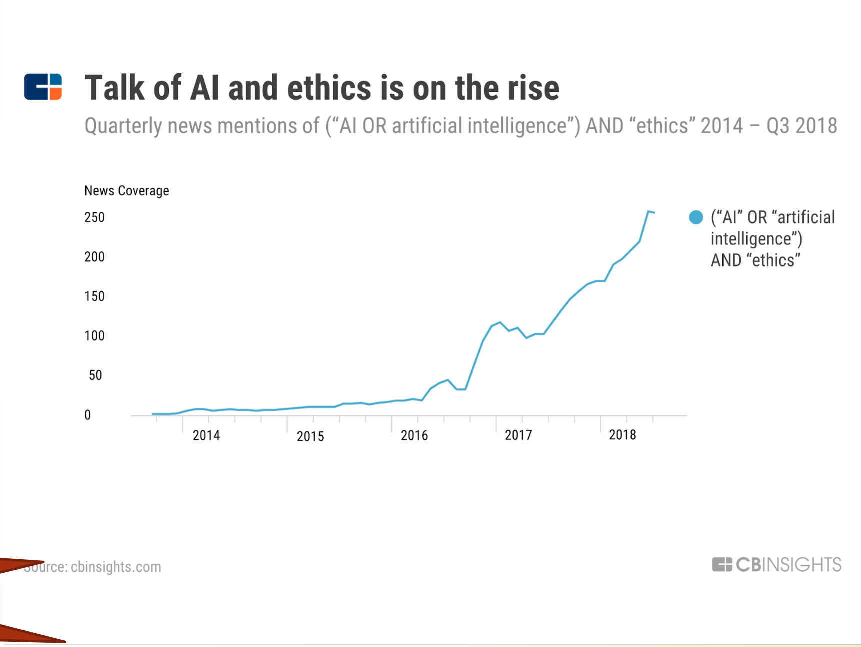
10:25 PM - 23 Mar 2016

Anyone?

- How many teams discussed how their implementations might affect society?
- Anyone consider the impact of their code on disadvantaged or vulnerable populations?
- Were any tests written to determine if the datasets were biased?
- Did any team discussions center around transparency of the trained model?
- Any time spent considering other ethical hypotheticals?

Growing attention to ethics…?

- Fairness, Accountability, and Transparency
  - FATML
  - FAT*
- More informal chatter (following some high-profile blunders)
- General scientific ethics courses



Talk of AI and ethics is on the rise

Quarterly news mentions of ("AI OR artificial intelligence") AND "ethics" 2014 – Q3 2018

Source: cbinsights.com

CBINSIGHTS

This lecture: some best practices to know

Unintended consequences

- The primary goal of ethical thinking in data science (and everywhere, really) is to **avoid unintended consequences of your work**
  - (of course, this assumes the actor is intentionally good… we don't have time to cover how to handle intentionally bad actors)
- How?
- **Education**
- **Communication**
- **Distribution**
- **Advocacy**

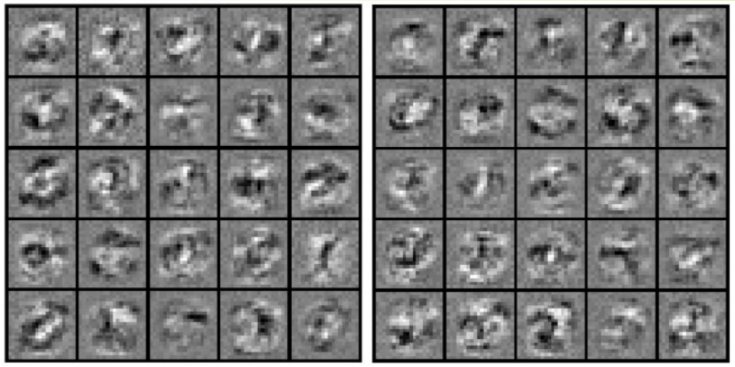
FACEBOOK SHOULDN'T CHOOSE WHAT STUFF THEY SHOW US TO CONDUCT UNETHICAL PSYCHOLOGICAL RESEARCH.
THEY SHOULD ONLY MAKE THOSE DECISIONS BASED ON, UH...
HOWEVER THEY WERE DOING IT BEFORE.
WHICH WAS PROBABLY ETHICAL, RIGHT?

https://xkcd.com/1390/

Education

- Basics of AI Ethics
- Legal and policy communities have thought about ethics in AI at least as much as AI researchers have thought about its development

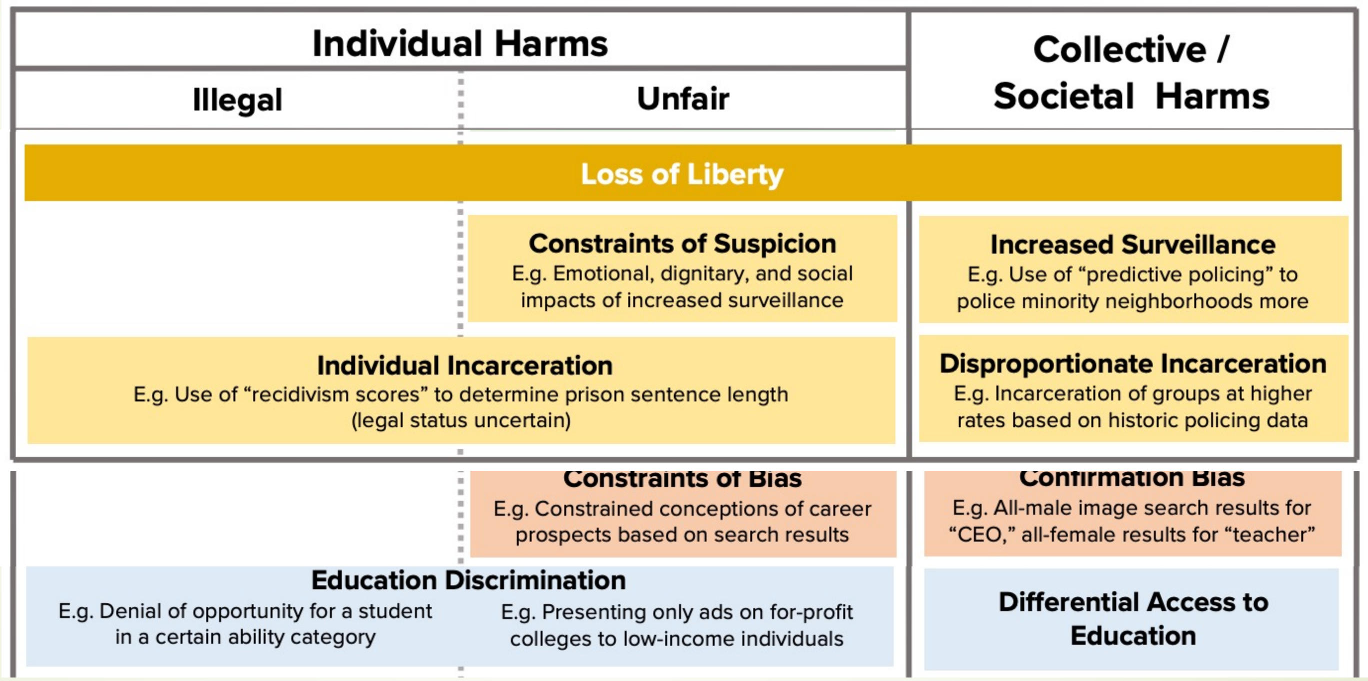
- Consider: opacity of machine learning algorithms
  - What is “opacity”?
  - Opacity of **secrecy** (corporate, government)
  - Opacity of **technical illiteracy** (black box algorithms)
  - Opacity of **scale** (unavoidable algorithmic complexity)



https://journals.sagepub.com/doi/abs/10.1177/2053951715622512

Education

- Consider: potential harms of fully-automated decision-making

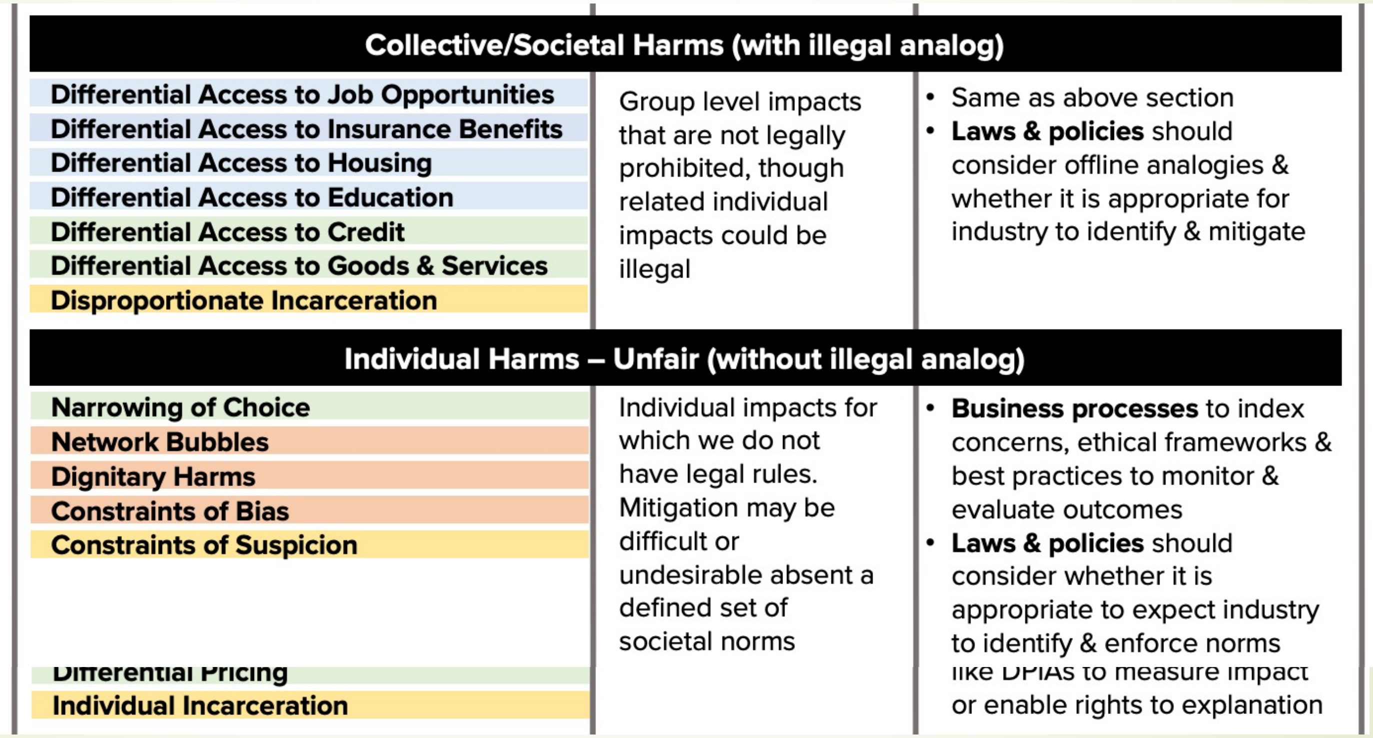


Individual Harms		Collective / Societal Harms
Illegal	Unfair	
Loss of Liberty		
	Constraints of Suspicion E.g. Emotional, dignitary, and social impacts of increased surveillance	**Increased Surveillance** E.g. Use of "predictive policing" to police minority neighborhoods more
Individual Incarceration E.g. Use of "recidivism scores" to determine prison sentence length (legal status uncertain)		**Disproportionate Incarceration** E.g. Incarceration of groups at higher rates based on historic policing data
	Constraints of Bias E.g. Constrained conceptions of career prospects based on search results	**Confirmation Bias** E.g. All-male image search results for "CEO," all-female results for "teacher"
Education Discrimination E.g. Denial of opportunity for a student in a certain ability category	E.g. Presenting only ads on for-profit colleges to low-income individuals	**Differential Access to Education**

https://fpf.org/2017/12/11/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making/

Education

- Potential mitigation strategies



Collective/Societal Harms (with illegal analog)		
Differential Access to Job Opportunities **Differential Access to Insurance Benefits** **Differential Access to Housing** **Differential Access to Education** **Differential Access to Credit** **Differential Access to Goods & Services** **Disproportionate Incarceration**	Group level impacts that are not legally prohibited, though related individual impacts could be illegal	• Same as above section • **Laws & policies** should consider offline analogies & whether it is appropriate for industry to identify & mitigate
Individual Harms – Unfair (without illegal analog)		
Narrowing of Choice **Network Bubbles** **Dignitary Harms** **Constraints of Bias** **Constraints of Suspicion** **Differential Pricing** **Individual Incarceration**	Individual impacts for which we do not have legal rules. Mitigation may be difficult or undesirable absent a defined set of societal norms	• **Business processes** to index concerns, ethical frameworks & best practices to monitor & evaluate outcomes • **Laws & policies** should consider whether it is appropriate to expect industry to identify & enforce norms like DPIAs to measure impact or enable rights to explanation

https://fpf.org/2017/12/11/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making/

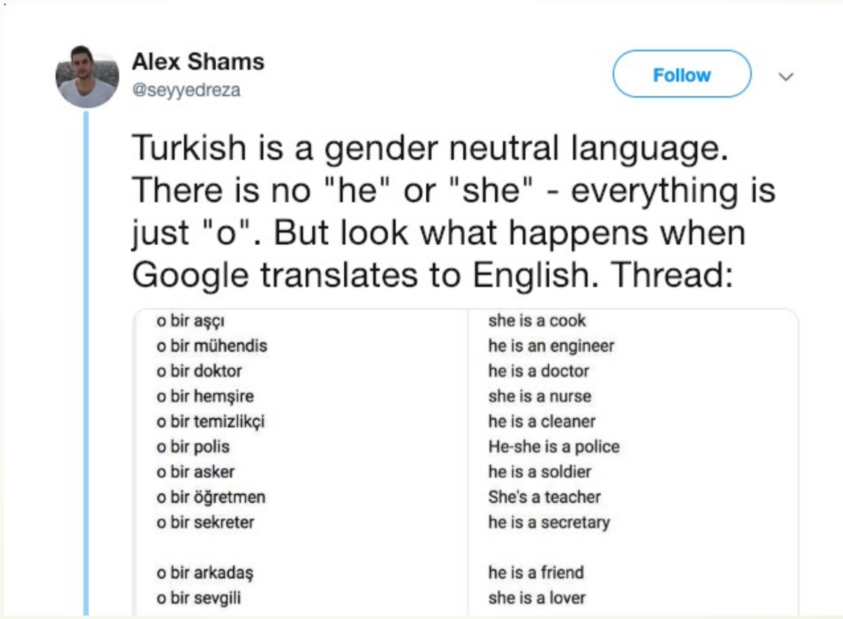
Education

- Consider: facial recognition in public places
  - City centers, airports
  - Concerns of error, function creep, and privacy
    - https://www.emeraldinsight.com/doi/pdfplus/10.1108/14779960480000246 (paywall)
  - Emotional privacy
    - Masking emotions
    - Social cohesion
    - http://blog.practicalethics.ox.ac.uk/2014/03/computer-vision-and-emotional-privacy/

https://www.emeraldinsight.com/doi/abs/10.1108/14779960480000246

Education

- Consider: societal impacts of natural language processing



Alex Shams
@seyyedreza

Turkish is a gender neutral language. There is no "he" or "she" - everything is just "o". But look what happens when Google translates to English. Thread:

o bir aşçı	she is a cook
o bir mühendis	he is an engineer
o bir doktor	he is a doctor
o bir hemşire	she is a nurse
o bir temizlikçi	he is a cleaner
o bir polis	He-she is a police
o bir asker	he is a soldier
o bir öğretmen	She's a teacher
o bir sekreter	he is a secretary
o bir arkadaş	he is a friend
o bir sevgili	she is a lover

Education

- Exclusion and demographic bias
- Overgeneralization and confirmation bias
- Topic overexposure (availability heuristic) and underexposure
- http://aclweb.org/anthology/P16-2096

- NLP ethical best practices http://aclweb.org/anthology/W17-1604.pdf

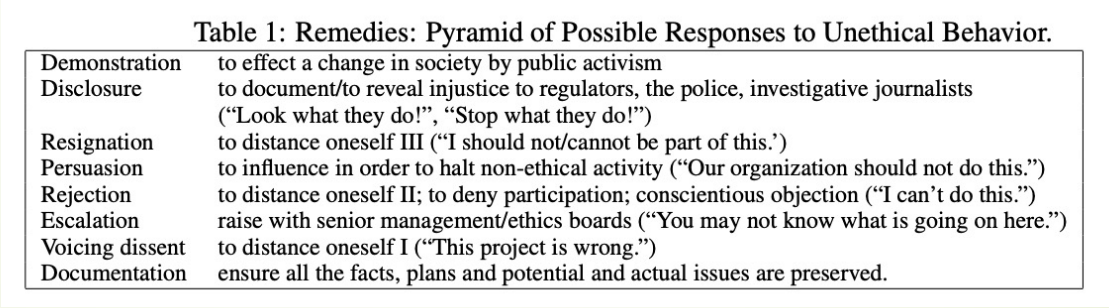
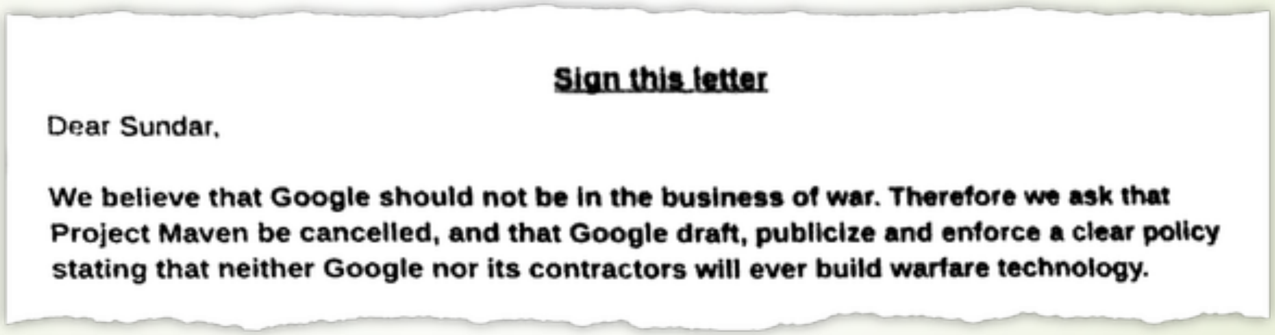


Table 1: Remedies: Pyramid of Possible Responses to Unethical Behavior.

Demonstration	to effect a change in society by public activism
Disclosure	to document/to reveal injustice to regulators, the police, investigative journalists (“Look what they do!”, “Stop what they do!”)
Resignation	to distance oneself III (“I should not/cannot be part of this.’)
Persuasion	to influence in order to halt non-ethical activity (“Our organization should not do this.”)
Rejection	to distance oneself II; to deny participation; conscientious objection (“I can’t do this.”)
Escalation	raise with senior management/ethics boards (“You may not know what is going on here.”)
Voicing dissent	to distance oneself I (“This project is wrong.”)
Documentation	ensure all the facts, plans and potential and actual issues are preserved.

Education

- Consider: “dual-use” technologies
  - Technologies designed for civilian use but which may have military applications
  - Google’s Project Maven, software for automated drone surveillance for the Pentagon



Sign this letter

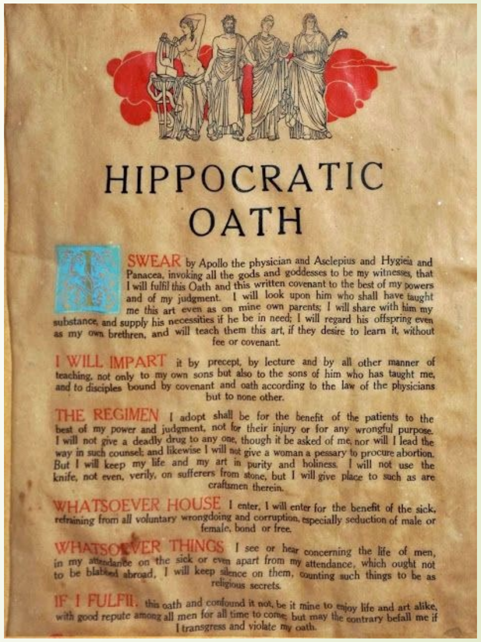
Dear Sundar,

We believe that Google should not be in the business of war. Therefore we ask that Project Maven be cancelled, and that Google draft, publicize and enforce a clear policy stating that neither Google nor its contractors will ever build warfare technology.

  - Microsoft employees have protested the company’s involvement in the same Department of Defense program
  - Amazon’s collaboration with the US Immigration and Customs Enforcement (ICE)

Education

- **Codes of Ethics**
- Some have advocated for a "Data Science Hippocratic Oath"
- IEEE and ACM organizations have explicit codes of ethics
- AI research is arguably unique, but NeurIPS 2018 has a Code of Conduct and the ML Ally Pledge is similarly well constructed
- Many other institutions and "influencing" organizations have begun making their own



HIPPOCRATIC OATH

I SWEAR by Apollo the physician and Asclepius and Hygieia and Panacea, invoking all the gods and goddesses to be my witnesses, that I will fulfil this Oath and this written covenant to the best of my powers and of my judgment. I will look upon him who shall have taught me this art even as on mine own parents; I will share with him my substance, and supply his necessities if he be in need; I will regard his offspring even as my own brethren, and will teach them this art, if they desire to learn it, without fee or covenant.

I WILL IMPART it by precept, by lecture and by all other manner of teaching, not only to my own sons but also to the sons of him who has taught me, and to disciples bound by covenant and oath according to the law of the physicians but to none other.

THE REGIMEN I adopt shall be for the benefit of the patients to the best of my power and judgment, not for their injury or for any wrongful purpose. I will not give a deadly drug to any one, though it be asked of me, nor will I lead the way in such counsel; and likewise I will not give a woman a pessary to procure abortion. But I will keep my life and my art in purity and holiness. I will not use the knife, not even, verily, on sufferers from stone, but I will give place to such as are craftsmen therein.

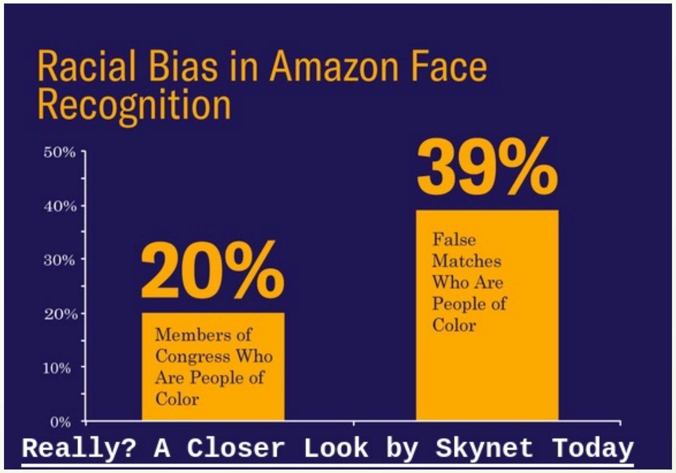
WHATSOEVER HOUSE I enter, I will enter for the benefit of the sick, refraining from all voluntary wrongdoing and corruption, especially seduction of male or female, bond or free.

WHATSOEVER THINGS I see or hear concerning the life of men, in my attendance on the sick or even apart from my attendance, which ought not to be blabbed abroad, I will keep silence on them, counting such things to be as religious secrets.

IF I FULFIL this oath and confound it not, be it mine to enjoy life and art alike, with good repute among all men for all time to come; but may the contrary befall me if I transgress and violate my oath.

Communication

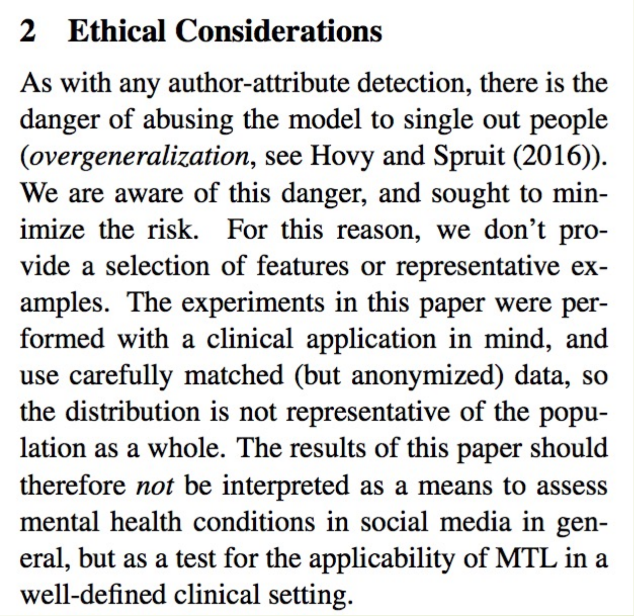
- Potential misuses and ethical considerations of new AI and data science algorithms / packages are rarely identified and pointed out, either in documentation or in academic papers
  - Amazon’s Rekognition product for facial recognition did not warn about the high false positive rate associated with its default parameters



Communication

- New “Ethical Considerations” section in published works (academic, code documentation, blog posts, etc)

- Margaret Mitchell, head of Google Research and Machine Intelligence at Google Brain
  - 2017 paper flagging patient suicide risk in clinical settings given their writings as input
  - Point out clear cases of potential ethical mis-use and how their study mitigated these concerns



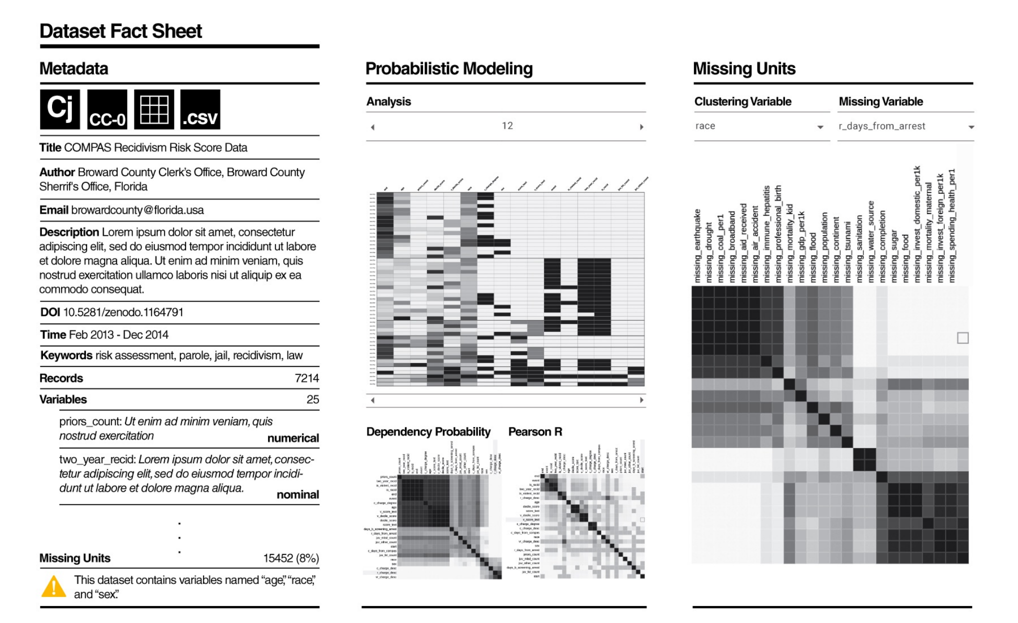
2 Ethical Considerations

As with any author-attribute detection, there is the danger of abusing the model to single out people (*overgeneralization*, see Hovy and Spruit (2016)). We are aware of this danger, and sought to minimize the risk. For this reason, we don’t provide a selection of features or representative examples. The experiments in this paper were performed with a clinical application in mind, and use carefully matched (but anonymized) data, so the distribution is not representative of the population as a whole. The results of this paper should therefore *not* be interpreted as a means to assess mental health conditions in social media in general, but as a test for the applicability of MTL in a well-defined clinical setting.

Communication

- Standardizing means of communicating aspects of new datasets and AI services

- Datasheets for datasets
- Data statements for NLP
- Policy certificates for RL
- Declarations of AI service conformity



Dataset Fact Sheet

Metadata

Cj CC-0 .csv

Title COMPAS Recidivism Risk Score Data

Author Broward County Clerk's Office, Broward County Sherrif's Office, Florida

Email browardcounty@florida.usa

Description Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

DOI 10.5281/zenodo.1164791

Time Feb 2013 - Dec 2014

Keywords risk assessment, parole, jail, recidivism, law

Records 7214

Variables 25

priors_count: *Ut enim ad minim veniam, quis nostrud exercitation* **numerical**

two_year_recid: *Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.* **nominal**

·
·
·

Missing Units 15452 (8%)

⚠ This dataset contains variables named "age," "race," and "sex."

Probabilistic Modeling

Analysis

12

Dependency Probability **Pearson R**

Missing Units

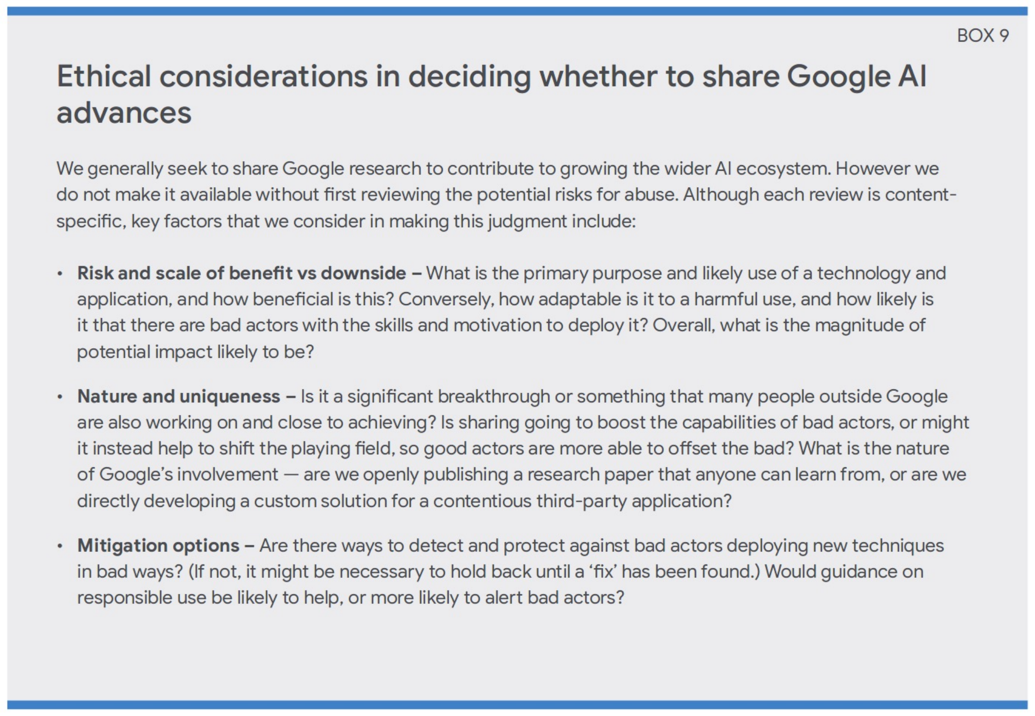
Clustering Variable race

Missing Variable r_days_from_arrest

Distribution

- Approval and Terms of Access for datasets, code, and models
- ImageNet
  - One of the most important computer vision datasets of the decade
  - Downloading it requires agreeing to terms of access!
  - Admittedly increases overhead for host lab or organization, but helps mitigate the dual-use problem
- A “Responsible AI License” for code and pre-trained models

Distribution



BOX 9

Ethical considerations in deciding whether to share Google AI advances

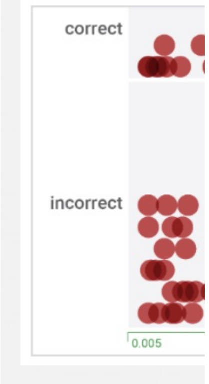
We generally seek to share Google research to contribute to growing the wider AI ecosystem. However we do not make it available without first reviewing the potential risks for abuse. Although each review is content-specific, key factors that we consider in making this judgment include:

- **Risk and scale of benefit vs downside –** What is the primary purpose and likely use of a technology and application, and how beneficial is this? Conversely, how adaptable is it to a harmful use, and how likely is it that there are bad actors with the skills and motivation to deploy it? Overall, what is the magnitude of potential impact likely to be?
- **Nature and uniqueness –** Is it a significant breakthrough or something that many people outside Google are also working on and close to achieving? Is sharing going to boost the capabilities of bad actors, or might it instead help to shift the playing field, so good actors are more able to offset the bad? What is the nature of Google's involvement — are we openly publishing a research paper that anyone can learn from, or are we directly developing a custom solution for a contentious third-party application?
- **Mitigation options –** Are there ways to detect and protect against bad actors deploying new techniques in bad ways? (If not, it might be necessary to hold back until a 'fix' has been found.) Would guidance on responsible use be likely to help, or more likely to alert bad actors?

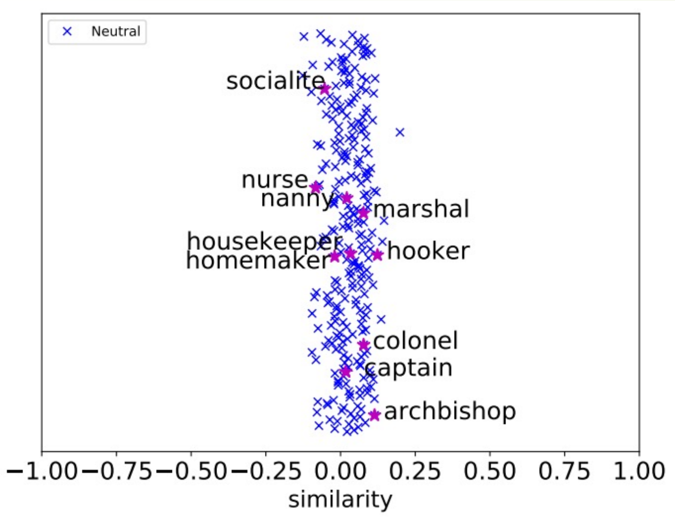
Distribution

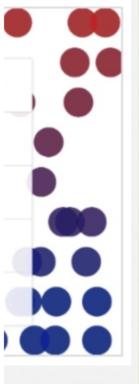
- Use, share, and create emerging tools to detect bias and explore datasets for ethical considerations

- IBM’s AI Fairness 360
- Google’s What-If
- gn_glove, a gender-neutral word2vec-like embedding



https:





https://arxiv.org/pdf/1809.01496.pdf

Advocacy

- **This is where we ALL come in**
- Bring up concerns in talks and classrooms (like this one!)
- Dedicate part of the syllabus (like this one)
- Take an entire class on AI and Ethics, Ethics and Philosophy

November 9, 2018
Alcohol Notice
SALON
Stanford ARTIFICIAL INTELLIGENCE
"Best Practices for Ethical AI Research"
with guests Margaret Mitchell and Timnit Gebru
Feedback
bit.ly/aisa

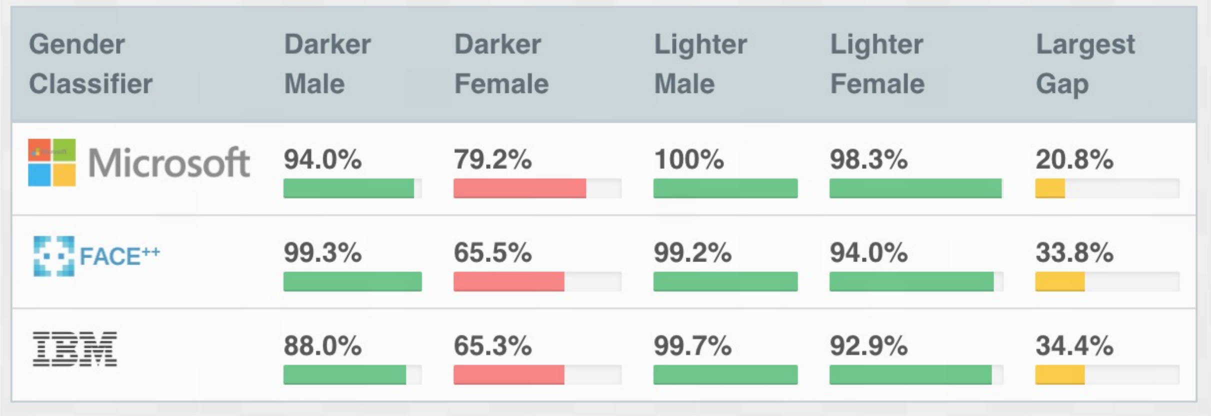
Image from one of Stanford AI Lab's 'AI Salon' events on Best Practices in doing Ethical AI Research

Advocacy

- Obtain and promote more diverse research perspectives
- In 2017, Joy Buolamwini found facial recognition platforms at Microsoft, IBM, and Face++ did very poorly when identifying women and minorities

Advocacy

- While each service touted an excellent overall accuracy, certain subgroups performed very poorly



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

- She created http://gendershades.org/ and contacted each company regarding their inclusion and diversity practices during development

Advocacy

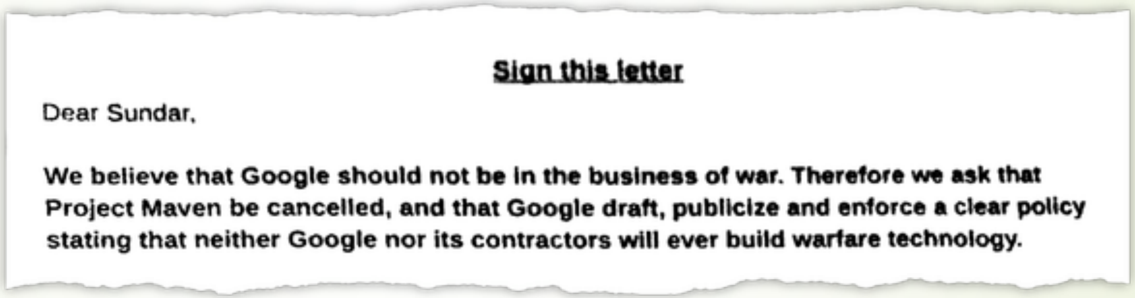
- Small and large-scale initiatives

- AI4ALL
- Women in AI
- Black in AI

- AI Now and NYU
- Human-Centered AI Institute at Stanford
- Ada Lovelace Institute
- AAAI/ACM Conference on AI, Ethics, and Society

Advocacy

- And don’t forget: **you**
  - **Take a stand** against unethical decisions



Sign this letter

Dear Sundar,

We believe that Google should not be in the business of war. Therefore we ask that Project Maven be cancelled, and that Google draft, publicize and enforce a clear policy stating that neither Google nor its contractors will ever build warfare technology.

  - Employees of Amazon and Microsoft have likewise worked to withdraw their respective companies from DoD contracts
  - Expand your own intellectual and research circles to include other viewpoints

Conclusions

- Data science and artificial intelligence are only going to become more intertwined with our daily lives (self-driving cars, smart homes, internet-of-things)
- Automated decision-making has the potential to shape our civilization on a large scale
- Understanding this technology and the strengths and limitations of its abilities is critical as we integrate it ever more deeply into our everyday routines
- Being able to interface not only with researchers, but with policymakers, legislators, and the public is going to be essential
- Can no longer afford to hide behind the ivory tower and ignore the implications of our work, and its unintended consequences

References

- Slides https://thegradient.pub/in-favor-of-developing-ethical-best-practices-in-ai-research/
- Gender Shades http://gendershades.org/index.html
- Stanford Sexual Orientation https://www.nytimes.com/2017/10/09/science/stanford-sexual-orientation-study.html
- Amazon Rekognition https://www.nytimes.com/2019/01/24/technology/amazon-facial-technology-study.html
- ACM FAccT Proceedings https://facctconference.org/
- AI Ethics Resources https://www.fast.ai/2018/09/24/ai-ethics-resources/

Questions?