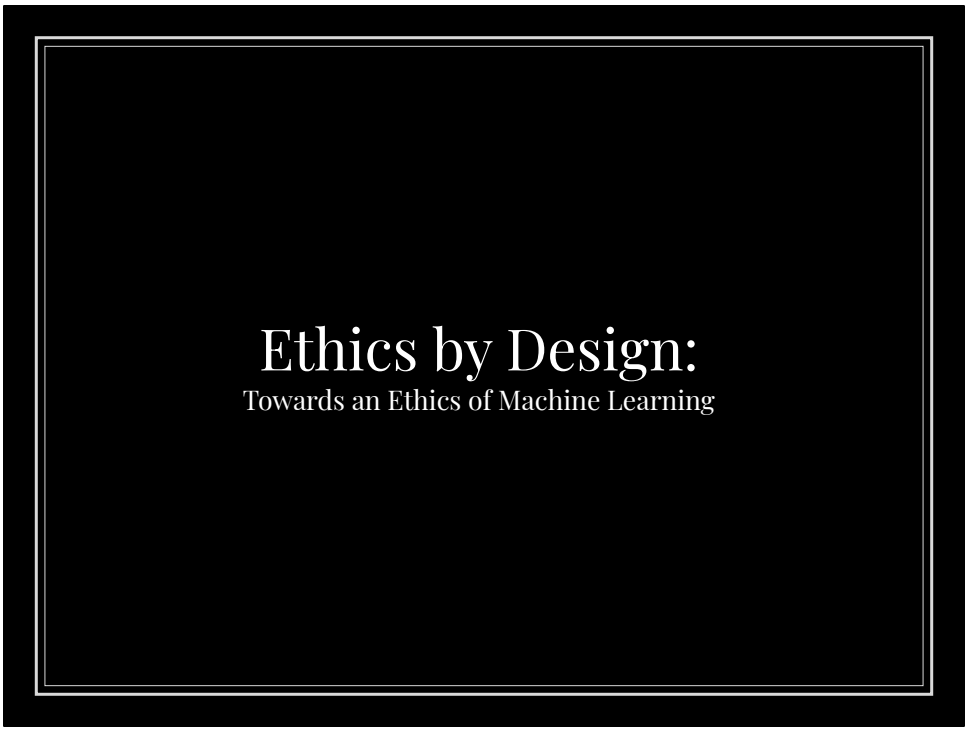


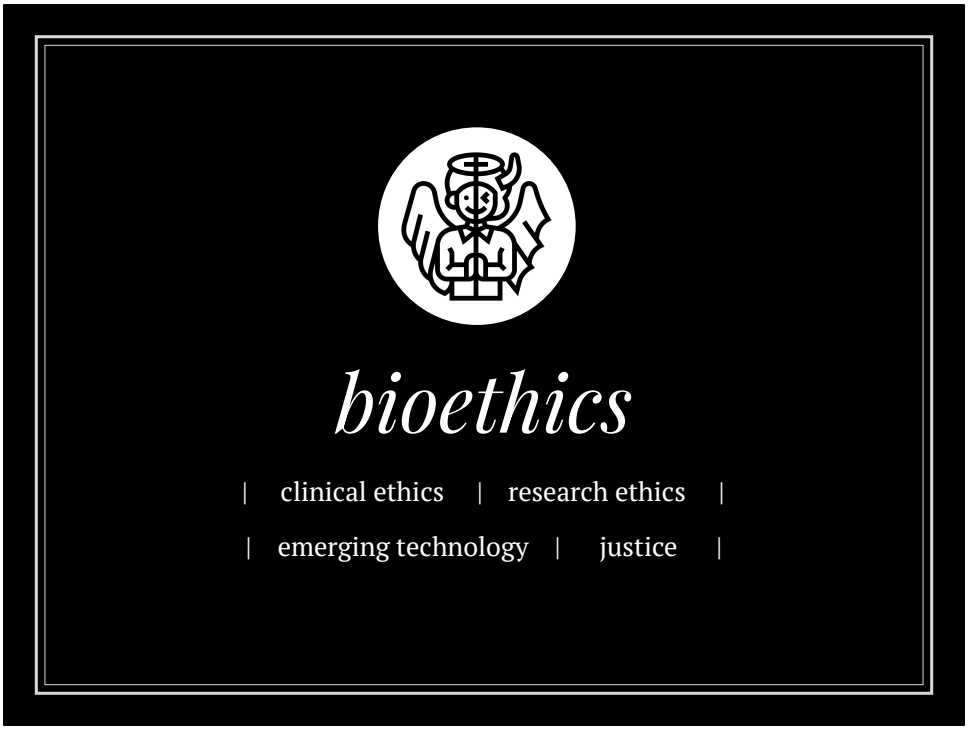
Title inspired by written interview of Jess Whittlestone, at https://www.technologyreview.com/2020/06/24/1004432/ai-help-crisis-new-kind-ethics-machine-learning-pandemic/

Second half of title inspired by podcast interview of Timnit Gebru, who speaks to trends towards fairness in ML, at https://twimlai.com/twiml-talk-336-trends-in-fairness-and-ai-ethics-with-timnit-gebru/

The opinions expressed in this article are the author's own and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, or the United States government.

2

On Twitter @ejardas
Best way to contact me: ejardas@uwalumni.com



Icon made by Eucalyp.

Agenda

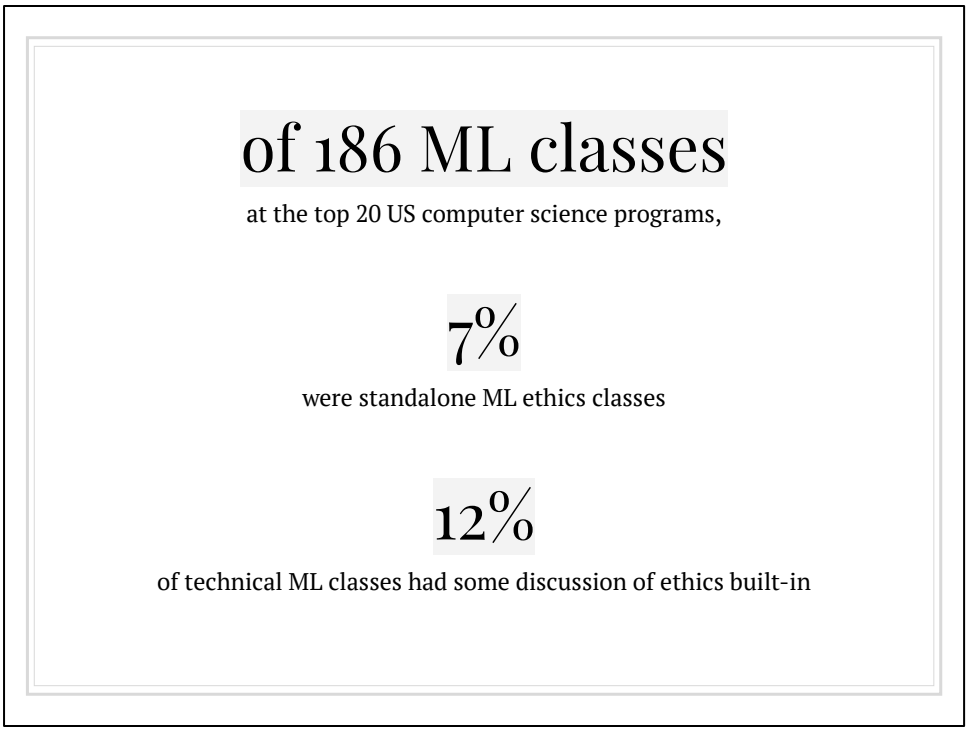
I. Introduction to the problem
II. The scope of ethics, machine learning, and society
III. 4 phases of ethics by design
IV. A few solutions…
V. Resources to connect with

ML/AI = machine learning / artificial intelligence

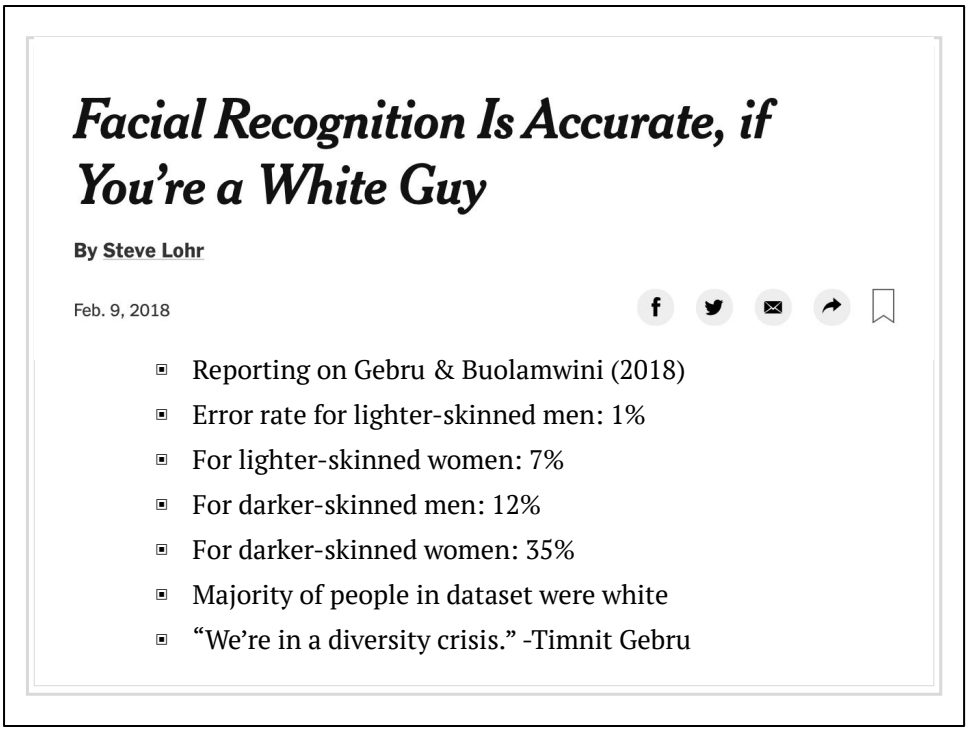
I.
introduction to the problem

is ethics integrated into machine learning?

Saltz et al. Integrating ethics with machine learning courses. (2019). https://dl.acm.org/doi/10.1145/3341164 (please email me if you cannot download the article!)



Saltz et al. Integrating ethics with machine learning courses. (2019). https://dl.acm.org/doi/10.1145/3341164 (please email me if you cannot download the article!)



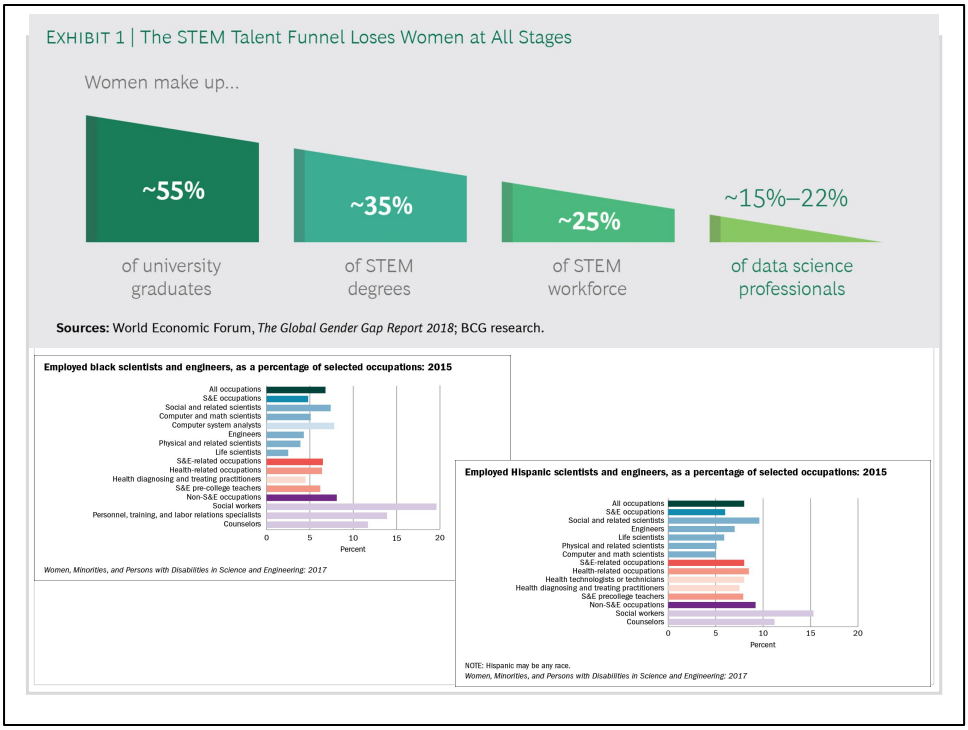
Facial Recognition Is Accurate, if You're a White Guy

By Steve Lohr

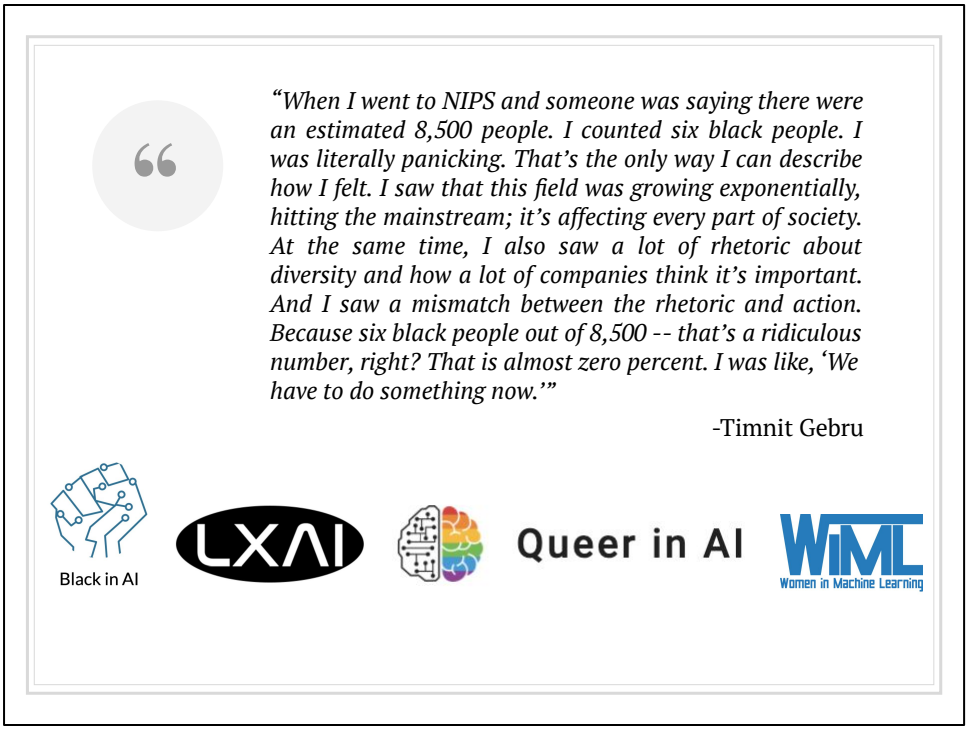
Feb. 9, 2018

- Reporting on Gebru & Buolamwini (2018)
- Error rate for lighter-skinned men: 1%
- For lighter-skinned women: 7%
- For darker-skinned men: 12%
- For darker-skinned women: 35%
- Majority of people in dataset were white
- "We're in a diversity crisis." -Timnit Gebru

Steve Lohr. Facial recognition is accurate, if you're a white guy. The New York Times (2018). https://nyti.ms/2BNurVq
Reporting on study by Gebru & Buolamwini (2018), Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Paper: http://proceedings.mlr.press/v81/buolamwini18a.html. Video: http://gendershades.org/
Buolamwini's TedTalk: https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms#t-9093



Duranton et al. What's keeping women out of data science? BCG (2020). https://www.bcg.com/publications/2020/what-keeps-women-out-data-science.aspx
NSF. Women, Minorities, and Persons with Disabilities in Science and Engineering. NSF.gov (2015). https://www.nsf.gov/statistics/2017/nsf17310/digest/occupation/blacks.cfm



Quote from
https://www.technologyreview.com/2018/02/14/145462/were-in-a-diversity-crisis-black-in-ais-founder-on-whats-poisoning-the-algorithms-in-our/
Black in AI: https://blackinai.github.io/
LatinX in AI: https://www.latinxinai.org/
Queer in AI: https://sites.google.com/view/queer-in-ai/
WiML: https://wimlworkshop.org/

II.
the scope of ethics, machine learning, and society

The Scope

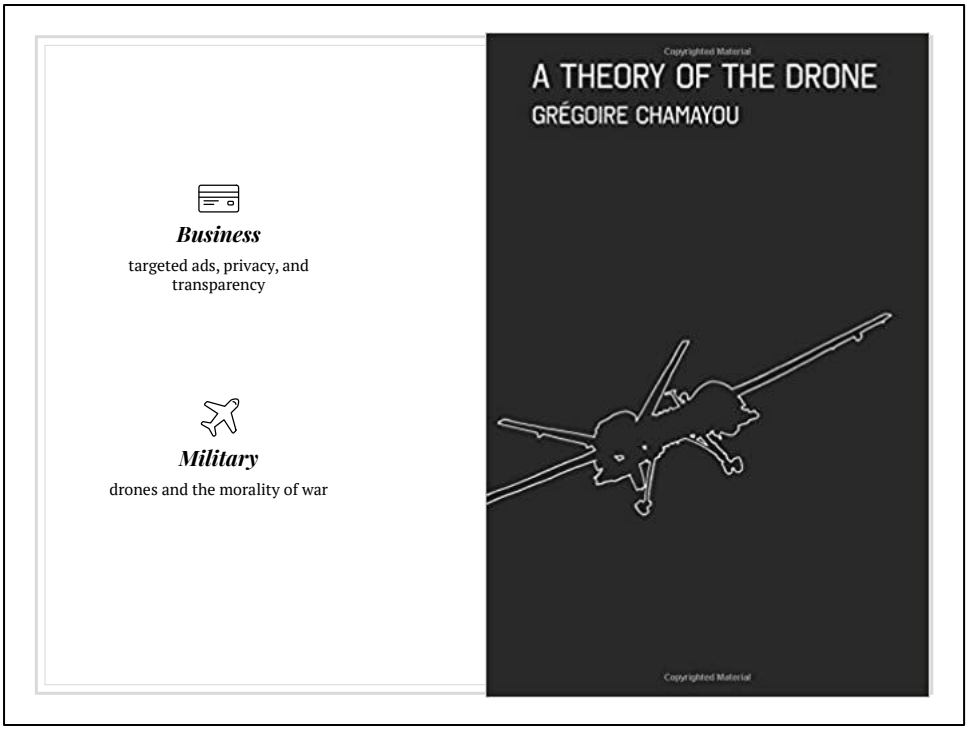
Business

targeted ads, privacy, and transparency

"x would like permission to track you across apps and websites owned by other companies. Your data will be used to deliver personalized ads to you."

Icons made by http://www.flaticon.com/authors/freepik

Kate O'Flaherty. Apple IOS 14: Is Facebook and Google's Worst Nightmare Coming True? Forbes (June 2020). https://www.forbes.com/sites/kateoflahertyuk/2020/06/25/apple-ios-14-is-facebook-and-googles-worst-nightmare-coming-true/#77eb55ac4335



Icons made by http://www.flaticon.com/authors/freepik
Gregorie Chamayou. A theory of the drone. (2015). https://thenewpress.com/books/theory-of-drone

The Scope

Business

targeted ads, privacy, and transparency

Technology

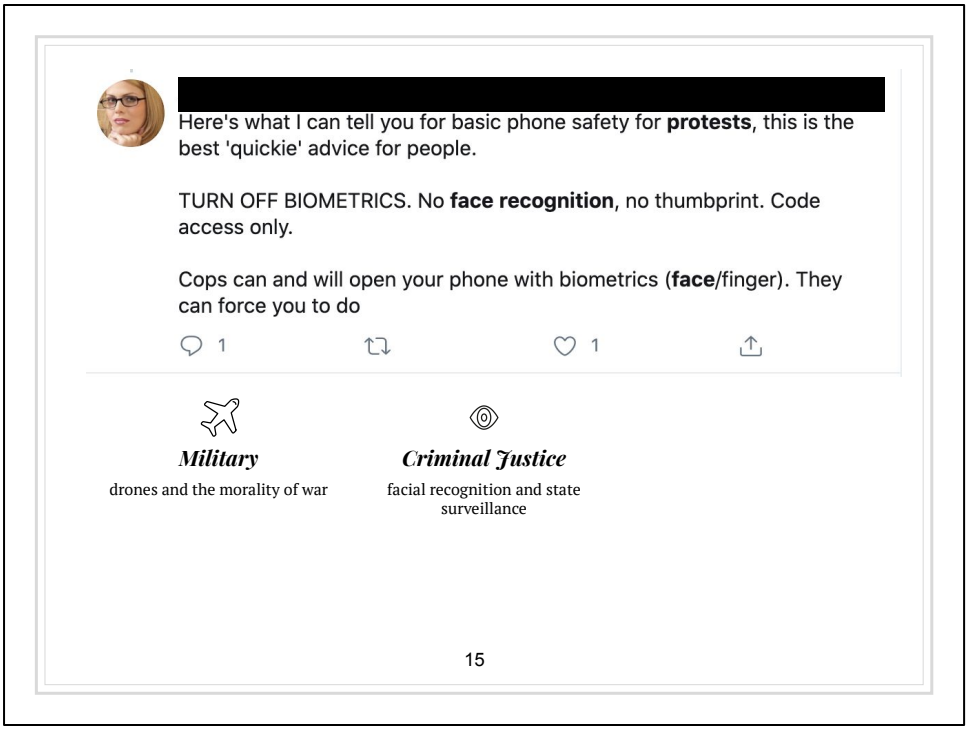
self-driving cars and moral responsibility

Military

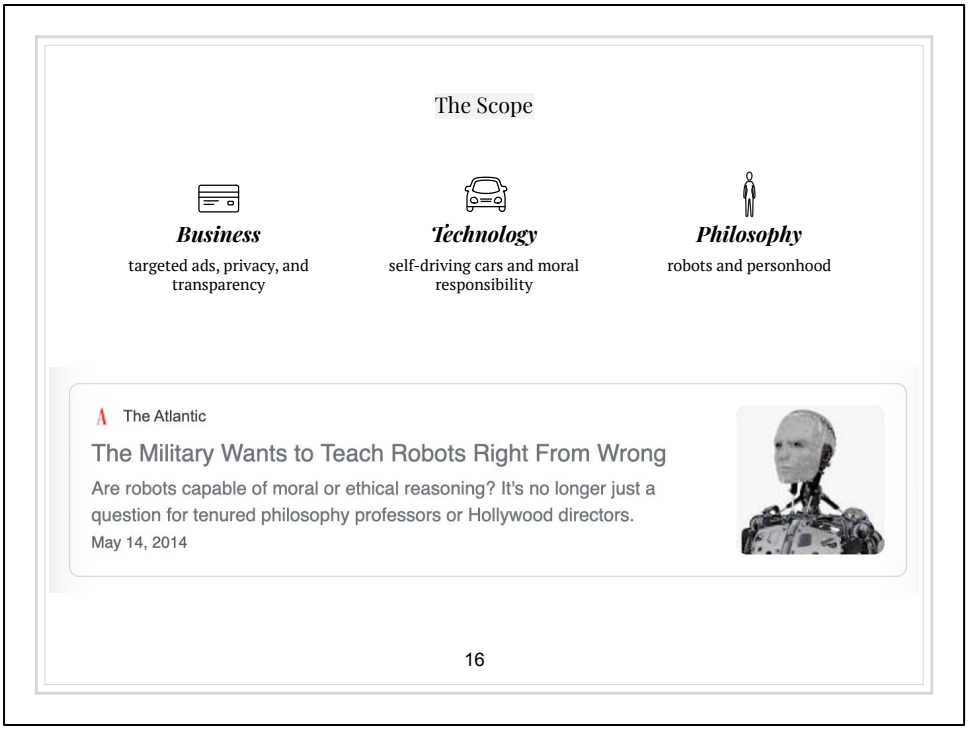
drones and the morality of war

Icons made by http://www.flaticon.com/authors/freepik

Jonathan Shaw. Confronting pitfalls of machine learning. Harvard Magazine (2019). https://www.harvardmagazine.com/2019/01/artificial-intelligence-limitations



Icons made by http://www.flaticon.com/authors/freepik
Gregory Fowler. Black Lives Matter could change facial recognition forever — if Big Tech doesn't stand in the way. The Washington Post (2020) https://www.washingtonpost.com/technology/2020/06/12/facial-recognition-ban/
Also recommend Radical AI Podcast: IBM, Microsoft, and Amazon Disavow Facial Recognition Technology: What Do You Need to Know? with Deb Raji https://radicalai.podbean.com/e/ibm-microsoft-and-amazon-disavow-facial-recognition-technology-what-do-you-need-to-know-with-deb-raji/ (Also available on any Podcast platform such as Spotify, Apple Podcasts etc.)



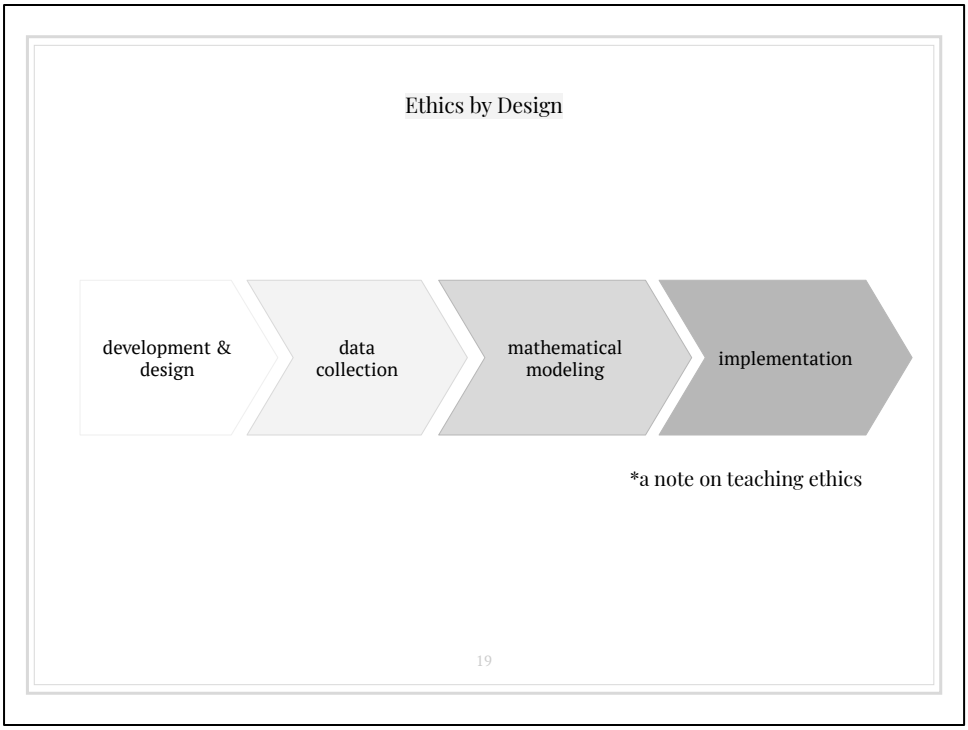
Icons made by "http://www.flaticon.com/authors/freepik"

Patrick Tucker and Defense One. The military wants to teach robots right from wrong. The Atlantic (2014). https://www.theatlantic.com/technology/archive/2014/05/the-military-wants-to-teach-robots-right-from-wrong/370855/

Icons made by http://www.flaticon.com/authors/freepik

Natasha Singer and Daisuke Wakabayashi. Google to Store and Analyze Millions of Health Records. The New York Times (2019). https://www.nytimes.com/2019/11/11/business/google-ascension-health-data.html

III.
4 phases of ethics by design



*A note on teaching ethics.
I have included a lot of news articles, podcasts, and videos that should be accessible to students.
A lot of the work of the ethicist is to reflect on and identify the ethical questions. You could have students come up with their own ideas: what are the important ethicals values at play in machine learning (privacy, justice, etc)? What questions should the developers of these algorithms be asking themselves?
Alternatively you could take one or more of the ethical questions that I identify myself in this presentation and talk about them as a class. For example, my question, “how will the data be stored?”, could be posed as: “how should private data be stored in order for this to be ethical?”

step 1.
development & design

Courtney Martin. The reductive seduction of other people's problems. BRIGHT Magazine (2016). https://brightthemag.com/the-reductive-seduction-of-other-people-s-problems-3c07b307732d

Also recommend: Jessica Lnew. SWEDOW: or why those kids don't need your crayons. Good Intentions, Bad Aid Blog (2014). https://goodintentionsbadaid.wordpress.com/2014/03/31/swedow-or-why-those-kids-dont-need-your-crayons/
and
Laura Freschi. Stuff We Don't Want Flowchart. Aid Watch Blog (2010). http://www.nyudri.org/aidwatcharchive/2010/05/the-%E2%80%9Cstuff-we-don%E2%80%99t-want%E2%80%9D-flow-chart

Development & Design

- What problem are you trying to fix?
- Who are you trying to serve?
- Are you solving problems with people, not for people?
- Do you have expertise on the problem?
- Do you have expertise on the methodology?
- Are you engaging community stakeholders?
- Is this what the community wants researchers or the government to spend time and money on?
- How could this tool be used for good? for evil?

POLICY | TECH | LAW

ICE rigged its algorithms to keep immigrants in jail, claims lawsuit

A 'secret no-release policy'

By Adi Robertson | @thedextriarchy | Mar 3, 2020, 12:59pm EST

RESEARCH ARTICLE

Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States

Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei

PNAS December 12, 2017 114 (50) 13108-13113; first published November 28, 2017
https://doi.org/10.1073/pnas.1700035114

Adi Robertson. Ice rigged its algorithms to keep immigrants in jail, claims lawsuit. The Verge (2020). https://www.theverge.com/2020/3/3/21163013/ice-new-york-risk-assessment-algorithm-rigged-lawsuit-nyclu-jose-velesaca

Gebru et al. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. PNAS (2017). https://doi.org/10.1073/pnas.1700035114 (Please email me if you cannot access this article!)

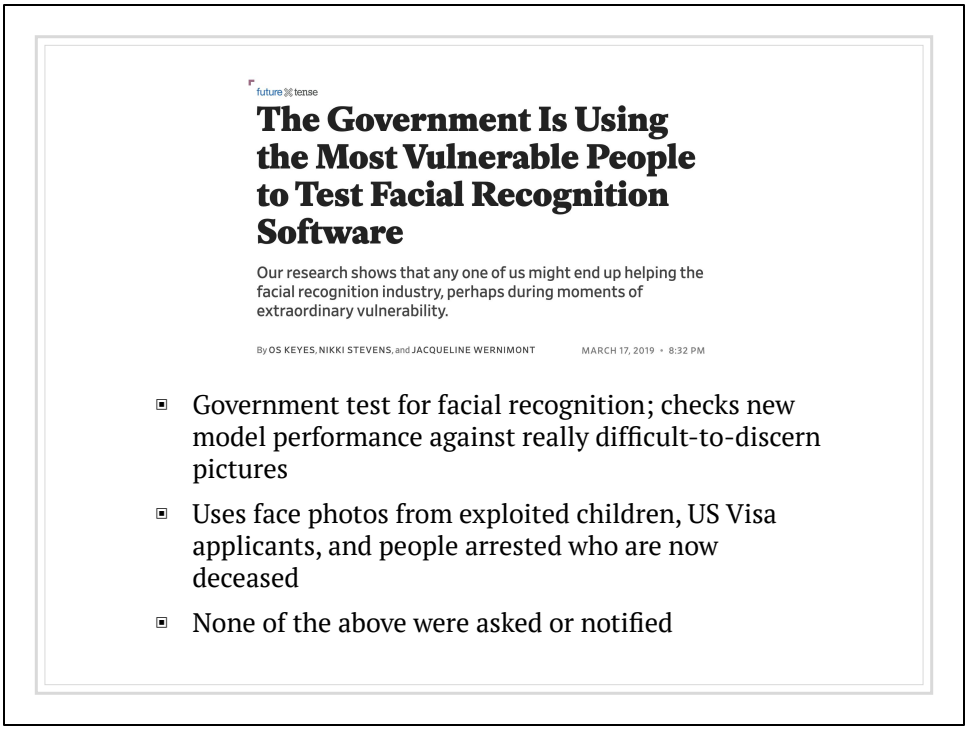
There are places in the world where the infrastructure is not there and the resources are not there to send people door to door and gather [census] data, [but where] having an understanding of the different types of populations that live in your country would be very helpful. But then again, this is exactly the thing that also made me want to study fairness. Because if I'm going to be continuing to do this line of work, I really need to have a better understanding of the potentially negative repercussions. What are the repercussions for surveillance? Also, what are the repercussions for a data-set bias?

-Timnit Gebru

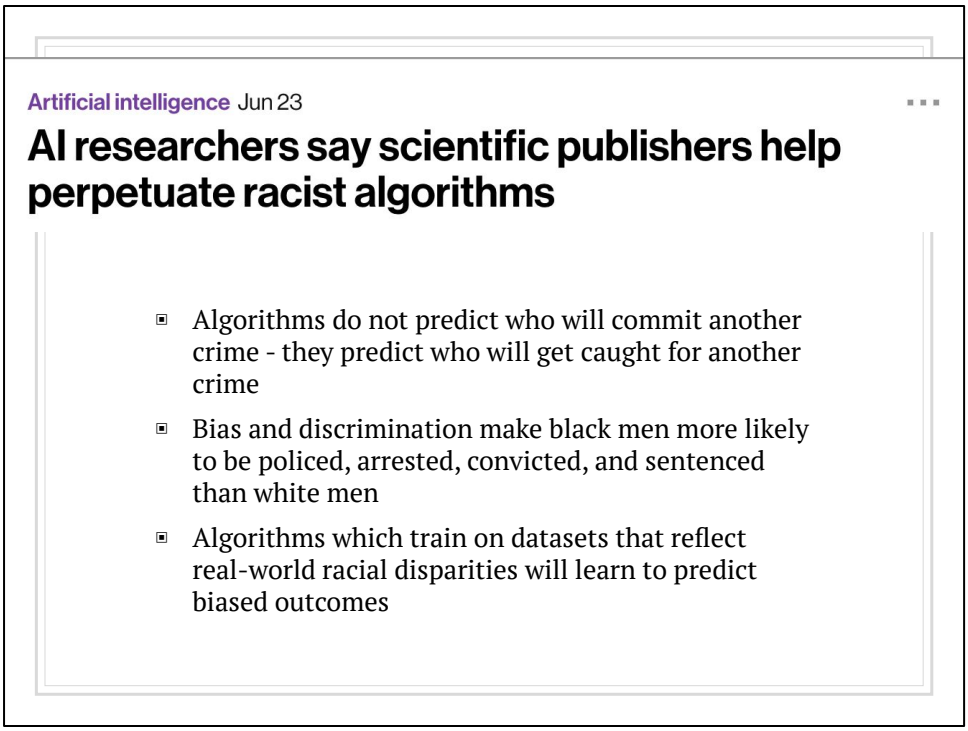
step 2.
data collection

Data Collection

- Are the data carefully collected?
- Are the data truly suitable to answer the question?
- Whose data are you using? How did you get them?
- Do you have ‘permission’ or consent to use their data?
- How will the data be stored?
- Are the data representative?
- Do the data reflect disparities?



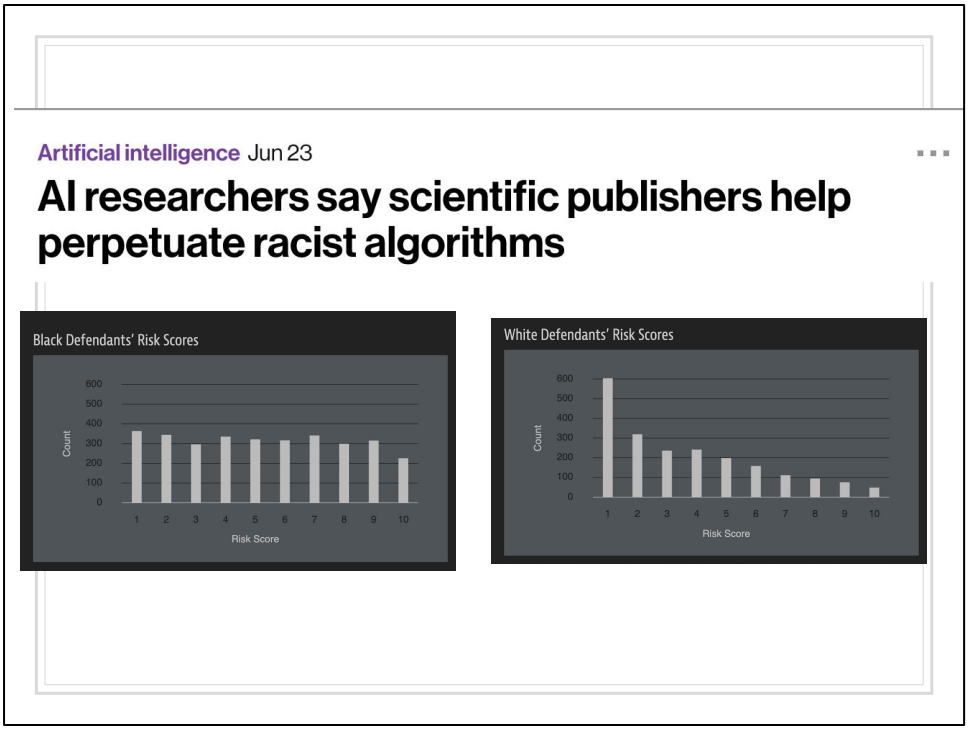
Keyes, Stevens, & Wernimont. The government is using the most vulnerable people to test facial recognition software.Slate (2019). https://slate.com/technology/2019/03/facial-recognition-nist-verification-testing-data-sets-children-immigrants-consent.html



Karen Hao. AI researchers say scientific publishers help perpetuate racist algorithms. MIT Technology Review (2020). https://www.technologyreview.com/2020/06/23/1004333/ai-science-publishers-perpetuate-racist-face-recognition/

Machine Bias. ProPublica (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Also see Joe Forward. The Loomis v. Wisconsin case: proprietary use of algorithms at sentencing. State Bar of Wisconsin (2017). https://www.wisbar.org/NewsPublications/InsideTrack/Pages/Article.aspx?Volume=9&Issue=14&ArticleID=25730



Karen Hao. AI researchers say scientific publishers help perpetuate racist algorithms. MIT Technology Review (2020). https://www.technologyreview.com/2020/06/23/1004333/ai-science-publishers-perpetuate-racist-face-recognition/

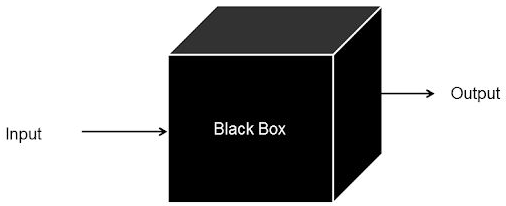
Machine Bias. ProPublica (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Also see Joe Forward. The Loomis v. Wisconsin case: proprietary use of algorithms at sentencing. State Bar of Wisconsin (2017). https://www.wisbar.org/NewsPublications/InsideTrack/Pages/Article.aspx?Volume=9&Issue=14&ArticleID=25730

step 3.
mathematical modeling

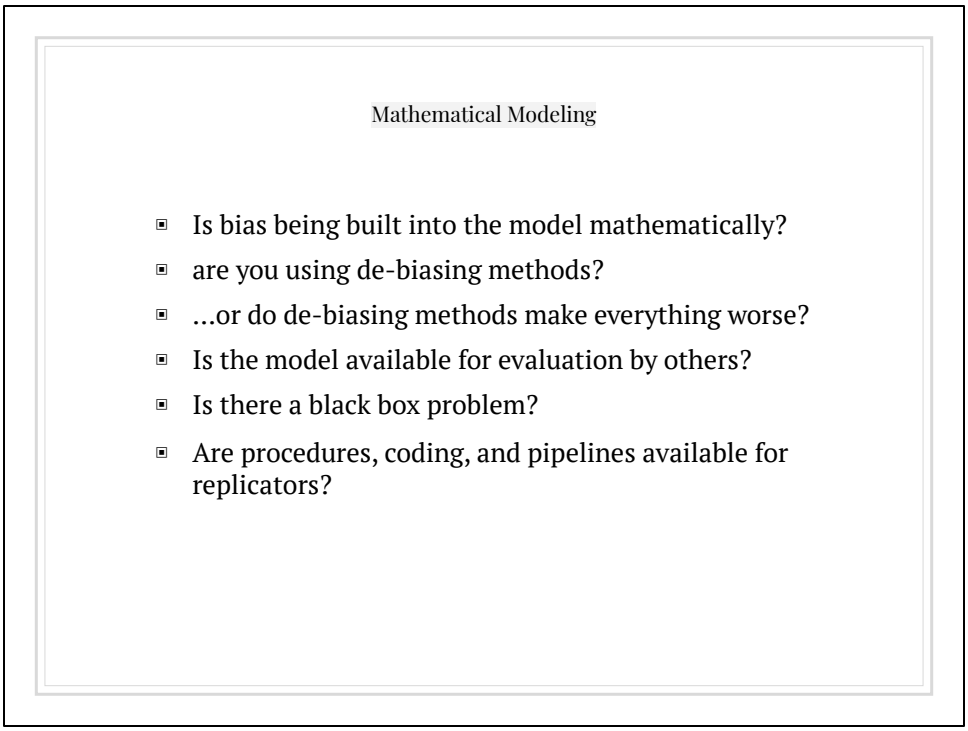
Mathematical Modeling

- Is bias being built into the model mathematically?
- are you using de-biasing methods?
- ...or do de-biasing methods make everything worse?
- Is the model available for evaluation by others?
- Is there a black box problem?



Internal behavior of the code is unknown

Gonen & Goldberg. Lipstick on a pig: debiasing methods cover up systematic gender biases in word embeddings but do not remove them. NAACL (2019). https://arxiv.org/abs/1903.03862v2



Gonen & Goldberg. Lipstick on a pig: debiasing methods cover up systematic gender biases in word embeddings but do not remove them. NAACL (2019). https://arxiv.org/abs/1903.03862v2

- Risk assessment to predict re-offense
- Tended to give higher risk scores to black men than white men, despite severity and frequency of past crimes suggesting the opposite
- *Loomis v. Wisconsin* (2016), court ruled:
  - permissible that COMPAS is not open-source
  - no evidence of gender bias
  - judge must be told about racial bias and other problems with COMPAS before shown risk score
  - Risk score cannot be only determining factor of sentencing

Also see Joe Forward. The Loomis v. Wisconsin case: proprietary use of algorithms at sentencing. State Bar of Wisconsin (2017). https://www.wisbar.org/NewsPublications/InsideTrack/Pages/Article.aspx?Volume=9&Issue=14&ArticleID=25730

State of Wisconsin v. Loomis. Harvard Law Review (2017). https://harvardlawreview.org/2017/03/state-v-loomis/

step 4.
implementation

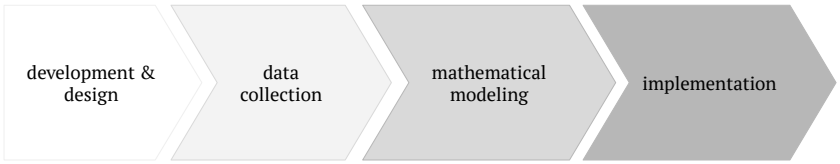
Implementation

- Does the model work outside of its development? (Is there a plan to test this?)
- Is the model's performance compared to the current baseline? Is any reported gain worth the trade-offs?
- Does the model work well for some people and not for others?
- Who has access to the model? Does it still serve the intended population?
- What do doctors, patients / clients, etc think about the use of the model?
- Are patients informed about the use of the model?
- Who is liable in case of a prediction error?
- Is the model regularly updated, fixed, or revised?
- Is the model cost-effective to maintain?

- There are currently 3 treatments for quitting smoking.
- At the doctor's office, the doctors have no way of knowing which will work best for you, so they give you one of the three to try first at random. If it doesn't work, they will give you the next one at random.
- None of the treatments are more harmful than the others.
- A model is developed that predicts which of the 3 treatments work best for you with great accuracy.
- Would you want this algorithm to be used by your doctor to help you quit smoking?

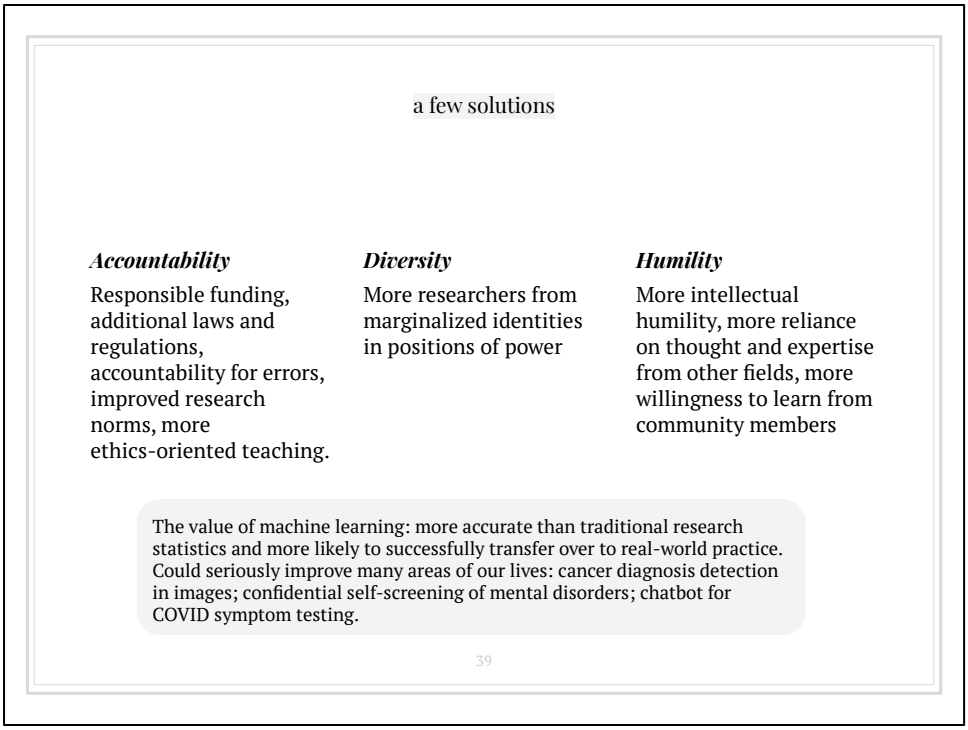
None of these data have been reported or published yet, so unfortunately there are no links. Study being conducted by Addiction Research Center at University of Wisconsin-Madison, see https://arc.psych.wisc.edu/

Ethics by Design

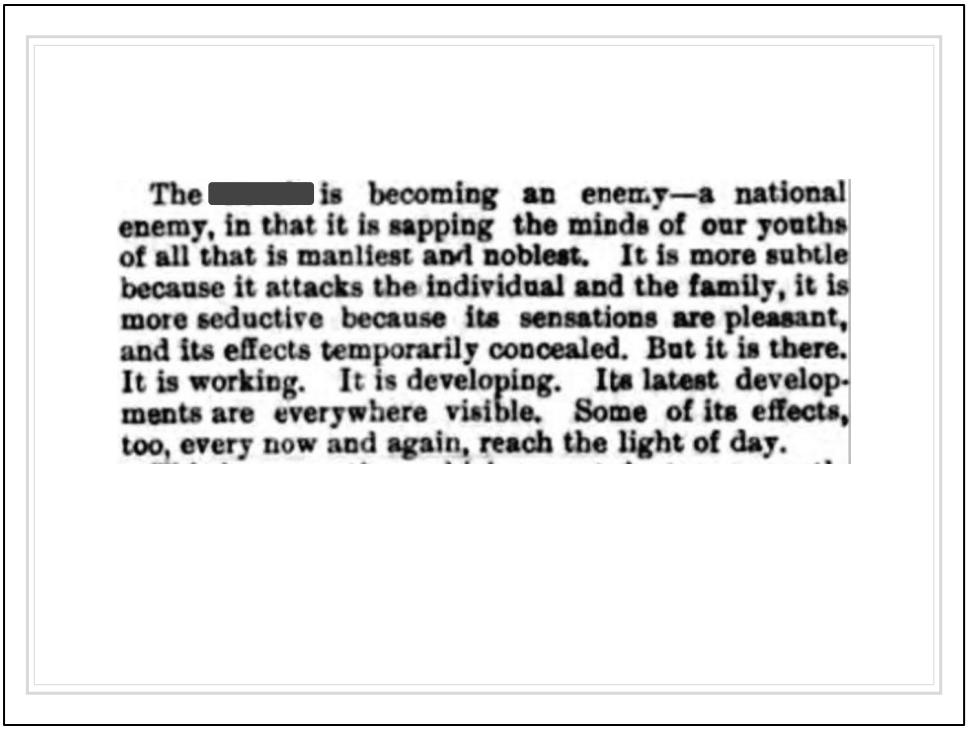


*most of these topics (privacy; justice) are common problems in many other areas of research. bioethicists have been working on frameworks for dealing with these issues in those contexts for many years.

IV.
a few solutions...



COVID chatbot: https://clevy.io/



The [redacted] is becoming an enemy—a national enemy, in that it is sapping the minds of our youths of all that is manliest and noblest. It is more subtle because it attacks the individual and the family, it is more seductive because its sensations are pleasant, and its effects temporarily concealed. But it is there. It is working. It is developing. Its latest developments are everywhere visible. Some of its effects, too, every now and again, reach the light of day.

A National Enemy. Thetford & Watton Times | February 10th, 1894. from https://pessimists.co/novel-archive/

V.
resources to connect with

Resources

Black in AI

Latinx in AI

Queer in AI

Women in Machine Learning

Jews in AI

Dis[ability] in AI

Black Girls Code

Data Science Africa

Algorithmic Justice League

Institute for Ethical AI and Machine Learning

ORCAA Risk

TWIML AI Podcast

Radical AI Podcast

Black in AI: https://blackinai.github.io/
LatinX in AI: https://www.latinxinai.org/
Queer in AI: https://sites.google.com/view/queer-in-ai/
WiML: https://wimlworkshop.org/
Diversity in AI brochure (including other AI groups)
https://media.neurips.cc/Conferences/NeurIPS2019/DiversityBrochure2019.pdf
Black Girls Code: https://www.blackgirlscode.com/
Data Science Africa: http://www.datascienceafrica.org/
Algorithmic Justice League https://www.ajlunited.org/
Institute for Ethical AI and Machine Learning: https://ethical.institute/
ORCAA Risk: https://orcaarisk.com/
TWIML AI Podcast: https://twimlai.com/
Radical AI Podcast: https://www.radicalai.org/

questions, comments, ideas?

ejardas@uwalumni.com

On Twitter @ejardas

43

Further academic reading. (Please email me if you have trouble accessing any of these!)

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. Ethics and Information Technology, 7(3), 149–155. https://doi.org/10.1007/s10676-006-0004-4

Barbosa, N. M., & Chen, M. (2019). Rehumanized Crowdsourcing. 1–12. https://doi.org/10.1145/3290605.3300773

Bostrom, N., & Yudkowsky, E. (n.d.). The ethics of artificial intelligence. In Cambridge Handbook of Artificial Intelligence. https://doi.org/10.1016/j.mpmed.2010.10.008

Buolamwini, J., & Gebru, T. (n.d.). Gender shades: intersectional accuracy disparities in commercial gender classification. https://doi.org/10.2147/OTT.S126905

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care ' addressing ethical challenges. New England Journal of Medicine, 378(11), 981–983. https://doi.org/10.1056/NEJMp1714229

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017). Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the United States. Proceedings of the National Academy of Sciences of the United States of America, 114(50), 13108–13113. https://doi.org/10.1073/pnas.1700035114

HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1, 609–614.

Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. Proceedings of the 52nd Hawaii International Conference on System Sciences, 2122–2131. https://doi.org/10.24251/hicss.2019.258

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines, 30(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 501–512. https://doi.org/10.1145/3351095.3372826

Hu, L., & Chen, Y. (2018). A short-term intervention for long-term fairness in the labor market. The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018, 2, 1389–1398. https://doi.org/10.1145/3178876.3186044

Jo, E. S., & Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 306–316. https://doi.org/10.1145/3351095.3372829

Linthicum, K. P., Schafer, K. M., & Ribeiro, J. D. (2019). Machine learning in suicide science: Applications and ethics. Behavioral Sciences and the Law, 37(3), 214–222. https://doi.org/10.1002/bsl.2392

Manrai, A. K., Patel, C. J., & Ioannidis, J. P. A. (2018). In the era of Precision medicine and big data, Who is normal? JAMA - Journal of the American Medical Association, 319(19), 1981–1982. https://doi.org/10.1001/jama.2018.2009

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science (New York, N.Y.), 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., Dewar, N., & Beard, N. (2019). Integrating ethics within machine learning courses. ACM Transactions on Computing Education, 19(4). https://doi.org/10.1145/3341164

van Wynsberghe, A., & Robbins, S. (2019). Critiquing the Reasons for Making Artificial Moral Agents. Science and Engineering Ethics, 25(3), 719–735.

S., Jonas, A., McAllister, K. S. L., Myles, P., Granger, D., Birse, M., Branson, R., Moons, K. G. M., Collins, G. S., Ioannidis, J. P. A., Holmes, C., & Hemingway, H. (2020). Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. The BMJ, 368, 1–12. https://doi.org/10.1136/bmj.l6927

Vyas, D., Eisenstein, L. G., & Jones, D. S. (2020). Hidden in Plain Sight: Reconsidering the Use of Race Correction in Clinical Algorithms. New England Journal of Medicinen, 1–9.

Yapo, A., & Weiss, J. (2018). Ethical Implications of Bias in Machine Learning. Proceedings of the 51st Hawaii International Conference on System Sciences, 9, 5365–5372. https://doi.org/10.24251/hicss.2018.668