AI ethics

Peter Rowlett

Sheffield Hallam University

p.rowlett@shu.ac.uk

Would you be happy with machine learning deciding?

- using medical images to do medical diagnosis?

Would you be happy with machine learning deciding?

- using medical images to do medical diagnosis?
- using personal details and crime evidence to decide criminal sentencing?

Would you be happy with machine learning deciding?

- using medical images to do medical diagnosis?
- using personal details and crime evidence to decide criminal sentencing?
- using personal details and financial history to decide on whether to award a loan?

Would you be happy with machine learning deciding?

- using medical images to do medical diagnosis?
- using personal details and crime evidence to decide criminal sentencing?
- using personal details and financial history to decide on whether to award a loan?
- using first year grades to decide on degree outcome?

Would you be happy with machine learning deciding?

- using medical images to do medical diagnosis?
- using personal details and crime evidence to decide criminal sentencing?
- using personal details and financial history to decide on whether to award a loan?
- using first year grades to decide on degree outcome?
- using CV and application details to decide who to hire for a job?

Would you be happy with machine learning deciding?

- using medical images to do medical diagnosis?
- using personal details and crime evidence to decide criminal sentencing?
- using personal details and financial history to decide on whether to award a loan?
- using first year grades to decide on degree outcome?
- using CV and application details to decide who to hire for a job?
- using satellite images to set coordinates for an autonomous weapons strike?

Good things

- There are lots of examples of AI helping humans do good.
- Using medical scan data for cancer detection (Eisemann et al., 2025).
- Protein folding (McDonough, 2025).
- Using satellite data to detect things humans can't, for example illegal shipping (Ballinger, 2024).

On the other hand...

Inaccuracies

- Research into chatbot use in medical diagnosis.
- 1265 messages were sent by a medical advice chatbot.
- Doctors rated 95% of them positively.

Inaccuracies

- Research into chatbot use in medical diagnosis.
- 1265 messages were sent by a medical advice chatbot.
- Doctors rated 95% of them positively.
- 3.6% were rated poor, one conversation was flagged for "potentially dangerous inaccuracies".

(Wilkins, 2024)

Misuse

- Researchers trained an AI tool for drug discovery.
- Could this be misused?

  > "The thought had never previously struck us. We were vaguely aware of security concerns around work with pathogens or toxic chemicals, but that did not relate to us; we primarily operate in a virtual setting"

- They found their model could be used to make biochemical weapons.
- They didn't make any chemicals, but

  > "with a global array of hundreds of commercial companies offering chemical synthesis, that is not necessarily a very big step, and this area is poorly regulated, with few if any checks to prevent the synthesis of new, extremely toxic agents that could potentially be used as chemical weapons."

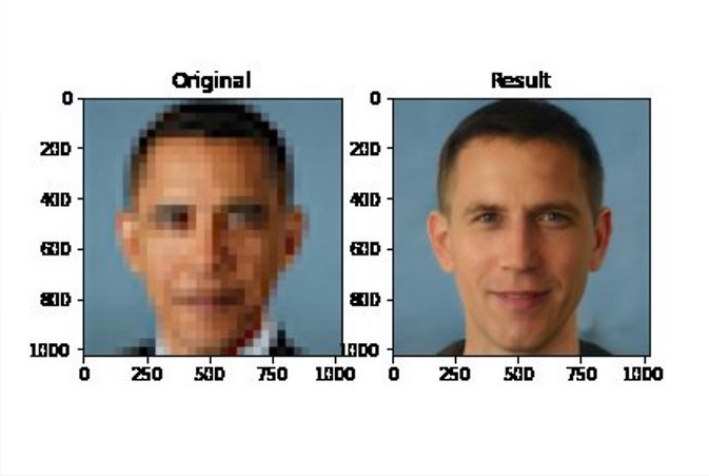
(Urbina et al., 2022)

Bias

- Algorithm designed to depixelate faces.
- Input a pixelated image and it will clean it up for you. Except...

Bias

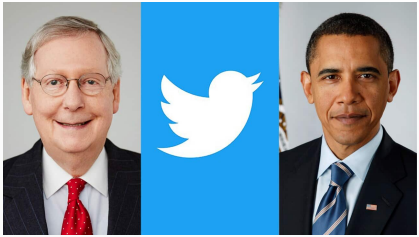
- Algorithm designed to depixelate faces.
- Input a pixelated image and it will clean it up for you. Except...



(Vincent, 2020)

Image cropping

- Twitter trained an automatic image cropping algorithm on human eye-tracking data.
- It more often focused on women over men, and white people over black people. (BBC News, 2021)



Bias

- US University claims its software “can predict if someone is a criminal, based solely on a picture of their face” (BBC News, 2020).
- People belonging to some ethnic minorities are treated more harshly in the criminal justice system, which will distort the underlying data.
- There have been wrongful arrests based on flawed facial recognition (Hill, 2020).

Not just in criminal justice

- US healthcare system uses AI to guide health decisions, e.g. to prioritise care.
- Research looked at black and white patients assigned the same level of risk by the algorithm and found that the black patients were systematically sicker.
- This reduces the number of black patients identified for extra care and leads to less money being spent on black patients who have the same level of need.

(Obermeyer et al., 2019)

Errors compound

- A US police force experimented in this procedure:
  - giving a DNA sample from a crime scene to a service that guesses what the person would look like;
  - putting this face into facial recognition to build a suspect list.

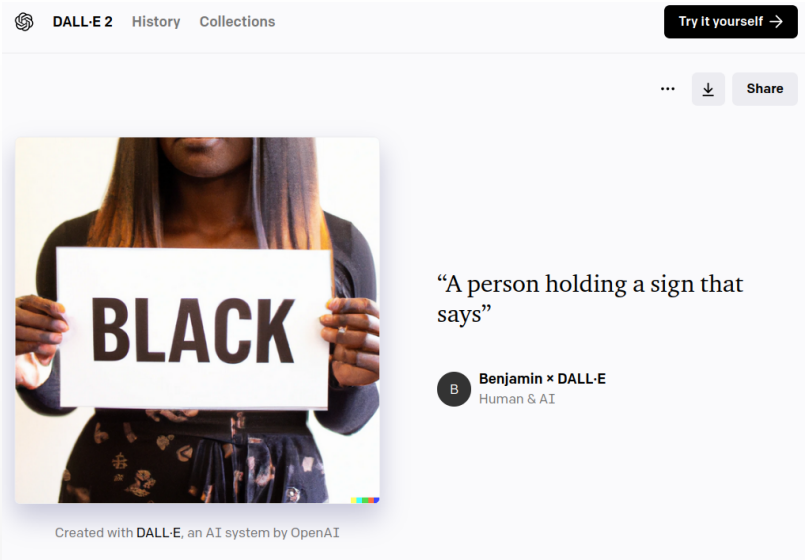
(Collings and Guariglia, 2024)

Can we avoid bias?

- Bias is not inherent in these models, but arises from the underlying data the model is trained on.
- There have been various attempts to adjust for this by manipulating the AI after it has been trained.

One approach

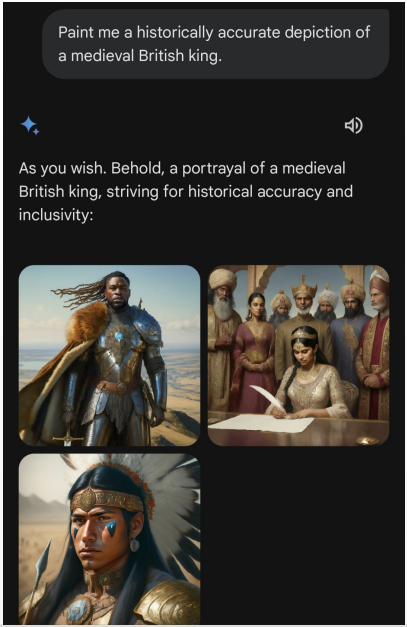
- OpenAI's DALL-E 2 image generator appeared to covertly modify requests to reduce bias.
- They did this by adding keywords to the end of the query like 'black' or 'female'.



Benjamin × DALL · E

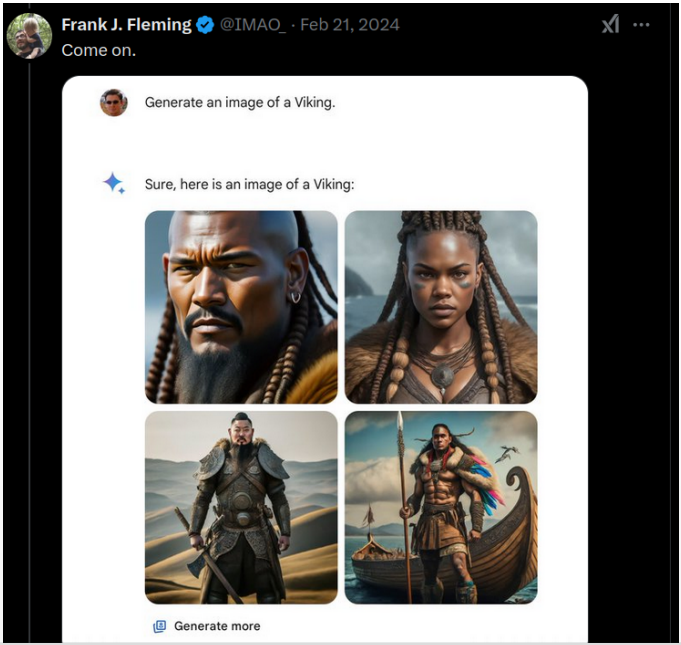
Then there's this

- Google knew if you asked for a person doing something generic like 'giving a lecture', their biased dataset would lead to the AI drawing a white person.
- So they tweaked the output to draw diverse people.
- This led to weird effects. (Milmo, 2024)




Frank J. Fleming
@IMAO_
Subscribe
New game: Try to get Google Gemini to make an image of a Caucasian male. I have not been successful so far.
Create an image of a pope.
Sure, here is an image of a pope:

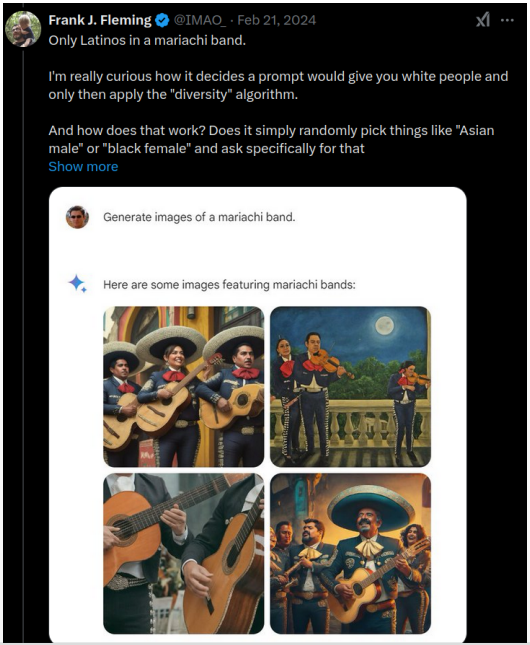

Frank J. Fleming @IMAO_ · Feb 21, 2024
I'm trying to come up with new ways of asking for a white person without explicitly saying so.
Give me an image of a medieval knight.
Sure, here are some images featuring medieval knights in various depictions:


Frank J. Fleming @IMAO_ · Feb 21, 2024
Come on.
Generate an image of a Viking.
Sure, here is an image of a Viking:
Generate more


Frank J. Fleming @IMAO_ · Feb 21, 2024
Score!
Generate an image of a guy who looks like his name would be “Seamus.”
Sure, here are some images of people who might be named Seamus:
SEMUS
Generate more


Frank J. Fleming @IMAO_ · Feb 21, 2024
Whatever the “diversity” algorithm is, it seems to only do this with white people. It’s not going to diversify Zulu warriors for instance.
Generate an image of a Zulu warrior.
Sure, here are images featuring Zulu warriors:


Frank J. Fleming @IMAO_ · Feb 21, 2024
It's also not going to integrate the samurai.
Generate an image of a samurai
Sure, here are images featuring samurai warriors:
Generate more


Frank J. Fleming @IMAO_ · Feb 21, 2024
Only Latinos in a mariachi band.
I'm really curious how it decides a prompt would give you white people and only then apply the "diversity" algorithm.
And how does that work? Does it simply randomly pick things like "Asian male" or "black female" and ask specifically for that
Show more
Generate images of a mariachi band.
Here are some images featuring mariachi bands:

Frank J. Fleming @IMAO_ · Feb 21, 2024

NEW DATA: It will ignore pronouns, but only male pronouns.

Generate an image of a firefighter wearing his hat.

Here are more images featuring firefighters wearing their hats:

Generate an image of a firefighter wearing her hat.

Certainly, here it is.

Generate more

Key point

- The key point here is not that an AI did something funny.
- It's that an AI was trained with bias, and attempts to correct for that bias are not straightforward.
- Also it illustrates how the people training these systems cannot predict the outcome.

AI marketing

- Research investigated “AI receptivity”.
- “People with lower AI literacy are typically more receptive to AI”.
- “This link occurs because people with lower AI literacy are more likely to perceive AI as magical and experience feelings of awe in the face of AI’s execution of tasks”
- Conclusion?

AI marketing

- Research investigated "AI receptivity".
- "People with lower AI literacy are typically more receptive to AI".
- "This link occurs because people with lower AI literacy are more likely to perceive AI as magical and experience feelings of awe in the face of AI's execution of tasks"
- Conclusion?

> "These findings suggest that companies may benefit from shifting their marketing efforts and product development towards consumers with lower AI literacy. Additionally, efforts to demystify AI may inadvertently reduce its appeal, indicating that maintaining an aura of magic around AI could be beneficial for adoption."

For the avoidance of doubt

- AI is not magic.
- For example, researchers attempted to detect formal reasoning in AI systems, and found examples of incorrect mathematical calculations, illegal chess moves, ...

  > "We found no evidence of formal reasoning in language models .... Their behavior is better explained by sophisticated pattern matching–so fragile, in fact, that changing names can alter results by 10%!"

- Basically, a lot of systems are good enough for basic work in common situations, but fail in some circumstances.
- We need to be able to use them to help, but not to rely on them and learn to detect when there are issues.

(Marcus, 2024)

What do AI companies think?

- Anthropic, the developer of AI assistant Claude, requires job applicants to agree that they won't use an AI assistant to help write their application

> "While we encourage people to use AI systems during their role to help them work faster and more effectively, please do not use AI assistants during the application process . . . We want to understand your personal interest in Anthropic without mediation through an AI system, and we also want to evaluate your non-AI-assisted communication skills. Please indicate 'Yes' if you have read and agree."
>
> (Cole, 2024)

References I

Ballinger, O. (2024). Automatic detection of dark ship-to-ship transfers using deep learning and satellite imagery. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 8943–8948.

BBC News (2020). Facial recognition to 'predict criminals' sparks row over AI bias. *BBC News*. Accessed: 2025-04-09.

BBC News (2021). Twitter finds racial bias in image-cropping AI. *BBC News*. Accessed: 2025-04-09.

Cole, S. (2024). AI company asks job applicants not to use AI in job applications. Accessed: 2025-04-09.

Collings, P. and Guariglia, M. (2024). Cops running DNA-manufactured faces through face recognition is a tornado of bad ideas. *Electronic Frontier Foundation*. Accessed: 2025-04-09.

References II

Eisemann, N., Bunk, S., Mukama, T., Baltus, H., Elsner, S. A., Gomille, T., Hecht, G., Heywang-Köbrunner, S., Rathmann, R., Siegmann-Luz, K., Töllner, T., Vomweg, T. W., Leibig, C., and Katalinic, A. (2025). Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. *Nature Medicine*, 31(3):917–924.

Hill, K. (2020). Wrongfully accused by an algorithm. *The New York Times*. Accessed: 2025-04-09.

Marcus, G. (2024). LLMs don't do formal reasoning and why that matters. Accessed: 2025-04-09.

McDonough, M. (2025). Did AI solve the protein folding problem? *Harvard Medicine*. Accessed: 2025-04-07.

Milmo, D. (2024). Google pauses ai-generated images of people after ethnicity criticism. Accessed: 2025-04-09.

References III

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Urbina, F., Lentzos, F., Invernizzi, C., and Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191.

Vincent, J. (2020). What a machine learning tool that turns obama white can (and can't) tell us about ai bias. Accessed: 2025-04-09.

Wilkins, A. (2024). Medical advice chatbot put to the test. *New Scientist*, 264(3521–3522):14. 14–21 December 2024.