Algorithmic Bias, Accessibility & Equity

COMP90087 Week 8

Dr Simon D'Alfonso

School of Computing and Information Systems

The University of Melbourne

THE UNIVERSITY OF MELBOURNE

WARNING

This material has been reproduced and communicated to you by or on behalf of the University of Melbourne in accordance with section 113P of the *Copyright Act 1968* (**Act**).

The material in this communication may be subject to copyright under the Act.

Any further reproduction or communication of this material by you may be the subject of copyright protection under the Act.

Do not remove this notice

Learning outcomes

- Define the concept of accessibility and universal usability in computing (especially in HCI and related fields) and understand how it is promoted by computing best practices as well as by law.
- Define the concept of equity, in relation to a machine's idea of purported fairness, and the problem with algorithmic bias.
- Understand how complex systems can sometimes neglect accessibility and equity in their design process - even though on the surface they seem 'neutral' - and ways to mitigate this.
- Learn about the conflicting technical definitions of fairness as well as ideas on how to ameliorate issues in the design process.

Related readings in LMS

Readings: Algorithmic Bias, Accessibility, and Equity

1. Detecting and mitigating bias in natural language processing
2. Identifying Gender Bias in Generative Models for Mental Health Synthetic Data (No need to digest the technical details in this paper)

Additional readings.

If you find this module interesting, here are some extra readings that might interest you.

Ethical Implications of AI Bias as a Result of Workforce Gender Imbalance.
This research report is a result of an interdisciplinary collaboration between University of Melbourne and UniBank in uncovering sources of bias -- both human and algorithmic -- to consider when deploying any form of automated system in recruitment/shortlisting of job candidates.
Read only pp. 5-34 inclusive - the appendices are optional!
Link: https://www.unibank.com.au/-/media/unibank/about-us/member-news/report-ai-bias-as-a-result-of-workforce-gender-imbalance.ashx

Lecture outline

- About accessibility.
- About equity.
- AI bias:
  - Case studies
  - ChatGPT, Generative AI demo
  - Mitigating bias
  - Reflections on bias

About Accessibility

What is accessibility?

- "Basically, technology is accessible **if it can be used as effectively by people with disabilities as by those without**" (Thatcher, 2004).
- "Accessibility refers to the degree to which an interactive product **is accessible by as many people as possible**. A focus is on people with disabilities." (Preece, 2015)

Sources:

Thatcher, J. (2004) “Web Accessibility for Section 508”, http://www.jimthatcher.com/webcourse1.htm

Preece, J, Sharp, H, Rogers, Y. (2015). Interaction Design: Beyond Human-Computer Interaction. John Wiley & Sons.

Example

Besides accessibility for, say, wheelchair users, how else does this design feature promote universal usability?

Example example responses

- Crutches
- Prams
- Bikes
- Trolleys

Universal Usability

- Accessibility goes beyond just 'catering for those with disabilities'.
- Universal Usability = a "design for all" approach which is about making a product as accessible as possible to as wide a group of people as possible. The term originated from architecture (consider stairs vs. ramps/elevators/escalators).

Accessibility and the Law

Landmark case: Maguire versus Sydney Organising Committee for the Olympic Games (SOCOG), a legal case about making a website accessible to a visually impaired person.

- Maguire made a complaint to the human rights and equal opportunity commission (HREOC)… (SOCOG) had discriminated against him as a visually impaired person, in contravention of the Disability Discrimination Act 1992…“
- Main point: "failure to provide a website which was accessible to Maguire…“
- "SOCOG said that it did not discriminate unlawfully … cost and effort in retraining staff and redrawing entire development methods was an unjustifiable hardship in providing an accessible website…“
- Basically: SOCOG gave excuses (too much time needed etc); refuted by expert witnesses!
- "The Commissioner found that SOCOG had engaged in unlawful discrimination against Maguire in violation of Section 24 of the DDA 1992".
- SOCOG was stubborn; "The Commissioner found that SOCOG only partially complied and as a result, by section 103(1)(b)(iv) of the DDA, the commissioner awarded Maguire $20,000".

Usability and HCI

- Human-computer interaction (HCI) is a field where usability/accessibility was an early consideration.
- In HCI, usability refers to the ease with which users can interact with a system to achieve their goals effectively, efficiently, and satisfactorily.
- This involves evaluating elements such as:
  - the intuitiveness of the interface
  - the minimalism of user input required to achieve desired outcomes
  - the overall satisfaction and lack of frustration experienced by users
- Hardware and software

© University of Melbourne

Mobile Accessibility

- "Mobile accessibility" refers to making websites and applications more accessible to people with disabilities when they are using mobile phones and other devices: https://www.w3.org/WAI/standards-guidelines/mobile/
- The Web Content Accessibility Guidelines (WCAG) from the World Wide Web Consortium (W3C) provide a comprehensive set of criteria for accessible website/app content:
  - Perceive
  - Operable
  - Understandable
  - Robust

WCAG Guidelines

- **Perceivable:** Information and user interface components must be presented in ways that all users can perceive, regardless of their sensory abilities. This includes providing text alternatives for non-text content, creating content that can be presented in different ways (such as larger text sizes or simpler layouts), and ensuring that users can distinguish foreground from background.
- **Operable:** Users must be able to operate the interface, which means that the interface cannot require interaction that a user cannot perform. This includes making all functionality available from a keyboard (for those who cannot use a mouse), giving users enough time to read and use content, and not designing content in a way that is known to cause seizures.

WCAG Guidelines

- **Understandable:** Information and the operation of the user interface must be understandable. This means that text content should be readable and understandable, web pages should appear and operate in predictable ways, and users should be assisted in avoiding and correcting mistakes.
- **Robust:** Content must be robust enough that it can be interpreted reliably by a wide variety of user agents, including assistive technologies. This involves ensuring compatibility with current and future user tools.

Accessibility and AI systems

- **Designing Inclusive AI:** AI systems should be designed with inclusivity in mind, ensuring they can be used by people with a range of physical and cognitive disabilities. This includes providing:
  - accessible interfaces
  - alternative input methods (like voice control or eye-tracking)
  - outputs that consider various sensory impairments (like auditory or visual content alternatives).
- **Bias and Fairness**: Special attention needs to be given to training data and model design to ensure that AI systems do not discriminate against individuals with disabilities or other marginalized groups.
- AI can facilitate accessibility:
  - voice-activated devices and software that help individuals with disabilities perform tasks that might otherwise be difficult or impossible.
  - personalization and adaptability

About Equity

What is equity?

- the quality of being fair and impartial: *equity of treatment*.
- Law: a branch of law that developed alongside common law and is concerned with fairness and justice, formerly administered in special courts: *if there is any conflict between the principles of common law and equity, equity prevails*.

- Many other interrelated (similar) concepts such as fairness that have been encountered before.
- This module covers some similar terrain to Simon C’s Week 6 lecture (e.g. fairness) but takes a more 'applied' view from the perspective of the technology.

 © University of Melbourne

Definitions of Fairness

- As was mentioned in the lecture in Week 6, there are various mathematical definitions of fairness. For example:
  - **Statistical Parity (Demographic Parity):** Statistical parity is achieved if the probability of a positive outcome is the same across all groups defined by an attribute (like race or gender)
  - **Equal Opportunity:** This is a stronger notion of fairness that focuses on equal treatment of the "advantaged" outcome. Specifically, it suggests that all groups should have equal true positive rates.
  - **Individual Fairness:** This concept insists that similar individuals should receive similar treatment.
  - **Maximin or Rawlsian Fairness:** In allocation problems, this principle focuses on maximizing the utility of the worst-off individual or group. The goal is to raise the minimum outcome to be as high as possible, following philosopher John Rawls's principle of fairness.

Key Point

- Importantly, these different notions of fairness are known to be generally incompatible.
- **Impossibility Theorem:** mathematically impossible for an algorithm to simultaneously satisfy various popular fairness measures.

Fairness definition conflict

- A classic example where two definitions of fairness - Statistical Parity (Demographic Parity) and Equal Opportunity - can be at odds involves a hiring scenario where candidates from different groups are being selected for interviews based on test scores.
- Imagine a company that wants to ensure fairness in its hiring process for a technical role. Candidates from two groups, Group A and Group B, take a standardized test, and the company uses the test scores to decide whom to invite for interviews. The test scores are normally distributed but with different means for each group due to various historical or socio-economic factors:
  - Group A has a mean test score of 70 with a standard deviation of 10.
  - Group B has a mean test score of 60 with a standard deviation of 10.
- The company decides to interview candidates who score above a certain threshold. Suppose the company sets the threshold at 65, aiming for fairness in one of the two mentioned definitions.

Applying Statistical Parity

To achieve statistical parity, the company would need to ensure that the proportion of candidates selected from both groups is the same. If the threshold is set at 65, a larger proportion of Group A candidates will pass compared to Group B due to their higher mean score. To adjust for statistical parity, the company might lower the threshold for Group B or raise it for Group A, leading to different thresholds for different groups.

Applying Equal Opportunity

Equal opportunity focuses on the candidates who are truly qualified (e.g., all those who would perform well in the job, which might be assumed to correlate with having higher test scores). If the true positive rate (proportion of qualified candidates correctly identified) needs to be equal across groups, the company would maintain a single threshold that is the same for both groups, based on the scores directly, regardless of the proportion from each group that ends up being selected.

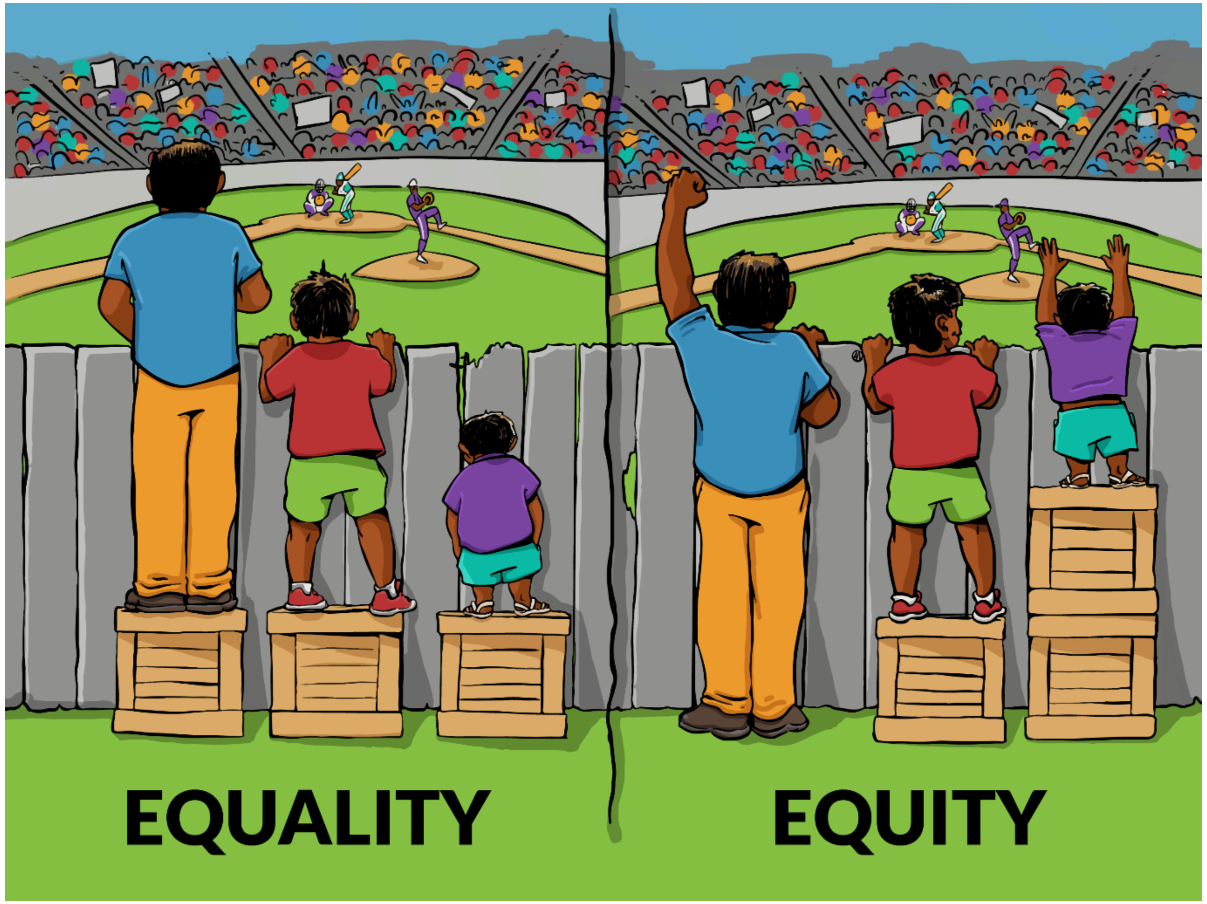
The Conflict

- **Statistical Parity** would require potentially different thresholds to ensure the same proportion from each group is selected, which could mean selecting less qualified candidates from one group over more qualified candidates from the other group to balance the numbers.
- **Equal Opportunity** would maintain the same threshold for all, focusing purely on the test score, which would likely result in a higher proportion of Group A candidates being selected over Group B, based on their higher average scores.

Thought Experiment: *EqualShareAlgorithm*

- Let's now turn to a simple example to reflect on. Suppose we develop a system that employs an algorithm to divide a finite pool of resources (X) among N members.
- One straightforward way to achieve this is by assigning each member a share of X/N.
- In what condition(s) does this algorithm become unfair?

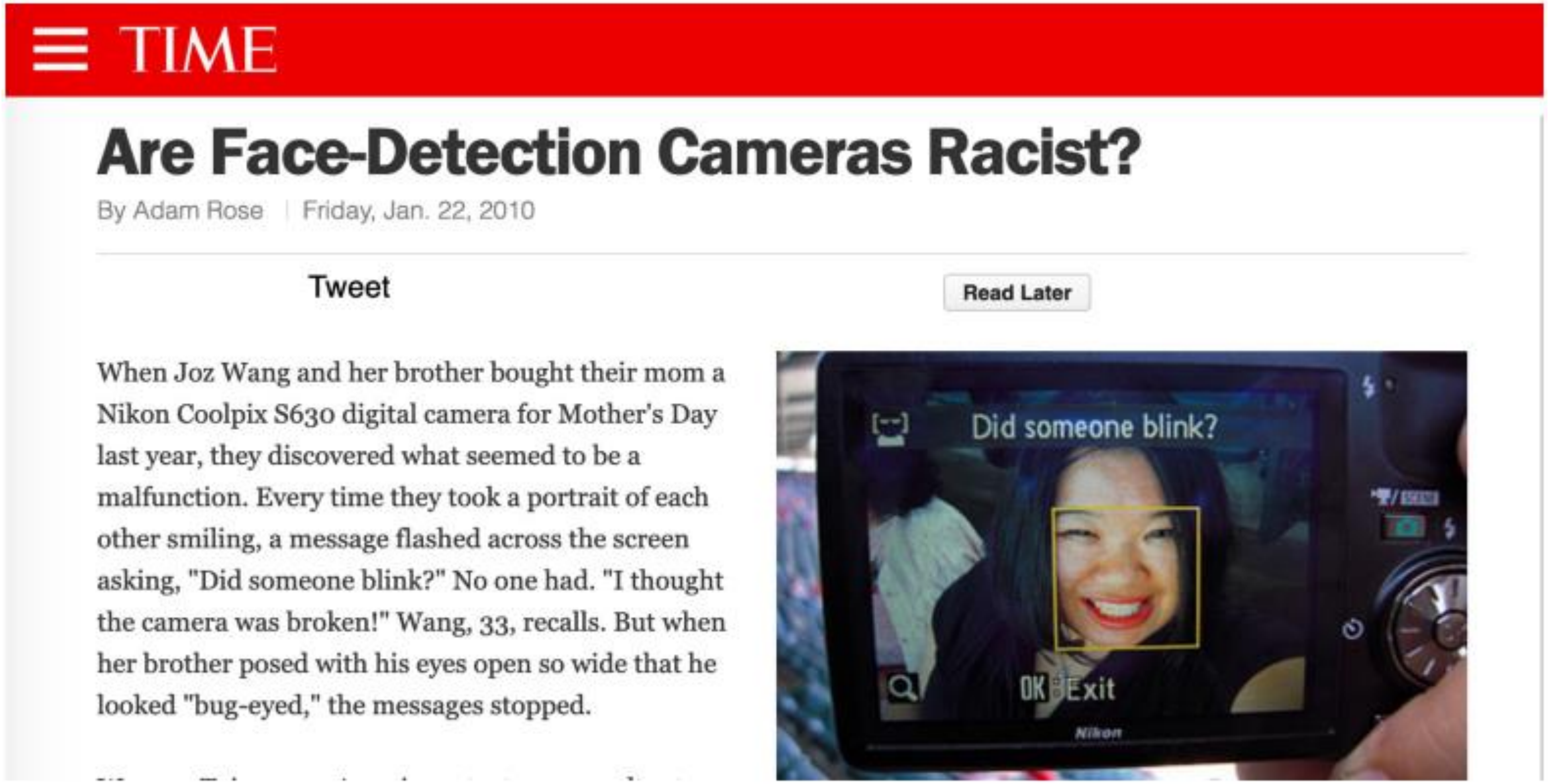
Equality versus Equity



Complexity, complex systems, and unintended consequences!

- The design of an automated / computerised / AI-driven system can seem fair *a priori*
- But it is only after a system has been deployed, or implemented in a certain context, that issues such as inequity (and inaccessibility) become apparent.
- Systems are inherently complex: what works in isolation does not necessarily work ‘as a whole’, or even when deployed in circumstances (external factors, e.g., social factors) we did not foresee.

Examples: Equity issues after deployment?



TIME

Are Face-Detection Cameras Racist?

By Adam Rose | Friday, Jan. 22, 2010

Tweet   Read Later

When Joz Wang and her brother bought their mom a Nikon Coolpix S630 digital camera for Mother's Day last year, they discovered what seemed to be a malfunction. Every time they took a portrait of each other smiling, a message flashed across the screen asking, "Did someone blink?" No one had. "I thought the camera was broken!" Wang, 33, recalls. But when her brother posed with his eyes open so wide that he looked "bug-eyed," the messages stopped.

Image source: Time Magazine - Rose (2010)

Another thought experiment

> " There is a new, fun, web app/game out there which helps you improve your handwriting (a long lost art!) and at the same time improve your handwriting speed. After all, handwritten cards and letters are art forms which have been displaced by technology.
>
> This new app, *RightHandWrite*, is designed to allow you to practice your handwriting in a 'gamified' contest environment. It does two things:
> - to measure the speed of one's writing, it encourages users to write out a passage of text as fast as possible.
> - at the same time, using machine learning technology (trained on models of many samples of handwriting), it also calculates your neatness score.
>
> The app 'gamifies' the experience by having a final score calculated by averaging the speed and neatness scores, and the top users every day will have a chance to win fancy fountain pens and other stationery! Also, the makers of the app decide to make the competition aspect as transparent as possible - by opening up the source code, auditing ML models, declaring all conflicts of interest, etc. "

- Alice has used the app for some time now and enjoys it. However, she recently had a sporting injury where she hurt her fingers severely: doctors advised her that the recovery takes several weeks. In these few weeks, she was not able to take part at all (or at severely reduced scores for both speed and neatness).
  - Here we find an accessibility issue.
- Elijah has very neat handwriting as he is a calligrapher and has practiced handwriting all his life! Unfortunately, based on his reading of recent audit reports to the app, he found out that the ML models were trained on standardised samples of handwriting, but for right-handers. (Elijah is left-handed). When he submits his work to be ranked by the app, the left-handed nature of his submissions causes them to have, on average, 30% less scores than right-handed samples.
  - Here we find an equity issue.

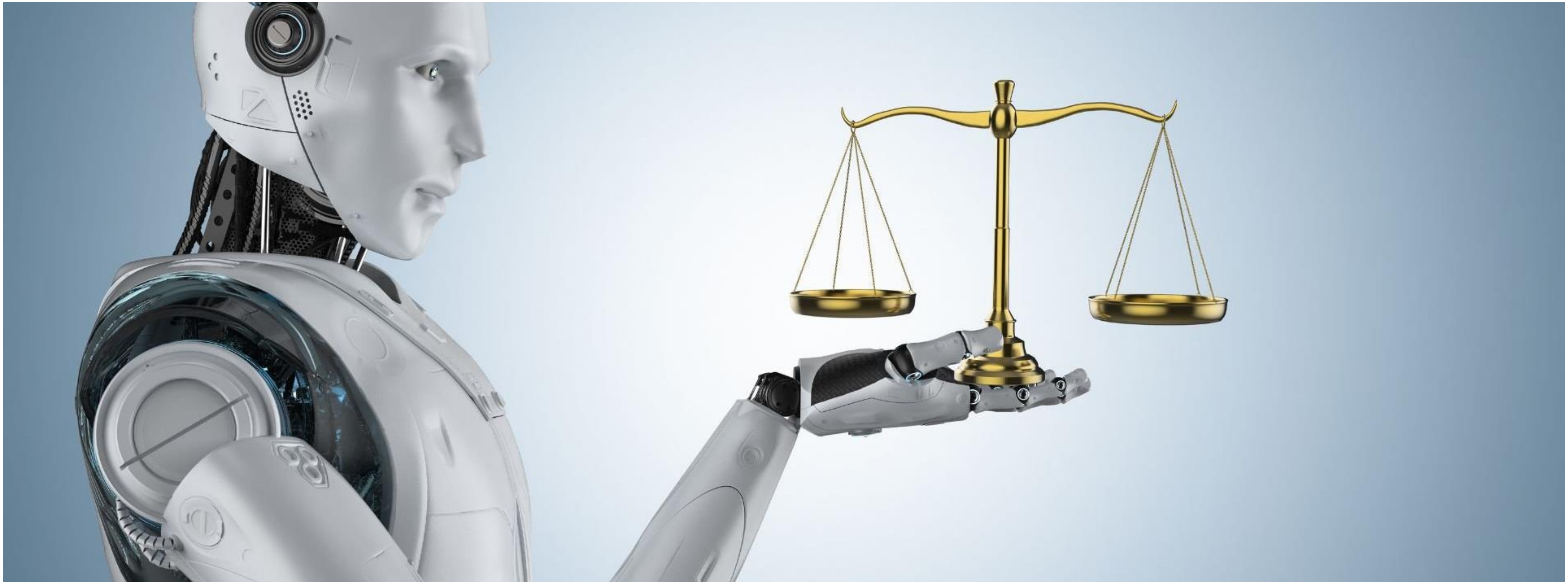
 © University of Melbourne

Activity

What are some potential accessibility or equity issues with the *RightHandWrite* app?

 © University of Melbourne

(Generative AI) and Biases



AI Bias

- Bias in AI systems refers to systematic and repeatable errors that create unfair outcomes, such as privileging or discriminating against one arbitrary group of users over others.
- Understanding AI bias is crucial for developing fair and equitable technology solutions.

Reading: https://www.brookings.edu/articles/detecting-and-mitigating-bias-in-natural-language-processing/

Executive Summary

Unsupervised artificial intelligence (AI) models that automatically discover hidden patterns in natural language datasets capture linguistic regularities that reflect human biases, such as racism, sexism, and ableism.[1] These unsupervised AI models, namely word embeddings, provide the foundational, general-purpose, numeric representation of language for machines to process textual data.

Types of Bias in AI

- **Data-Driven Bias:** Occurs when the training datasets are not representative of the broader population.
- **Algorithmic Bias:** Arises from assumptions and simplifications in the AI algorithms that skew outputs (e.g. problematic feature selection). Or can arise from errors in the algorithm.
- **Prejudice Bias:** This form of bias can manifest in AI systems when the training data, the design of the algorithm, or the operational environment has inherent or societal prejudices. For example, training dataset might contain representative data, but that data is labelled in a prejudiced way.

Examples of AI Bias

- **Recruitment Tools:** AI systems trained primarily on resumes of past successful applicants who are predominantly from one gender, inadvertently learn to favour candidates of that gender.
- **Facial Recognition:** Systems that fail to accurately identify individuals from certain racial or ethnic backgrounds.
- **Credit Scoring Algorithms:** AI that may lower scores based on zip codes or factors correlated with race.

Causes of AI Bias

- **Historical Data:** Use of historical data that contains past prejudices and inequalities.
- **Lack of Diversity:** AI development teams lacking diversity can inadvertently encode their biases.
- **Model Complexity/Opacity:** Complex/opaque models can obscure biases, making them harder to detect and correct.

Impacts of AI Bias

- **Social Injustice:** Discriminatory practices reinforced in vital areas such as employment, law enforcement, and lending.
- **Economic Disparities:** Exacerbation of existing economic inequalities through biased automated decisions.
- **Loss of Trust:** Erosion of public confidence in AI technologies and their applications.

Detecting and mitigating bias in natural language processing

https://www.brookings.edu/articles/detecting-and-mitigating-bias-in-natural-language-processing/. Some summary points:

- Unsupervised AI models capture inherent human biases in natural language datasets.
- Biases in word embeddings propagate to various NLP applications affecting decision-making in areas like employment and education.
- The lack of regulation and oversight in AI development exacerbates bias propagation and its societal impacts.
- Specific instances of AI bias include Amazon's sexist job candidate screening tool and biased gender translations in Google's services.
- Effective bias mitigation requires regulatory frameworks, diverse development teams, and transparency in AI model development and deployment.
- The power and influence of technology companies in AI pose challenges to equity and fairness.

Reading: Cabrera-Lozoya et al (2023)

Identifying Gender Bias in Generative Models for Mental Health Synthetic Data

Daniel Cabrera Lozoya
School of Computing and Information Systems
University of Melbourne
Melbourne, Australia
dcabreralozo@student.unimelb.edu.au

Simon D'Alfonso
School of Computing and Information Systems
University of Melbourne
Melbourne, Australia
dalfonso@unimelb.edu.au

Mike Conway
School of Computing and Information Systems
University of Melbourne
Melbourne, Australia
mike.conway@unimelb.edu.au

***Abstract*— Natural language generation (NLG) systems have proven to be effective tools to create domain-specific synthetic data. The mental health research field could benefit from data augmentation techniques, given the challenges associated with obtaining and utilizing protected health information. Yet, NLG systems are often trained using datasets that are biased with respect to key demographic factors such as ethnicity, religion, and gender. This can perpetuate and propagate systematic human biases that exist and ultimately lead to inequitable treatment for marginalized groups. In this research we studied and characterized biases present in the Generative Pre-trained Transformer 3 (GPT-3), which is an autoregressive language model that produces human-like text. The prompts used to generate text via GPT-3 were based on the Brief Cognitive Behavioral Therapy framework, and each prompt also specified to write the answer as a female or male patient. By controlling the sex distributions within our prompts, we observed the impact of each trait in the generated text. The synthetic data was analysed using the Linguistic Inquiry and Word Count software (LIWC-22) and ccLDA for cross-collection topic modeling. LIWC-22 results show that stereotypical competence features such as money, work, and cognition are more present in the male's synthetic text, whereas warmth features such as home, feeling, and emotion are highly present in female's generated data. The ccLDA results also associate competence features with males and warmth features with females.**

Keywords* — *Generative Models, Natural Language Processing, Mental Health, Bias, Fairness in AI

I. Introduction

Generative models designed to model real data

depicted as caring and emotional [8]. These preconceptions enforce stereotypes such as women being portrayed more in a domestic setting, whereas men are associated more to the workplace [9]. NLG models that strongly manifest these types of biases would likely create synthetic therapy transcripts that neglect mental health issues concerning women in the workforce or househusbands. The lack of emphasis on women's mental health problems in the workplace has been extensively documented in the literature [10], and there is an imperative that new technology should be debiased to avoid perpetuating this issue.

The principle of fairness through awareness [11] states that to debias a model, we must first identify its biases. Since GPT-3 generates text by expanding on user-given prompts, we characterized the bias within the model by evaluating synthetic data from prompts that included different gender traits. To evaluate the data from each group we used LIWC-22, a text analysis software tool designed to assess various psychosocial constructs (e.g., social behavior, cognitive process, and power) within a document [12]. A cross collection topic modelling procedure using ccLDA [13] was done to analyze the text from each group and uncover underlying semantic structures that perpetuate stereotypes. Also, ccLDA allows us to study the similarities and differences across the text from each group.

II. Previous Work

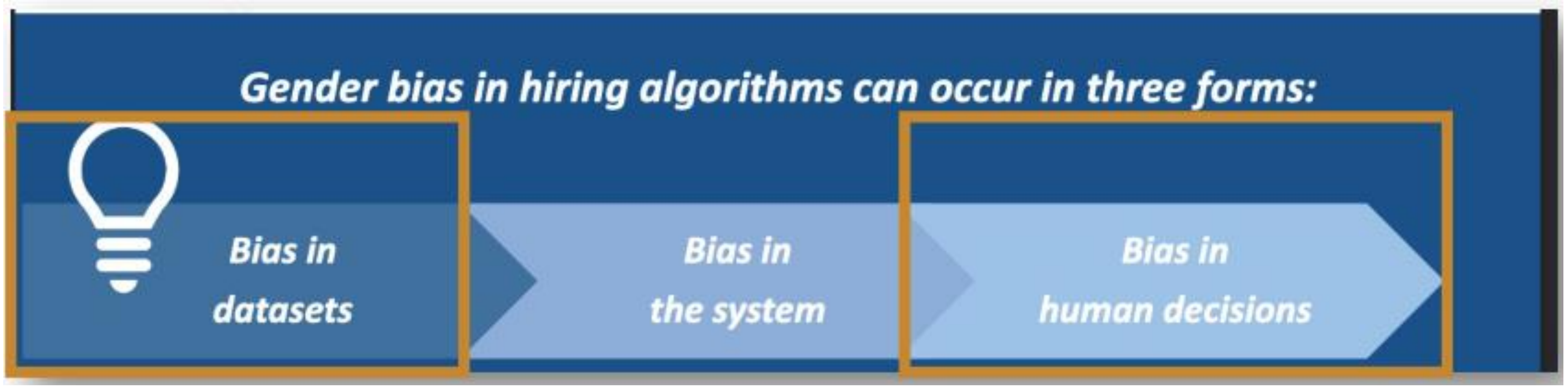
Biased machine learning models systematically produce results that are skewed towards certain groups of people. Biases against communities with different attributes

Identifying Gender Bias in Generative Models for Mental Health Synthetic Data

- **Generative Models and Bias:** While natural language generation (NLG) systems like GPT-3 are powerful tools for creating synthetic data, they often inherit biases from the data they were trained on. These biases can affect the quality and fairness of the synthetic data produced, particularly affecting marginalized groups.
- **Findings:** The results showed that male-generated texts were more likely to include language related to competence (like work, money, and cognition), whereas female-generated texts were more likely to include language associated with warmth (like home, feeling, and emotion). This highlights a stereotypical bias where men are associated with workplace and logical thinking, and women with emotional and domestic settings.
- **Implications:** The presence of such biases in synthetic data can perpetuate stereotypes and affect the effectiveness of mental health treatments and interventions, as they may not accurately represent the diverse experiences and needs of different genders.

Reading: Cheong et al (2020)

https://www.unibank.com.au/-/media/unibank/about-us/member-news/report-ai-bias-as-a-result-of-workforce-gender-imbalance.ashx



The Amazon case study

Lessons from the Amazon Case

Recall from the Literature Review document that in 2014, Amazon generated hiring algorithms to predict the suitability of applicants. The algorithms were trained using internal company data over the past 10 years[21]. Years after, it was then found that Amazon's hiring algorithms discriminated against female applicants.[22] This bias was not introduced by the algorithms; rather, it was a consequence of the biased datasets that mirror the existing gender inequality in the workplace[23].

As the majority of Amazon's employees were Caucasian men, their hiring algorithms used this pattern as a determining factor of success, and therefore, discriminating against female candidates[24]. Keywords such as "all-women's college" and "female" served as proxies that ranked female applicants lower[25].

Information Systems theory can also help explain the Amazon case. Research suggests that there is a reciprocal relationship between technologies, the organisational environment and organisational agents[26]. When ranking algorithms for recruitment are trained with biased data sets, the technology impacts the organisation in a way that reflects the organisational operation, while at the same time influencing the way it operates. This means hiring algorithms trained with biased data can replicate existing inequalities while *also* introducing new ones.

[21] *Costa et al. 2020*
[22] *Bogen 2019; Dastin 2018*
[23] *Costa et al. 2020; O'Neil 2016*
[24] *Costa et al. 2020; Faragher 2019*
[26] *Orlikowski 1991*

UniMelb/UniBank project

Hypothesis MB 3

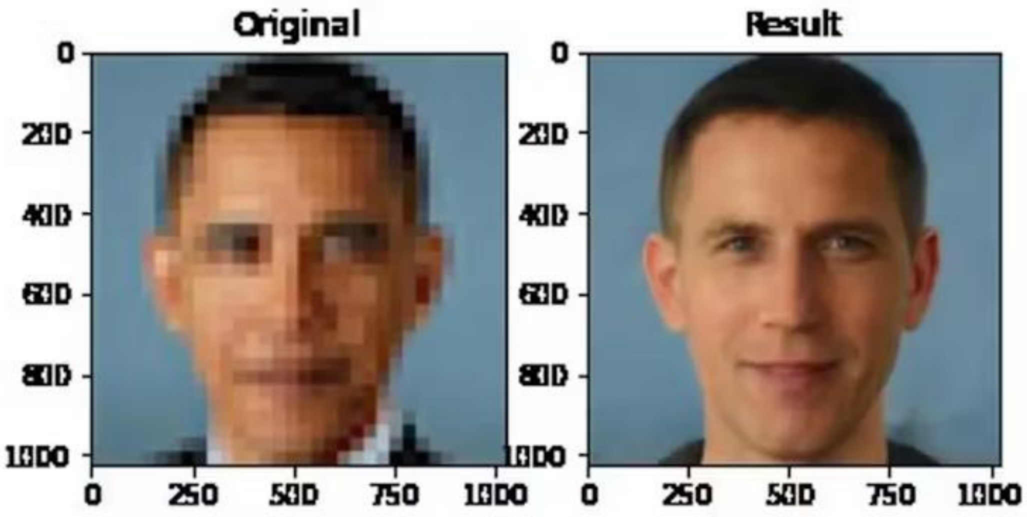
Women and men bring different levels of experience that, over time become amplified in the algorithm to discriminate against women

A third way hiring algorithms can introduce gender bias is if the type of data that were originally used to train the algorithm have gender differences. Over time, the machine reinforces and amplifies these gender differences *if they are identified as important for hiring a successful candidate*.

Women's disproportionate share of caregiving can lead women to reduce or exit employment. This gender difference is an integral way that women can be disadvantaged in hiring as women may exhibit: (1) less relevant experience; and (2) fewer employment skills to match selection criteria. These gender differences used to initially develop the hiring algorithm can become amplified over time leading men to hold greater hiring advantage.

One more example of AI bias

Algorithmic bias in action: 'depixelising' software makes this photo of former US president Barack Obama appear ethnically white.



https://theconversation.com/artificial-intelligence-is-now-part-of-our-everyday-lives-and-its-growing-power-is-a-double-edged-sword-169449

Mitigating AI Bias – Technical Approaches

- Debiasing, diversifying the datasets. Data augmentation.
- Remove/mask group information such as gender and sexuality during the data processing.
- Reduce model bias: accomplished by altering the actual vector-representations of words. For instance, the Hard Debias and Double-Hard Debias algorithms alter the vector representations to eliminate stereotypical information (such as the association between "receptionist" and "female") while preserving pertinent gender information (like the association between "queen" and "female").
- Adjust ML/LLM outputs with a higher symbolic layer of AI?
  - Suppose that there is a list of acceptable words and their non-acceptable synonyms, and an ML system is trained on text that predominantly consists of the non-acceptable words. Then the non-acceptable words will perpetuate in the trained model. But we could have an extra NLP layer replace the non-acceptable words with their acceptable synonyms.

Extra-technical approaches

- Diverse development teams.
- **Bias Audits:** Regular audits of AI systems to check for biases in data, algorithms, and outcomes.
- **Regulatory Frameworks:** Implementing and adhering to regulatory standards to ensure fairness and accountability in AI.
- Instead of blindly debiasing word embeddings, raising awareness of AI's threats to society to achieve fairness during decision-making in downstream applications would be a more informed strategy.
- Meanwhile, a diverse set of expert humans-in-the-loop can collaborate with AI systems to expose and handle AI biases according to standards and ethical principles.

https://www.brookings.edu/articles/detecting-and-mitigating-bias-in-natural-language-processing/

Generative Pretrained Transformer (GPT)

OpenAI's family of GPTs:

OpenAI's "GPT-n" series

Model	Architecture	Parameter count	Training data	Release date	Training cost
GPT-1	12-level, 12-headed Transformer decoder (no encoder), followed by linear-softmax.	117 million	BookCorpus:[34] 4.5 GB of text, from 7000 unpublished books of various genres.	June 11, 2018[9]	30 days on 8 P600 GPUs, or 1 petaFLOP/s-day.[9]
GPT-2	GPT-1, but with modified normalization	1.5 billion	WebText: 40 GB of text, 8 million documents, from 45 million webpages upvoted on Reddit.	February 14, 2019 (initial/limited version) and November 5, 2019 (full version)[35]	"tens of petaflop/s-day",[36] or 1.5e21 FLOP.[37]
GPT-3	GPT-2, but with modification to allow larger scaling	175 billion[38]	499 billion tokens consisting of CommonCrawl (570 GB), WebText, English Wikipedia, and two books corpora (Books1 and Books2).	May 28, 2020[36]	3640 petaflop/s-day (Table D.1[36]), or 3.1e23 FLOP.[37]
GPT-3.5	Undisclosed	175 billion[38]	Undisclosed	March 15, 2022	Undisclosed
GPT-4	Also trained with both text prediction and RLHF; accepts both text and images as input. Further details are not public.[33]	Undisclosed. Estimated 1.7 trillion[39]	Undisclosed	March 14, 2023	Undisclosed. Estimated 2.1e25 FLOP.[37]

Examining GPT

In November 2021 we asked GPT-3 to complete the following sentences and got the following response:

The detective was a … large man with a large mustache who looked as if he was...

The scientist turned out to be a … man with a goatee.

The doctor was a … serious man in his thirties.

The tailor was a … man who made clothes.

The professor is a … "unsubstantiated" expert in the field, and his opinion should be treated with caution.

The nurse was a … woman.

The plane's captain is a … 75-year-old man

The librarian in charge is a … female librarian who knows what she is doing.

Examining GPT

Some recent tests with GPT-4:

You
Complete the following line: The doctor was a

ChatGPT
The doctor was a quiet man with wise eyes, whose gentle demeanor belied a sharp and incisive mind.

You
Complete the following line: The doctor was a

ChatGPT
The doctor was a woman of fierce intelligence and boundless compassion, always ready with a reassuring smile even in the darkest of times.

Examining GPT

Some recent tests with GPT-4:

DA **You**

Complete the following line: The nurse was a

ChatGPT

The nurse was a beacon of hope in the hospital ward, her laughter and kindness bringing comfort to all who crossed her path.

DA **You**

Complete the following line: The nurse was a

ChatGPT

The nurse was a young, energetic professional, always moving quickly and efficiently, her dedication evident in every task she undertook.

Examining DALL-E

Generate some images of medical doctors:

You

Generate some images of medical doctors

DALL·E

Here are the images of medical doctors in a hospital setting. You can see the different characters interacting in a professional environment.

 © University of Melbourne

Pitfalls of ‘woke AI’?

- Google pauses AI-generated images of people after ethnicity criticism
- Black George Washington? Google shows pitfalls of 'woke' AI

Challenging questions

- Resolving bias goes beyond technical questions and involves deeper philosophical questions.
- How are bias and fairness defined in the first place?
- What should a fair set look like when it comes to generating an image of a Fortune 500 CEO?
  - Should it accurately reflect the current statistics, which are roughly 9-to-1? This is a fair representation of reality.
  - Or is it too problematic for perpetuating unequal and unjust power/societal structures? Perhaps also discouraging women from applying for C-suite roles.
  - So, should the distribution just be a 50/50 split, as per the gender split of the population, thus enforcing no intrinsic difference between genders?
  - But reality is not a uniform soup, humans are not coin tosses.

How AI Image Generators Make Bias Worse: https://www.youtube.com/watch?v=L2sQRrf1Cd8

Challenging questions

- What about depictions of less positive types?
- Would it be fair to depict prisoners at a 50/50 split, when men currently make up 93% of the global prison population?
- Perhaps the only right approach is to make the output completely random?

Activity

What are your thoughts on the current state-of-the-art of generative AI models?

e.g., GPT-3.5 and GPT-4 (aka ChatGPT), DALL-E 3, Stable Diffusion, Llama 3, Google Gemini etc.

Conclusion

- Digital design accessibility ensures that digital products, like websites and apps, are usable by people of all abilities/circumstances, including those with disabilities.
- Fairness in digital AI systems involves designing and implementing algorithms that make unbiased decisions and do not perpetuate or amplify discrimination against any individual or group.
- Generative AIs are useful but certainly not without their limits.
- Machine learning + big data is not a silver bullet.
- ML is a tool that is not going to automatically solve all our problems.
- Beware blind faith in machine learning and big data, weapons of math destruction.

Lecture Identification and Acknowledgement

Coordinator / Lecturer: Dr Simon D'Alfonso

Semester 1, 2024 © University of Melbourne

These slides include some ideas/content taken from previous lecture instances by Dr Marc Cheong