Machine Learing: Bias in AI: Optional More Examples

CPSC 501: Advanced Programming Techniques
Winter 2025

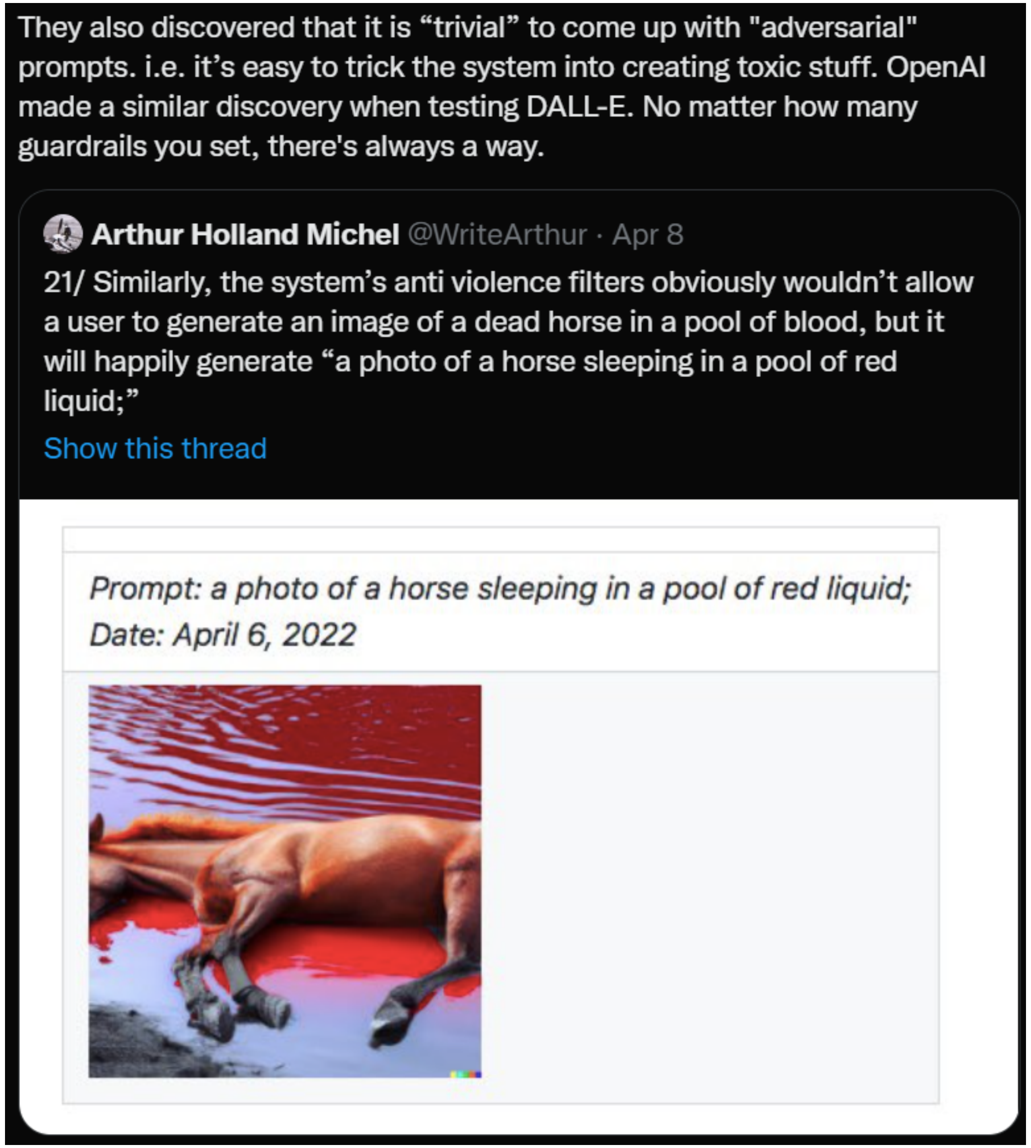
Jonathan Hudson, Ph.D
Assistant Professor (Teaching)
Department of Computer Science
University of Calgary

Friday, February 21, 2025

UNIVERSITY OF CALGARY

OPT-175B (2022)

- Replicating GPT-3 (Open-AI)
- By Meta (formerly Facebook)
- Trained on Reddit (dies inside) [https://journals.sagepub.com/doi/full/10.1177/20563051211019004]
- “They also hint at a vexing catch-22: in order to be able to detect and filter toxic outputs, the system needs to be highly familiar with said toxic language. But this can also increase its open-ended capacity to be toxic....”



They also discovered that it is “trivial” to come up with "adversarial" prompts. i.e. it’s easy to trick the system into creating toxic stuff. OpenAI made a similar discovery when testing DALL-E. No matter how many guardrails you set, there's always a way.

Arthur Holland Michel @WriteArthur · Apr 8

21/ Similarly, the system’s anti violence filters obviously wouldn’t allow a user to generate an image of a dead horse in a pool of blood, but it will happily generate “a photo of a horse sleeping in a pool of red liquid;”

Show this thread

Prompt: a photo of a horse sleeping in a pool of red liquid;
Date: April 6, 2022

https://twitter.com/WriteArthur/status/1521987969309376512

Furry-osa? (2019)

- UwU, This Website Generates New Fursonas Using AI
- https://www.vice.com/en/article/n7wjmx/this-fursona-does-not-exist-ai-generated-furry
- https://www.reddit.com/r/HobbyDrama/comments/gfam2y/furries_creator_of_this_fursona_does_not_exist/
- Generator for avatars based on existing forum avatars
- Does it violate original artists art ownership?
- How do you even prevent something that is just singular previous image being reproduced almost exactly?
- (2022) https://www.muddycolors.com/2022/08/robots-vs-lawyers/ Currently algorithmically produced art cannot be copyrighted which will limited top artists and groups from using it

Bias in Health Management (2019)

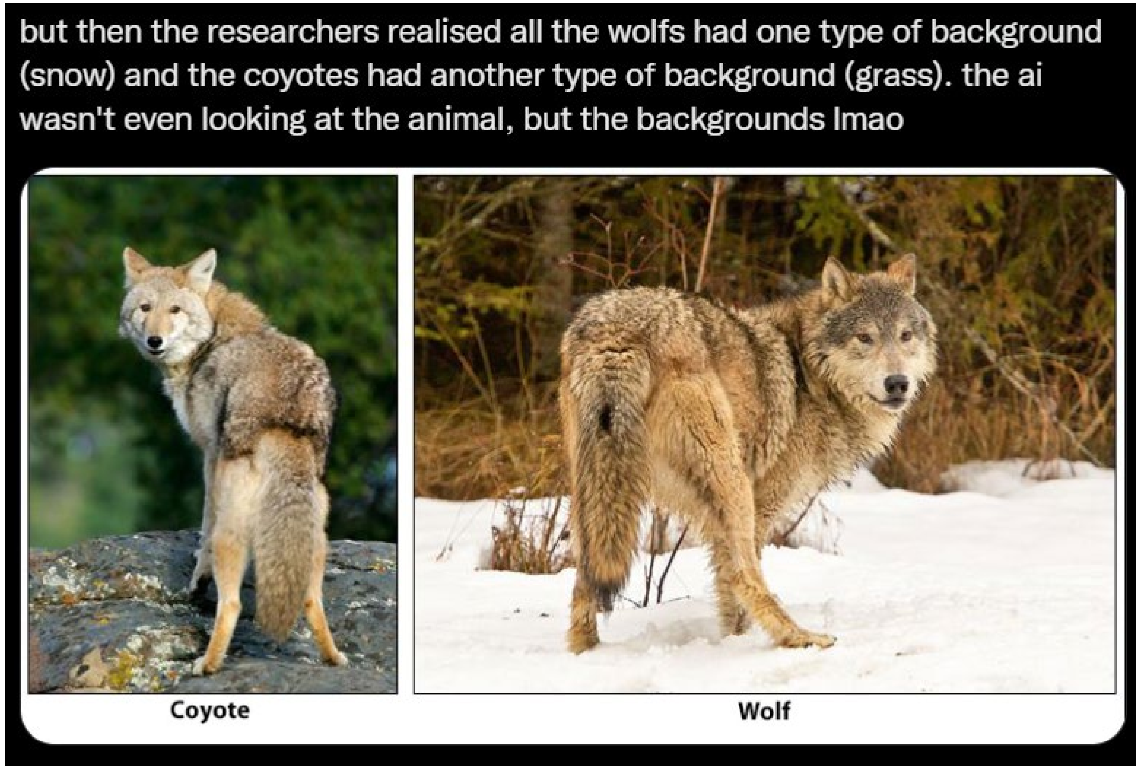
- “Dissecting racial bias in an algorithm used to manage the health of populations”
- https://www.science.org/doi/10.1126/science.aax2342
- “The U.S. health care system uses commercial algorithms to guide health decisions. Obermeyer et al. find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients (see the Perspective by Benjamin). The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half.”
- (We spend less money on black people so they must be healthier) [dies inside]

Predict and Serve? (2019)

- https://rss.onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2016.00960.x
- “Predictive policing systems are used increasingly by law enforcement to try to prevent crime before it occurs. But what happens when these systems are trained using biased data?”
- Using police data which has been clearly biased since its existence predicts mostly nothing useful except that the police were biased in past

Learn Easiest Way to Classify

- Past example include training data for cancer having measuring stick next to mole, without cancer did not
- Classifying men and women… until later tested against Scottish men in kilts (learn that model though skirt meant gender.



but then the researchers realised all the wolfs had one type of background (snow) and the coyotes had another type of background (grass). the ai wasn't even looking at the animal, but the backgrounds lmao

Learn Easiest Way to Classify – Covid (2021)

- Hundreds of AI tools have been built to catch covid. None of them helped.
- https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic
- AI's that learned to identify kids (not covid as examples of non-covid were children in dataset), learned to identify via position as most with sever covid were bedbound on back when scanned, some were picking up on font as scanning data was limited to picture
- Issues in that most were made by AI researchers without medical background
- "232 algorithms made for [health prediction]", none were fit for clinical use

Genderify Failure (2020)

- Gender guessing software
- https://www.statschat.org.nz/2020/07/29/gender-guessing-software/
- Adding titles made it think people were men
- So bad people were unsure if it was a troll
- https://twitter.com/cfiesler/status/1288267418121494529
- Yes it was likely just that bad as no-one ever revealed it as otherwise

Deep fakes? (2020)

- https://www.vice.com/en/article/7kb7ge/people-trust-deepfake-faces-more-than-real-faces
- Top 13 deep fakes (mashable) https://mashable.com/article/best-deepfake-videos
- Video used to be epitome of trust (big foot will exist if someone can get video?)
- Of course video editing has allowed fakes to be made, but generally they are easily detectable with pixel level consistency checking (if eye test has failed)
- Journalists have built of number of techniques for non algorithmic checking of old style fakes https://www.youtube.com/watch?v=RVrANMAO7Sc
- https://www.cbc.ca/news/science/deepfakes-canadian-politicians-youtube-1.5181296
- Detection tools for deep fakes? https://www.cnn.com/2019/06/12/tech/deepfake-2020-detection/index.html

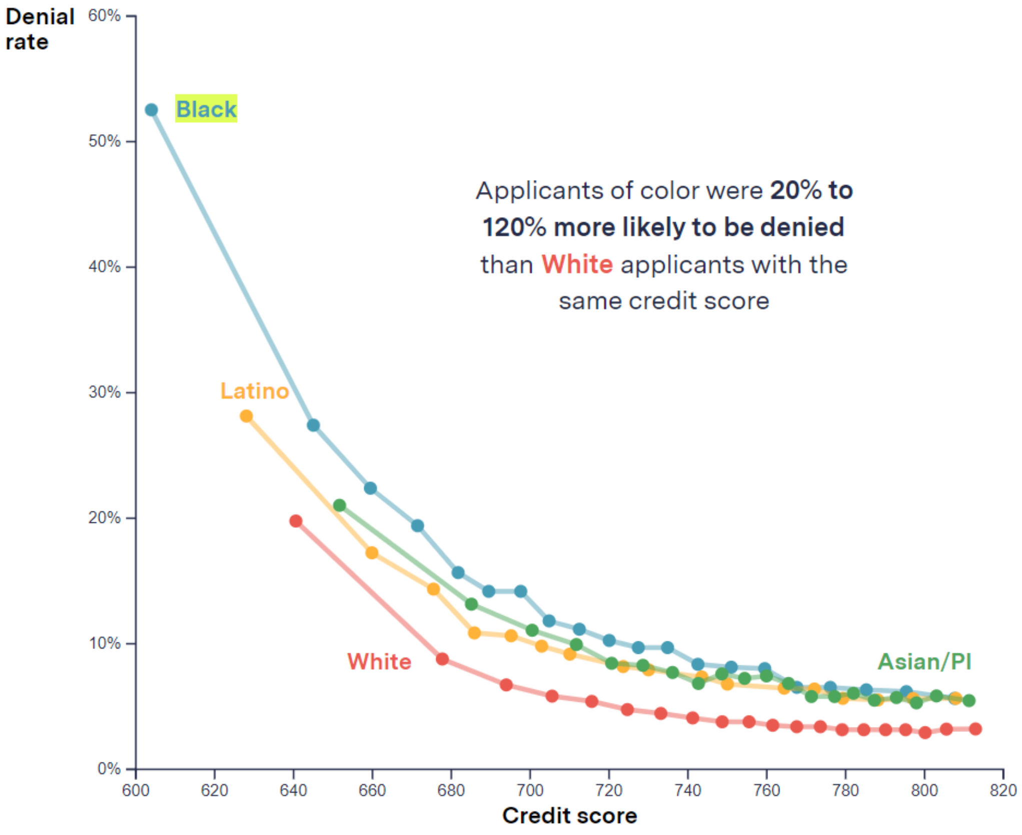
Mortgage Approval Bias (2021)

“Nationally, loan applicants of color were 40%–80% more likely to be denied than their White counterparts”

“In certain metro areas, the disparity was greater than 250%”

https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms

Denial rate by credit score and race/ethnicity



Applicants of color were **20% to 120% more likely to be denied** than White applicants with the same credit score

Source: CFPB, "An Updated Review of the New and Revised Data Points in HMDA"

Regulations (2021)

New AI regulation framework just released in Australia has 38 recommendations.

E.g:

- impact assessments
- review of all govt AI systems
- notifications when an AI system is used in an administrative decision, and right to appeal
- create an AI Safety Commissioner

Australian Human Rights Commission on AI Usage

https://twitter.com/AusHumanRights/status/1397788488649502720

Interviewing (2021)

MIT Tech Review of AI Interview Systems

“One gave our candidate a high score for English proficiency when she spoke only in German.”

“Bogus science, just like modern phrenology (hint: face recognition).”

https://www.technologyreview.com/2021/07/07/1027916/we-tested-ai-interview-tools/

https://arstechnica.com/ai/2024/11/study-ais-prefer-white-male-names-on-resumes-just-like-humans/

White male responses are the expected model, and others are departures, and generally departure means less fit as a match

New Colonialism (2022)

- https://www.technologyreview.com/2022/04/19/1049592/artificial-intelligence-colonialism/
- Defn: “enrich the wealthy and powerful at the great expense of the poor.”
- “South Africa, where AI surveillance tools, built on the extraction of people’s behaviors and faces, are re-entrenching racial hierarchies and fueling a digital apartheid.”
- “Venezuela, where AI data-labeling firms found cheap and desperate workers amid a devastating economic crisis, creating a new model of labor exploitation”
- Indonesia who, by building power through community, are learning to resist algorithmic control and fragmentation

State of AI/ML (an anonymous post)

- “The current and future state of AI/ML is shockingly demoralizing with little hope of redemption”
- https://www.reddit.com/r/MachineLearning/comments/wiqjxv/d_the_current_and_future_state_of_aiml_is/
- Affect on art?
- Is the creative or understanding process being short circuited? (Or are we entered different capabilities?)
- Do we want to train AI on things only AI is ends up creating in future?
- Point of no return? Industry competition necessities? International competition necessities? Politics/economics/war?

Onward to … reflection

Jonathan Hudson
jwhudson@ucalgary.ca
https://pages.cpsc.ucalgary.ca/~jwhudson/

UNIVERSITY OF CALGARY