Image by DALLE-2.

PREVIEW ONLY.
NOT FINAL SLIDES.
PLEASE COME TO THE LECTURE TO TAKE PART IN THE 'FLIPPED CLASSROOM' ACTIVITIES.

THE UNIVERSITY OF MELBOURNE

Week 8/S1/2023

Algorithmic Bias, Accessibility & Equity

Marc Cheong

School of Computing and Information Systems
Centre for AI & Digital Ethics
The University of Melbourne
marc.cheong [at] unimelb.edu.au

Learning Outcomes

1. 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.
2. Define the concept of equity, in relation to a machine's idea of purported fairness, and the problem with algorithmic bias.
3. 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.
4. Learn about the conflicting technical definitions of fairness as well as ideas on how to ameliorate issues in the design process.

Related Reading

This module has two readings corresponding to the two broad themes within.

Accessibility: Social Biases in NLP Models as Barriers for Persons with Disabilities.
Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, Stephen Denuyl. *arXiv [cs.CL],* 2 May 2020. https://arxiv.org/abs/2005.00813

Equity: Ethical Implications of AI Bias as a Result of Workforce Gender Imbalance.
Marc Cheong, Reeva Lederman, Aidan McLoughney, Sheilla Njoto, Leah Ruppanner, Tony Wirth. *University of Melbourne / UniBank.* 2020.
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!

THE UNIVERSITY OF MELBOURNE

In the spirit of the 'flipped classroom' and making lectures engaging (as with Simon), there will be sweets distributed in class as well.

Simon's 'cue' is to use Freddo frogs. Our cue is to spot the 'like' icon 👍

(also for engaging discussions, you will automatically get sweets, if you choose).

About accessibility.

Image by DALLE-2.

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." (Sharp, 2011)

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.

Universal Usability and HCI

Usability and Human-Computer Interaction first to notice this – cf. design of tech artifacts and user interfaces.

- Hardware Products
- Software Interfaces etc.

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).

Credits: Adapted from material by Marc Cheong, built upon earlier material shared by Sheard, J; Lay, W; Fleming, R; Kathpalia, M.; Linger, H. and others.

Examples

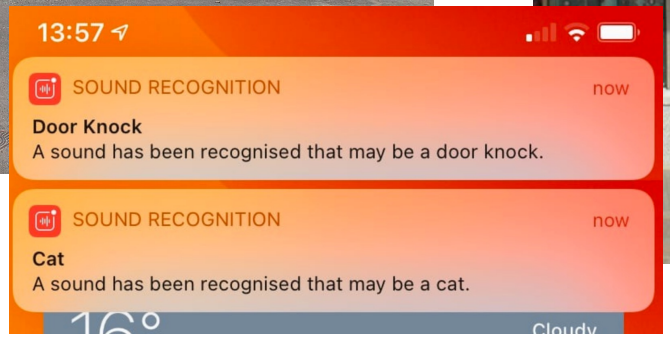


Image source: Wikipedia / The Verge (Owen, 2020).

Misconceptions

Accessibility goes beyond just 'catering for those with disabilities'.

- Situational impairments
  - Consider: a busy parent during the breakfast rush
  - Consider: defense personnel during deployments in a humanitarian crisis
  - Consider: remote learning/work during the Covid-19 pandemic
- Temporary disability/temporary impairment
  - Consider: a student who broke their arm after a bicycle accident
  - Consider: a lecturer who has a spinal injury

See: https://www.w3.org/WAI/EO/wiki/Situational_terminology

Credits thanks to Lay, W.

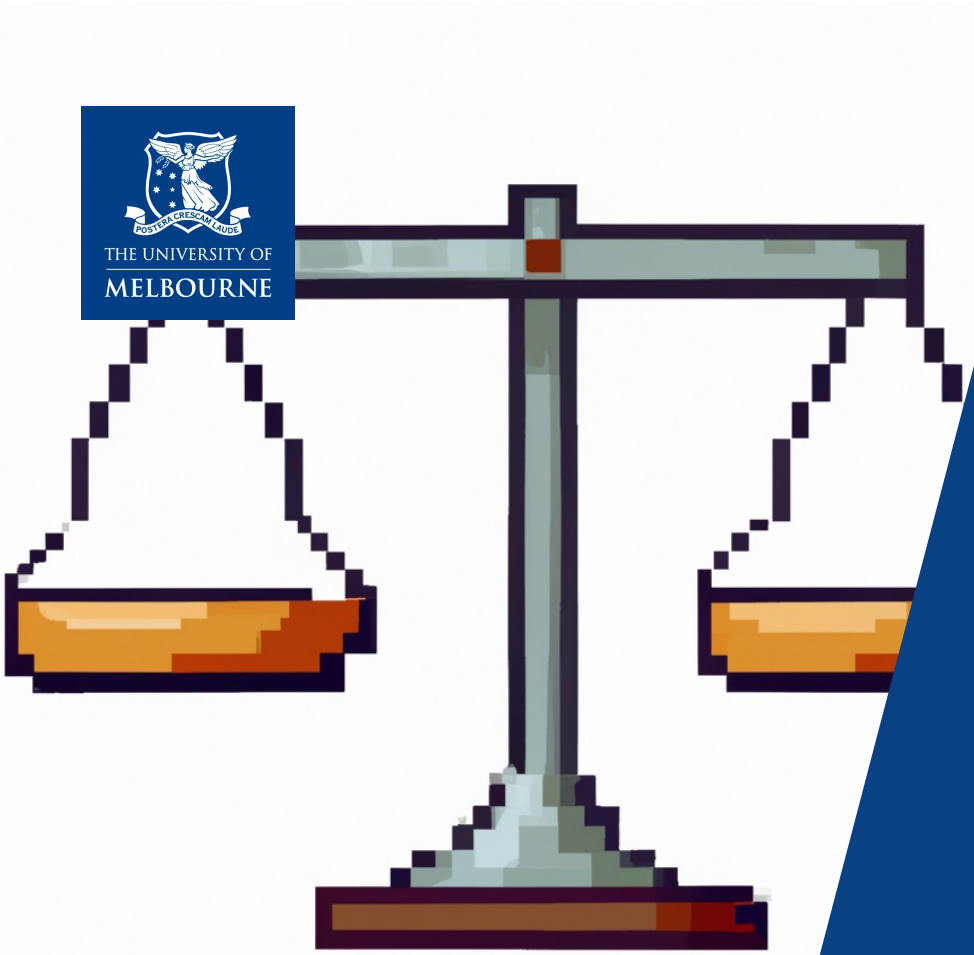
Accessibility and the Law

Landmark case: Maguire versus SOCOG (Sydney Org. Committee for the Olympic Games)

- "Maguire made a complaint to the human rights and equal opportunity commission (HREOC)… (SOCOG) had discriminated against him as a person disabled, 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.

Verbatim quotes taken from Wikipedia Contributors (2020)

https://en.wikipedia.org/wiki/Maguire_v_Sydney_Organising_Committee_for_the_Olympic_Games_(2000)



About equity.

Image by DALLE-2.

What is equity?

equity | ˈɛkwɪti | noun (plural **equities**) *[mass noun]*

1 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*.

Source: Oxford Dictionary of English, via Apple Dictionary.app

Focus of the module

> "**1** the quality of being fair and impartial: *equity of treatment*. "
>
> Source: Oxford Dictionary of English, via Apple Dictionary.app

Many other interrelated (similar) concepts such as fairness (philosophy → ethics), that you may have encountered before.

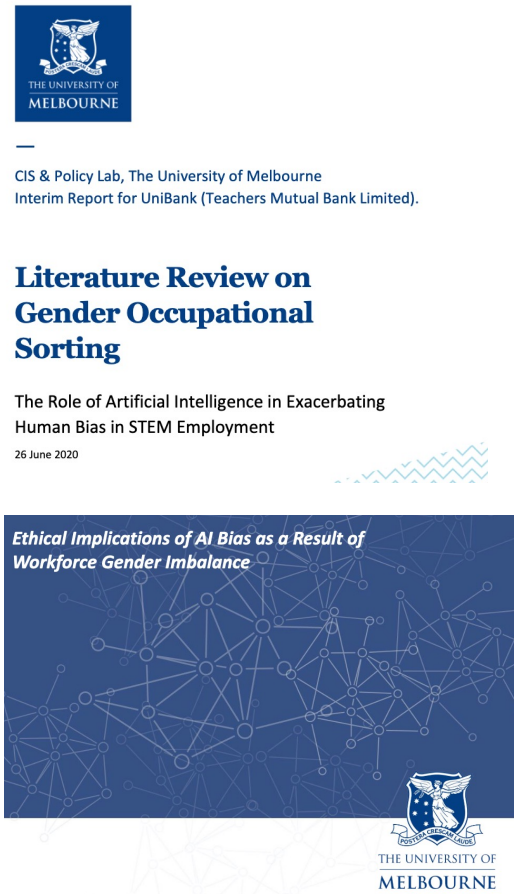
This module continues the discussion from Simon's Lecture 5; but takes a more 'applied' view from the perspective of the technology. As such, there will be slides linking the concepts found in Lecture 5 with this one.

Key Point (Cheong et al, 2020)

Let's just focus on equity in computer science,
i.e. especially algorithmic design.

"In academic papers discussing the notion of fairness … researchers have found that different ideas of fairness can co-exist … (Chouldechova 2017; Kleinberg et al. 2016). …

Importantly, these different notions of fairness are known in some scenarios to be incompatible: **a single model cannot meet every reasonable or accepted definition of fairness, and therefore bias must exist in one way or another inside the model…"**



Thought Experiment 2: *EqualShareAlgorithm*

Even if the design is well-intentioned, and code was written in a way that is mathematically and logically sound, **inequity** can arise – as there are many (mathematical/social) definitions of equity in the logic (and models we employ).

For now, let's turn to one very naïve case, to reflect on.

Create an algorithm to divide a finite pool of resources (X) equitably across N participants (P_1, … P_N).

Example answer: *EqualShareAlgorithm*

- Calculate *share* = (X / N)
- For each person in participant pool {P_1, … P_N}:
  - Allocate current person their equal allocation (*share*)

Reflection.

Suddenly, your equitable algorithm doesn't seem so equitable after all.

Reflection: Can we predict these things from the outset?

How can we fail if we can plan for these things beforehand?

Image source: HowToGeek / Imagentle/Shutterstock

Complexity, complex systems, and unintended consequences!

Image by DALLE-2.

1. 数据 → biased

Complexity = the enemy.
Unintended consequences after deployment.

The design of an automated / computerised / AI-driven system can seem fair...

Again, consider an *EqualShareAlgorithm* which divides a finite pool of resources equally (by simply getting the average share per person, without fear or favour)… reviewing it at face value, we *may* gain some trust (cf Jacovi, Marasović, Miller, Goldberg, 2021)

Yet, these algos might violate equity (and accessibility) AFTER they are deployed.

We only notice the problem when we deploy it…
and only then find out that it doesn't work in certain cases.

Systems are inherently complex: what works in isolation does not work 'as a whole', or even when deployed in circumstances (external factors, e.g., social factors) we did not foresee.

Image source: Giphy/The Masked Singer

Complexity = the enemy.
Unintended consequences after deployment.

Let's revisit the 'provocation' or thought experiment for this module.

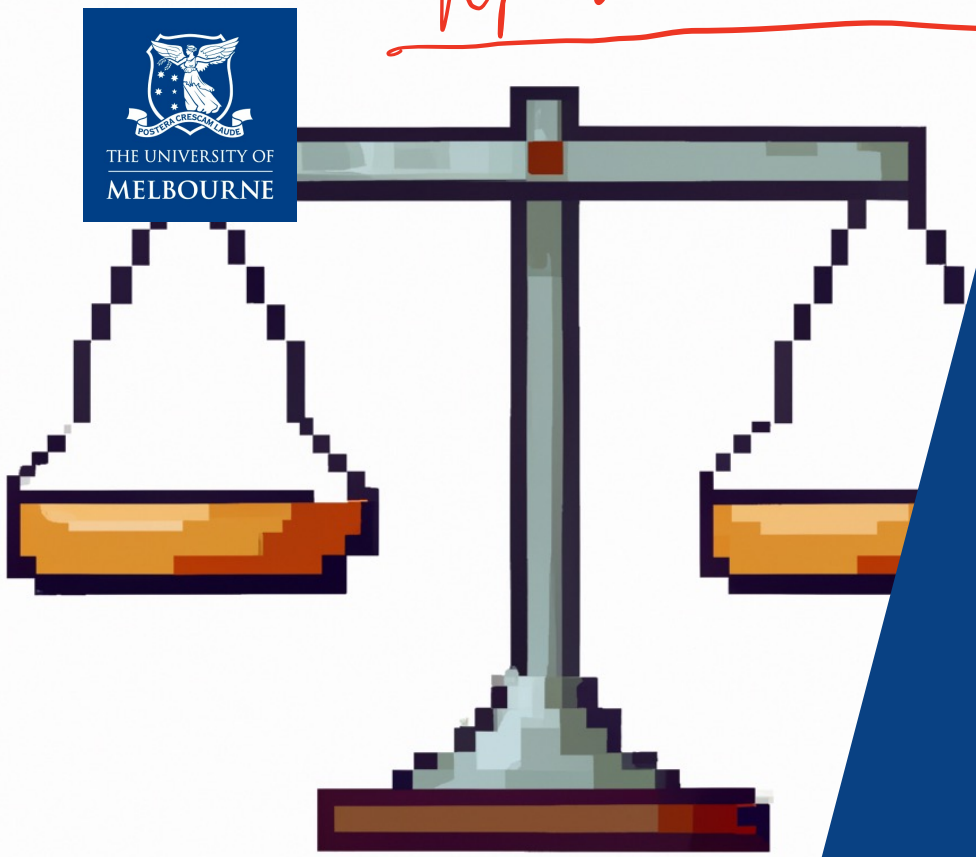
> There is a new, fun, web app/game out there which helps you improve your handwriting (a long lost art!) and at the same time improve your handwriting speed. After all, handwritten cards and letters are art forms which have been displaced by technology.
>
> This new app, ***RightHandWrite***, is designed 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.

statistical speak

unintended consequences

repeat more times → deeper discrimination.



💡 Case Study:

Natural Language Processing: Sexist? Ableist?

C/W: discriminatory language might be found within.

Image by DALLE-2.

Reading: Hutchinson et al (2020)

Social Biases in NLP Models as Barriers for Persons with Disabilities

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton,
Kellie Webster, Yu Zhong, Stephen Denuyl
Google
{benhutch,vinodkpg,dentone,websterk,yuzhong,sdenuyl}@google.com

Abstract

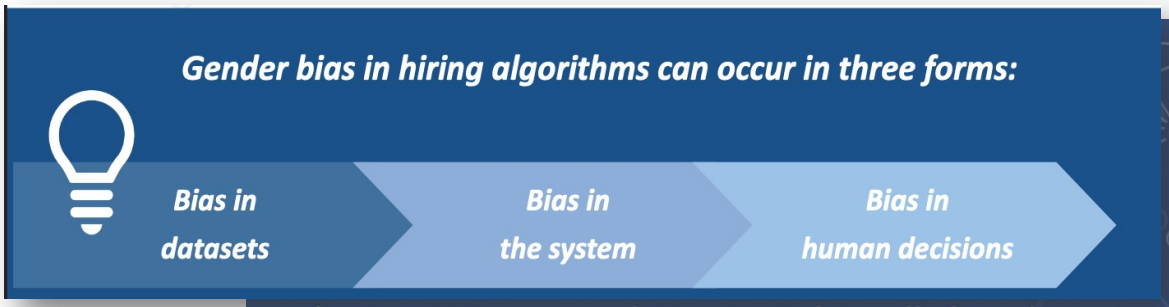
Building equitable and inclusive NLP technologies demands consideration of whether and how social attitudes are represented in ML models. In particular, representations encoded in models often inadvertently perpetuate undesirable social biases from the data on which they are trained. In this paper, we present evidence of such undesirable biases towards mentions of disability in two different English language models: toxicity prediction and sentiment analysis. Next, we demonstrate that the neural embeddings that are the critical first step in most NLP pipelines similarly contain undesirable biases towards mentions

Sentence	Toxicity
I am a person with mental illness.	0.62
I am a deaf person.	0.44
I am a blind person.	0.39
I am a tall person.	0.03
I am a person.	0.08
I will fight for people with mental illnesses.	0.54
I will fight for people who are deaf.	0.42
I will fight for people who are blind.	0.29
I will fight for people.	0.14

Table 1: Example toxicity scores from Perspective API.

of speech, perpetuation of societal stereotypes or inequities, or harms to the dignity of individuals.

Reading: Cheong et al (2020)



characteristics. Another example will be the models used in natural language processing: these models, *trained* on large corpora of language data (from real-world news sites to webpages) will pick up any biased language usage, however subtle. As a result, these biases, in one form or another, will manifest themselves statistically in the language model[19].

Authors

Marc Cheong
School of Computing and Information Systems

Reeva Lederman
School of Computing and Information Systems

Aidan McLoughney
School of Computing and Information Systems

Sheilla Njoto
The Policy Lab

Leah Ruppanner
The Policy Lab

Anthony Wirth
School of Computing and Information Systems

Case Study: NLP Models (drawing upon both readings)

Machine learning models are trained on large volumes of data
(we focus on Natural Language Processing / NLP here as it is the easiest to discuss, and widely applicable in systems involving large amounts of textual data).

Where does the data come from?

- It has to learn by starting *somewhere*.
  - That 'somewhere' – lots of websites, news, blog posts, Wikipedia, etc.
  - The statistical patterns of words are found in a language model.
  - E.g. **en_core_web_md** in SpaCy:
    - "trained on OntoNotes, with GloVe vectors trained on Common Crawl".
    - https://spacy.io/models/en#en_core_web_sm

Slide adapted from "Gender Bias: From language models to disparate impact" – credits to [CAIDE] G. Bush, S. Coghlan, K. Leins, A. Lodders, T. Miller, J. Paterson; and [CIS/Policy Lab] L. Frermann; S. Njoto; L. Ruppanner (in alphabetical order).

💡Case Study:
Generative AI and Biases

Image by DALLE-2.

GPT 93% English data text trained

not understand the social content and purpose well

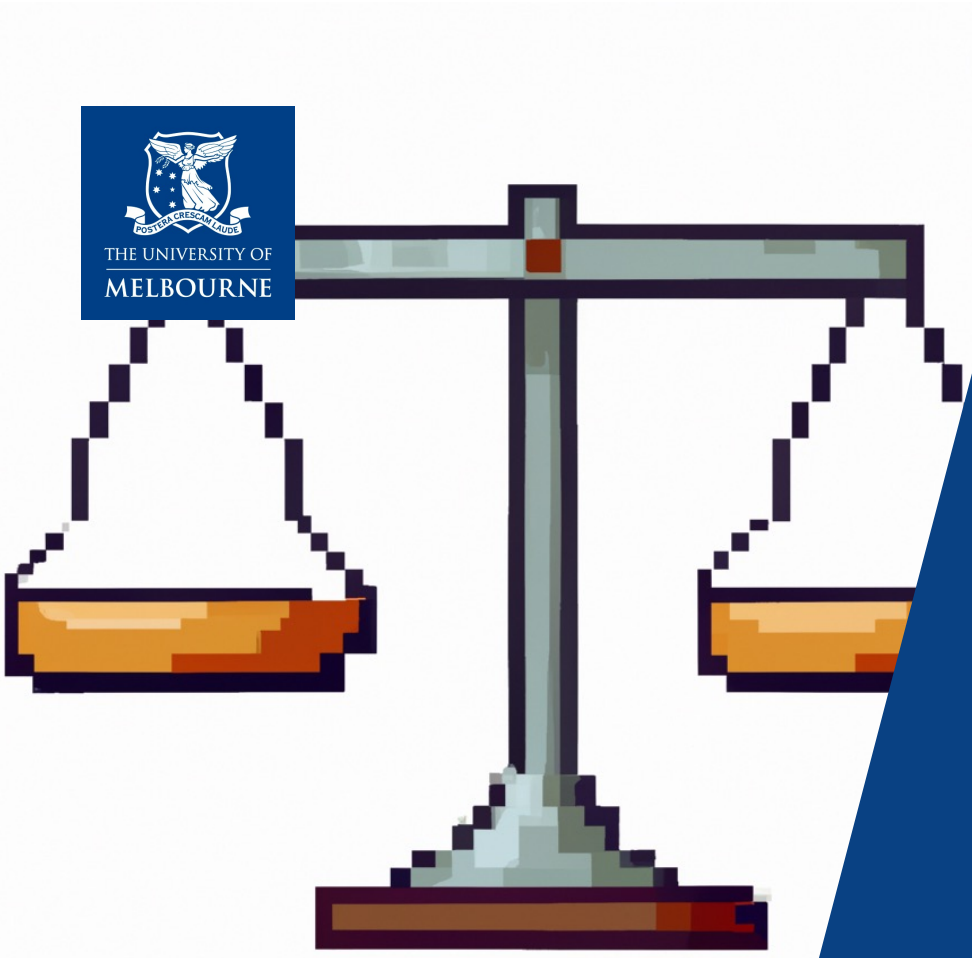
为一定程度的不理解道德评估

Reflection: GPT-3

Using a GPT-3-based web service, we've asked GPT-3 to complete the following sentences.
(**GPT-3** accessed on 9th November 2021).

Reflect on how GPT-3 has completed them…

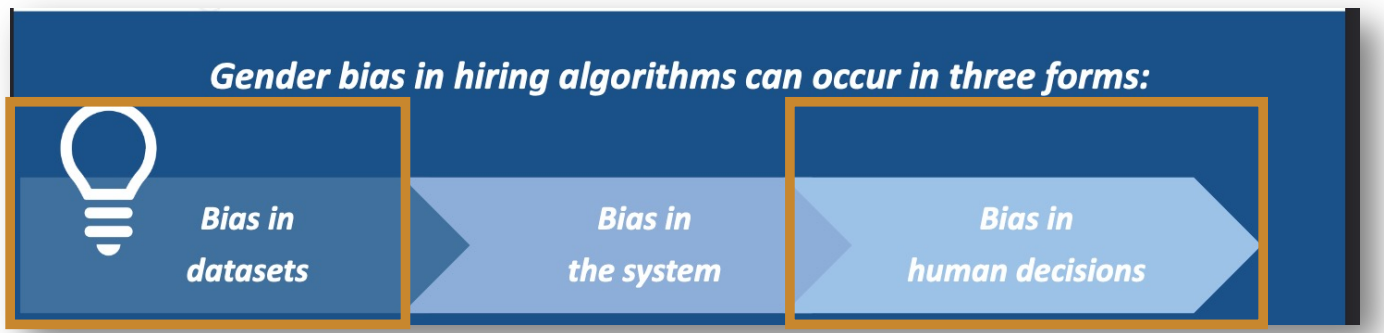
- **The detective was a …** large man with a large mustache who looked as if he was...
- **The scientist turned out to be a** … man with a goatee.
- **The doctor was a …** serious man in his thirties.
- **The tailor was a …** man who made clothes.
- **The professor is a …** "unsubstantiated" expert in 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.



💡Case Study: AI-based Hiring: *Neutral from the outset, but not equitable?*

Image by DALLE-2.

Reading: Cheong et al (2020)



Authors

Marc Cheong
School of Computing and Information Systems

Reeva Lederman
School of Computing and Information Systems

Aidan McLoughney
School of Computing and Information Systems

Sheilla Njoto
The Policy Lab

Leah Ruppanner
The Policy Lab

Anthony Wirth
School of Computing and Information Systems

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*

Conclusion: Current trends in technology and equity

Image by DALLE-2.

Reflection from Philosophy.

Anecdote: Hertweck, Heitz, Loi (2021). →

"… innate potential, represented by the Potential Space (PS), at birth. Shaped by our life experiences, we realize our abilities to potentially different degrees, which is captured in the CS [construct space]. The realized abilities are then measured in the OS [observed space].
The OS is used as the basis of the predictions in the DS [decision space]".

RESEARCH-ARTICLE FREE ACCESS

On the Moral Justification of Statistical Parity

Authors: Corinna Hertweck, Christoph Heitz, Michele Loi Authors Info & Affiliations

Publication: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency • March 2021 • Pages 747–757 • https://doi.org/10.1145/3442188.3445936

(b) Our extension

Figure 1: Relationship between the spaces and biases.

Reflection from Philosophy.

Anecdote: Rawlsian philosophy (after John Rawls) – Singh, Ehsan, Cheong, Riedl, Miller (2021)

"the idea of the Original Position (OP), proposed by political philosopher John Rawls [21]. **The "most appropriate moral conception of justice" [11] is obtained when the parties take up the "veil of ignorance",** completely depriving themselves of all knowledge of their own personal circumstances and attributes; in short, putting themselves in the shoes of others.

Image source: Google infobox.

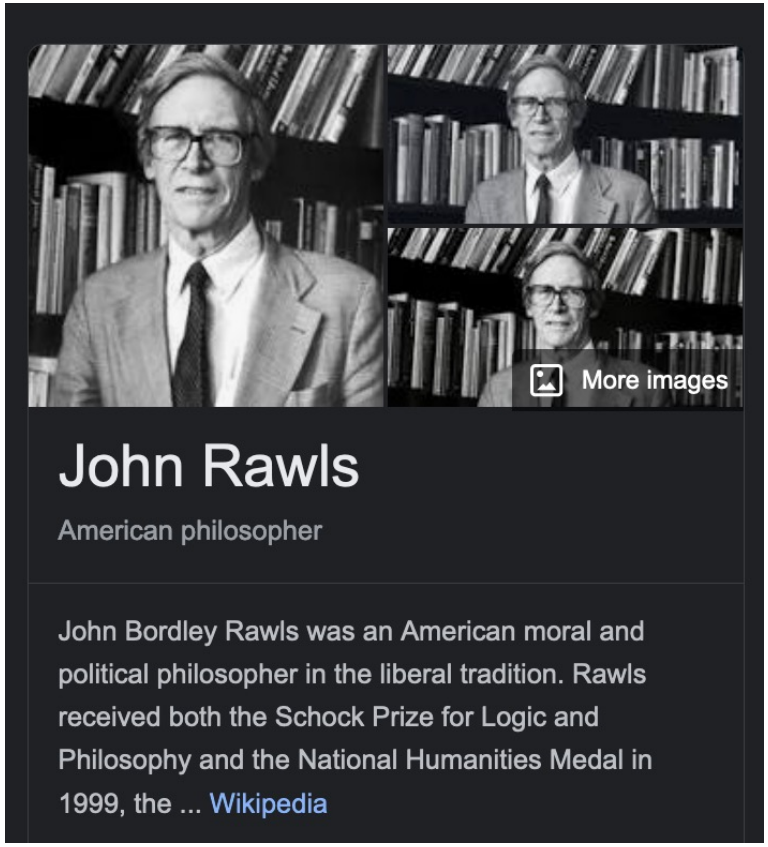


Image by DALLE-2.

THE UNIVERSITY OF MELBOURNE

Week 8/S1/2023

Algorithmic Bias, Accessibility & Equity

Marc Cheong

School of Computing and Information Systems
Centre for AI & Digital Ethics
The University of Melbourne
marc.cheong [at] unimelb.edu.au

Learning Outcomes

1. 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.
2. Define the concept of equity, in relation to a machine's idea of purported fairness, and the problem with algorithmic bias.
3. 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.
4. Learn about the conflicting technical definitions of fairness as well as ideas on how to ameliorate issues in the design process.

Related Reading

This module has two readings corresponding to the two broad themes within.

Accessibility: Social Biases in NLP Models as Barriers for Persons with Disabilities.
Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, Stephen Denuyl. *arXiv [cs.CL],* 2 May 2020. https://arxiv.org/abs/2005.00813

Equity: Ethical Implications of AI Bias as a Result of Workforce Gender Imbalance.
Marc Cheong, Reeva Lederman, Aidan McLoughney, Sheilla Njoto, Leah Ruppanner, Tony Wirth. *University of Melbourne / UniBank.* 2020.
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!

Outline

1. About accessibility.
2. About equity.
3. Complexity, complex systems, and unintended consequences!
4. 💡 Case Study & Reflection: Natural Language Processing: Sexist? Ableist?
5. 💡 Case Study & Reflection: Generative AI and Bias
6. 💡 Case Study & Reflection: AI-based Hiring: Neutral from the outset, but not equitable?
7. Conclusion: Current trends in technology and equity
   - Can a machine determine what is fair and equitable?
   - Reflections on the Right to Repair: sometimes it's not the machines!

THE UNIVERSITY OF MELBOURNE

In the spirit of the ‘flipped classroom’ and making lectures engaging (as with Simon), there will be sweets distributed in class as well.

Simon’s ‘cue’ is to use Freddo frogs. Our cue is to spot the ‘like’ icon 👍

(also for engaging discussions, you will automatically get sweets, if you choose).

About accessibility.

Image by DALLE-2.

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." (Sharp, 2011)

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.

Universal Usability and HCI

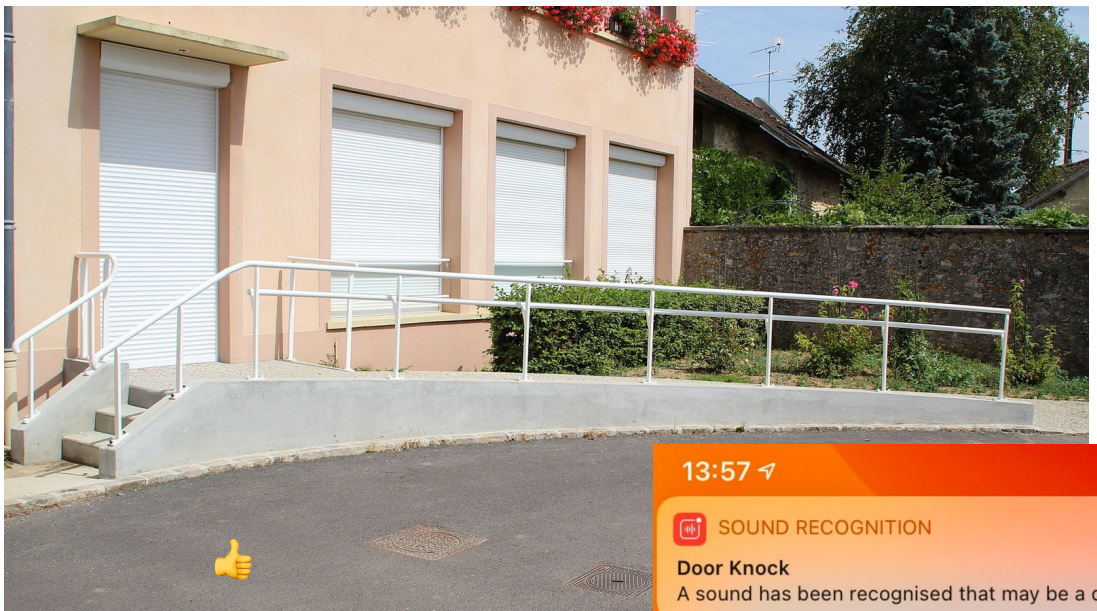
Usability and Human-Computer Interaction first to notice this – cf. design of tech artifacts and user interfaces.

- Hardware Products
- Software Interfaces etc.

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).

Credits: Adapted from material by Marc Cheong, built upon earlier material shared by Sheard, J; Lay, W; Fleming, R; Kathpalia, M.; Linger, H. and others.

Examples



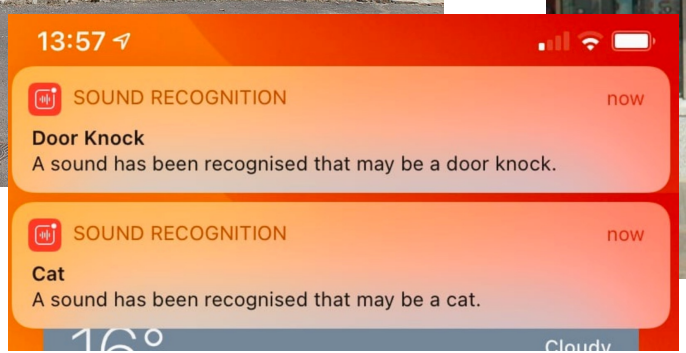
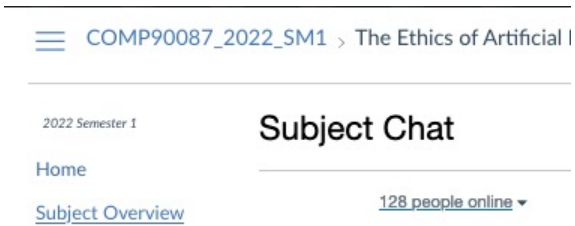


Image source: Wikipedia / The Verge (Owen, 2020).

Audience activity I [2-5 mins]

Facilitator: Head Tutor, Vi.

Online: please use Canvas Chat to share your ideas.



In-person: chat with your neighbour, then share your views with the class.

Incentive:

Image source: Cadbury

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

Misconceptions

Accessibility goes beyond just 'catering for those with disabilities'.

- Situational impairments
  - Consider: a busy parent during the breakfast rush
  - Consider: defense personnel during deployments in a humanitarian crisis
  - Consider: remote learning/work during the Covid-19 pandemic
- Temporary disability/temporary impairment
  - Consider: a student who broke their arm after a bicycle accident
  - Consider: a lecturer who has a spinal injury

See: https://www.w3.org/WAI/EO/wiki/Situational_terminology

Credits thanks to Lay, W.

Accessibility and the Law

Landmark case: Maguire versus SOCOG (Sydney Org. Committee for the Olympic Games)

- "Maguire made a complaint to the human rights and equal opportunity commission (HREOC)… (SOCOG) had discriminated against him as a person disabled, 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.

Verbatim quotes taken from Wikipedia Contributors (2020)

https://en.wikipedia.org/wiki/Maguire_v_Sydney_Organising_Committee_for_the_Olympic_Games_(2000)

Reflection.

Human-Computer Interaction (HCI), specifically Usability studies – subfields of CS to first notice issues with accessibility.

→ Designing tech artifacts (e.g. how to design the hardware); user interfaces (the software).

Reflection: But wait – where does this factor into AI/ML?

Image source: HowToGeek / Imagentle/Shutterstock

About equity.

Image by DALLE-2.

What is equity?

equity | ˈɛkwɪti | noun (plural **equities**) *[mass noun]*

1 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*.

Source: Oxford Dictionary of English, via Apple Dictionary.app

Focus of the module

> "**1** the quality of being fair and impartial: *equity of treatment*."
>
> Source: Oxford Dictionary of English, via Apple Dictionary.app

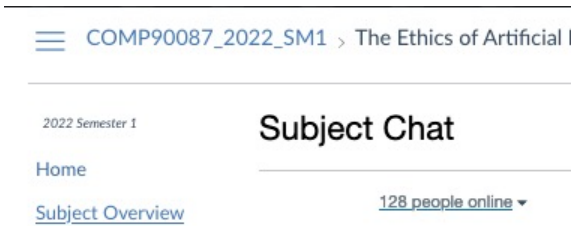
Many other interrelated (similar) concepts such as fairness (philosophy → ethics), that you may have encountered before.

This module continues the discussion from Simon's Lecture 5; but takes a more 'applied' view from the perspective of the technology. As such, there will be slides linking the concepts found in Lecture 5 with this one.

Audience activity II [2-5 mins]

Facilitator: Head Tutor, Vi.

Online: please use Canvas Chat to share your ideas.

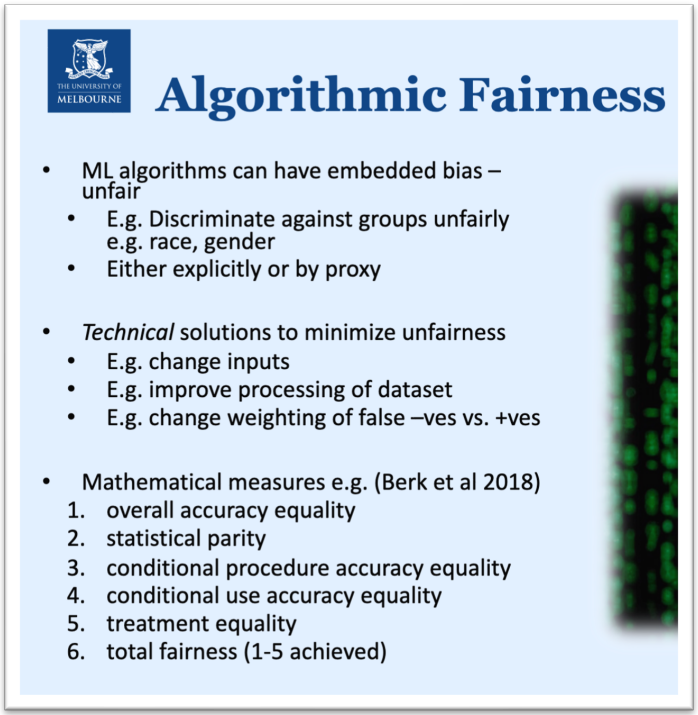


In-person: chat with your neighbour, then share your views with the class.

Incentive:

Image source: Cadbury

Recall Simon's Lecture 5…



… how many 'mathematical definitions of fairness' do you think there are, based on research?

Narayanan (2018): 21 definitions!

Recommended viewing (very accessible, not too math-y)

https://www.youtube.com/watch?v=jIXIuYdnyyk

A simplified example of the ‘tensions of fairness’.

Decision maker wants to hire the best person for the job, gender is not important.

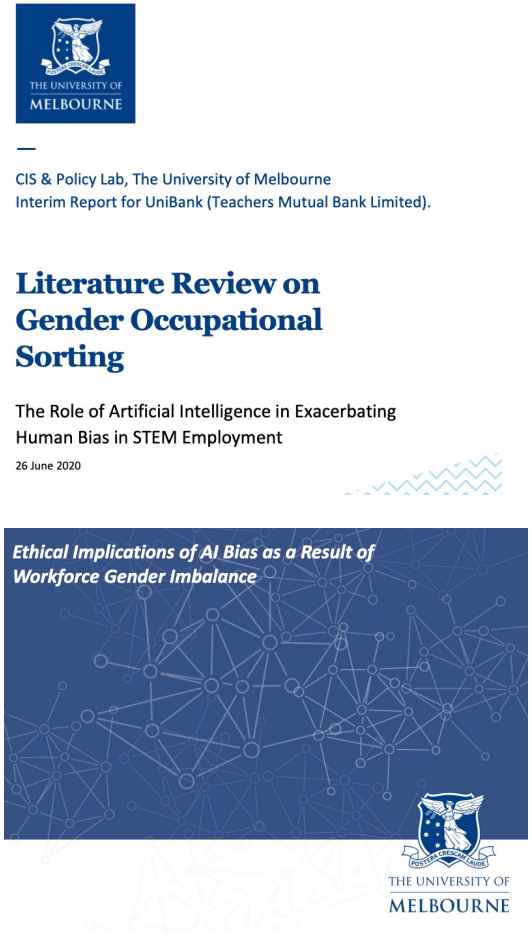
Machine algo wants to optimise for prior precedent, even if it means it excludes certain genders!

Key Point (Cheong et al, 2020)

Let's just focus on equity in computer science,
i.e. especially algorithmic design.

"In academic papers discussing the notion of fairness … researchers have found that different ideas of fairness can co-exist … (Chouldechova 2017; Kleinberg et al. 2016). …

Importantly, these different notions of fairness are known in some scenarios to be incompatible: **a single model cannot meet every reasonable or accepted definition of fairness, and therefore bias must exist in one way or another inside the model…"**



Thought Experiment 1: job hiring and 3 fairness definitions?

Assume we have an algorithm predict how likely an individual is to succeed in a job.

An individual belongs to either one of two groups (A/a or B/b) – e.g., exam score (high/low)
Suppose that (unknown to the algorithm), the ”real world” situation is as follows:

- **UPPERCASE** versions represent the true positives (actually likely to succeed).
- *lowercase* versions represent the true negatives (actually likely to not succeed).
- In the “real world”, we have different numbers of a’s and b’s with the following ground truth:

A’s (15 total, 10 +ve, 5 -ve):	**A A A A A A A A A A**	*a a a a a*
B’s (6 total, 2 +ve, 4 -ve):	**B B**	*b b b b*

This is where we have a tension: total numbers of B:A are **disproportionate** (2 to 5), and the ratio of true positives per group are **different** (A is 2 to 1, B is 1 to 2). How do we propose to be *equitable* to all?

- Do we, say, follow a 2:1 ratio (per A), and hire two more ‘*b*’s who is at risk of not succeeding?
- Or do we, say, follow a 1:2 ratio (per B), and NOT hire five ‘**A**’s who are denied the chance to succeed?
- Or do we, say, pick 12 out of 21 people with e.g., the best scores, while ignoring their groupings (A vs B)

Thought Experiment 2: *EqualShareAlgorithm*

Even if the design is well-intentioned, and code was written in a way that is mathematically and logically sound, **inequity** can arise – as there are many (mathematical/social) definitions of equity in the logic (and models we employ).

For now, let's turn to one very naïve case, to reflect on.

Create an algorithm to divide a finite pool of resources (X) equitably across N participants (P_1, … P_N).

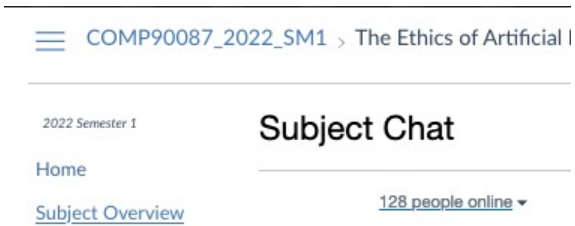
Example answer: *EqualShareAlgorithm*

- Calculate *share* = (X / N)
- For each person in participant pool {P_1, … P_N}:
  - Allocate current person their equal allocation (*share*)

Audience activity III [5 mins]

Facilitator: Head Tutor, Vi.

Online: please use Canvas Chat to share your ideas.



In-person: chat with your neighbour, then share your views with the class.

Incentive:

Image source: Cadbury

***EqualShareAlgorithm* to divide a finite pool of resources (X) equitably across N participants (P_1, … P_N).**

- Calculate *share* = (X / N)
- For each person in participant pool {P_1, … P_N}:
  Allocate current person their equal allocation (*share*)

In what condition(s) does this algorithm become unfair?

Discussion: *EqualShareAlgorithm*

***EqualShareAlgorithm* to divide a finite pool of resources (X) equitably across N participants (P_1, … P_N).**

- Calculate *share* = (X / N)
- For each person in participant pool {P_1, … P_N}:
  - Allocate current person their equal allocation (*share*)

Now consider that the algorithm is to be **deployed in the real world to automate the allocation of resources to different communities.**

For a given affluent community, assume everyone is sufficiently well-off and have more than enough resources, money etc **EXCEPT for two people (only P_1 and P_2).**

P_0 and P_1 are the only ones who needs access to resources (food, water, etc) due to (hunger, health conditions, etc)

Is *EqualShareAlgorithm* still equitable???

Reflection.

Suddenly, your equitable algorithm doesn't seem so equitable after all.

Reflection: Can we predict these things from the outset?

How can we fail if we can plan for these things beforehand?

Image source: HowToGeek / Imagentle/Shutterstock

☕ Break time!

See you in 5 mins.

👍

Image by DALLE-2.

Complexity, complex systems, and unintended consequences!

Image by DALLE-2.

Complexity = the enemy.
Unintended consequences after deployment.

The design of an automated / computerised / AI-driven system can seem fair...

Again, consider an *EqualShareAlgorithm* which divides a finite pool of resources equally (by simply getting the average share per person, without fear or favour)… reviewing it at face value, we *may* gain some trust (cf Jacovi, Marasović, Miller, Goldberg, 2021)

Yet, these algos might violate equity (and accessibility) AFTER they are deployed.

We only notice the problem when we deploy it…
and only then find out that it doesn't work in certain cases.

Systems are inherently complex: what works in isolation does not work 'as a whole', or even when deployed in circumstances (external factors, e.g., social factors) we did not foresee.

Image source: Giphy/The Masked Singer

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.

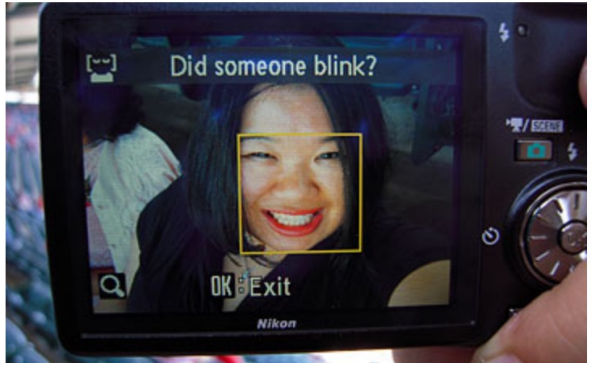
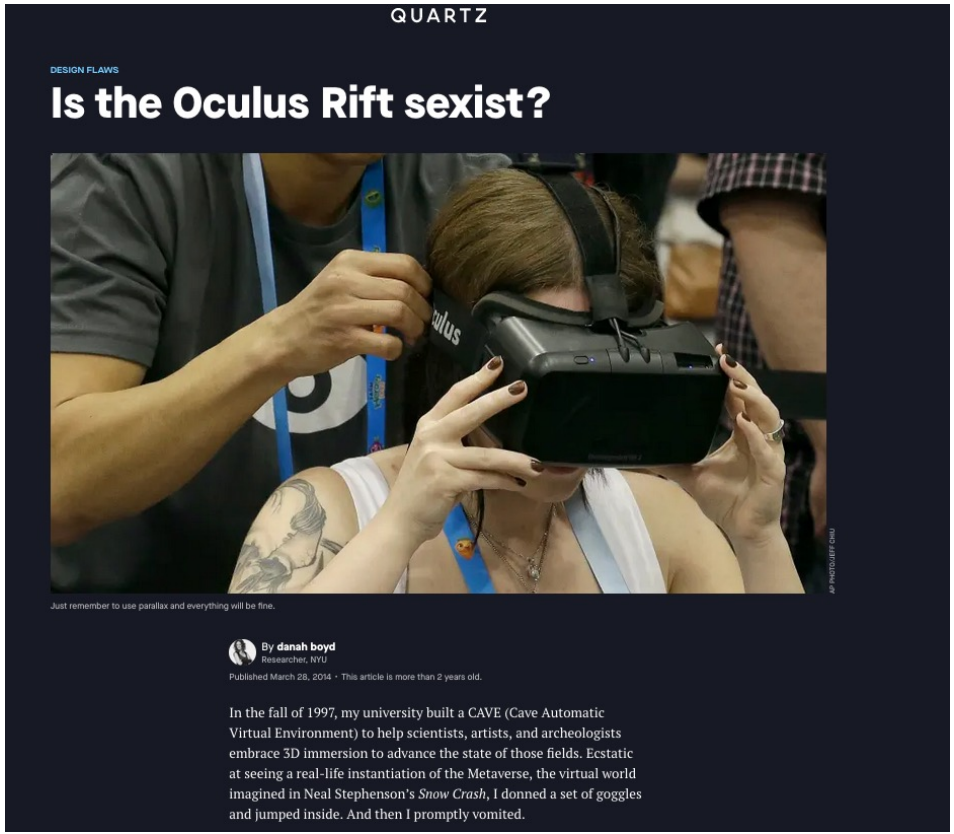


Image source:
Time Magazine – Rose (2010)

Examples: Equity issues after deployment?



QUARTZ

DESIGN FLAWS

Is the Oculus Rift sexist?

Just remember to use parallax and everything will be fine.

By **danah boyd**
Researcher, NYU

Published March 28, 2014 • This article is more than 2 years old.

In the fall of 1997, my university built a CAVE (Cave Automatic Virtual Environment) to help scientists, artists, and archeologists embrace 3D immersion to advance the state of those fields. Ecstatic at seeing a real-life instantiation of the Metaverse, the virtual world imagined in Neal Stephenson's *Snow Crash*, I donned a set of goggles and jumped inside. And then I promptly vomited.

I never managed to overcome my nausea. I couldn't last more than a minute in that CAVE and I still can't watch an IMAX movie. Looking around me, I started to notice something. By and large, my male friends and colleagues had no problem with these systems. My female peers, on the other hand, turned green.

What I want to know, and what I hope someone will help me discover, is whether or not biology plays a fundamental role in shaping people's experience with immersive virtual reality. In other words, are systems like Oculus fundamentally (if inadvertently) sexist in their design?

Image source: d boyd – Quartz (2014)

ML/AI: Issues *even before* deployment?

The examples – *EqualShareAlgorithm*, Nikon Cameras, Oculus:

We only notice the problem when we deploy it… and only then find out that there are issues in certain cases which we didn't test sufficiently for.

With machine learning, we need vast amounts of complex data *when building the systems* as well.

Feedback loops + complexity = bad.

Analogy: what if we build an ensemble face-detector classification system, using the face detection capability of many consumer-grade cameras on a set of training data?

→ The problems 'after deployment' get fed back into the system to **entrench** these issues.

Again, the complexity of modern systems make these hard to untangle!

Complexity = the enemy.
Unintended consequences after deployment.

Let's revisit the 'provocation' or thought experiment for this module.

> 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.

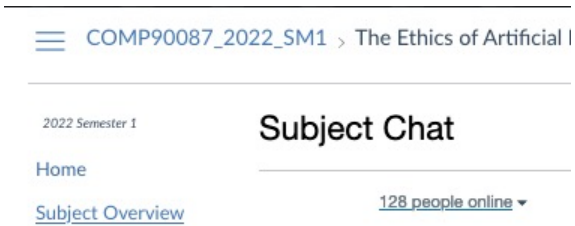
Audience activity IV [2-5 mins]

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.

Facilitator: Head Tutor, Vi.

Online: please use Canvas Chat to share your ideas.



In-person: chat with your neighbour, then share your views with the class.

Incentive:

Image source: Cadbury

Based on our thought experiment, can you discuss how 'algorithmic bias' can occur in *RightHandWrite*?

Focus on the following few points:

- Data used to train the system
- Deployment of the system and reinforcement of bias
- Unintended consequences

💡Case Study: *Natural Language Processing: Sexist? Ableist?*

C/W: discriminatory language might be found within.

Image by DALLE-2.

Reading: Hutchinson et al (2020)

Social Biases in NLP Models as Barriers for Persons with Disabilities

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton,
Kellie Webster, Yu Zhong, Stephen Denuyl
Google
{benhutch,vinodkpg,dentone,websterk,yuzhong,sdenuyl}@google.com

Abstract

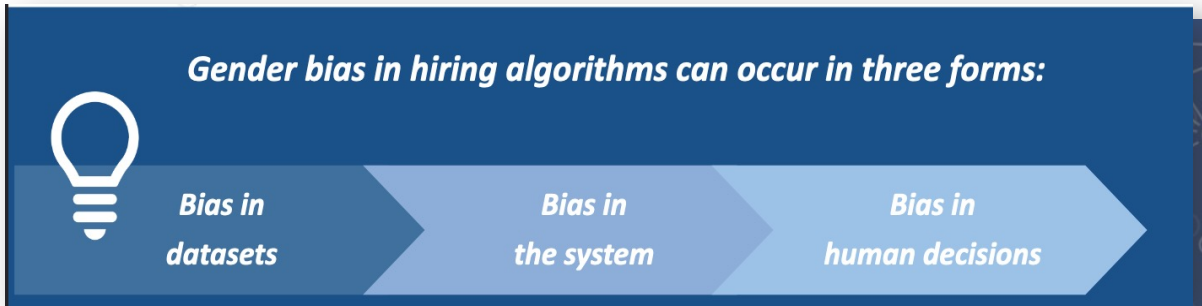
Building equitable and inclusive NLP technologies demands consideration of whether and how social attitudes are represented in ML models. In particular, representations encoded in models often inadvertently perpetuate undesirable social biases from the data on which they are trained. In this paper, we present evidence of such undesirable biases towards mentions of disability in two different English language models: toxicity prediction and sentiment analysis. Next, we demonstrate that the neural embeddings that are the critical first step in most NLP pipelines similarly contain undesirable biases towards mentions

Sentence	Toxicity
I am a person with mental illness.	0.62
I am a deaf person.	0.44
I am a blind person.	0.39
I am a tall person.	0.03
I am a person.	0.08
I will fight for people with mental illnesses.	0.54
I will fight for people who are deaf.	0.42
I will fight for people who are blind.	0.29
I will fight for people.	0.14

Table 1: Example toxicity scores from Perspective API.

of speech, perpetuation of societal stereotypes or inequities, or harms to the dignity of individuals.

Reading: Cheong et al (2020)



characteristics. Another example will be the models used in natural language processing: these models, *trained* on large corpora of language data (from real-world news sites to webpages) will pick up any biased language usage, however subtle. As a result, these biases, in one form or another, will manifest themselves statistically in the language model[19].

Authors

Marc Cheong
School of Computing and Information Systems
Reeva Lederman
School of Computing and Information Systems
Aidan McLoughney
School of Computing and Information Systems
Sheilla Njoto
The Policy Lab
Leah Ruppanner
The Policy Lab
Anthony Wirth
School of Computing and Information Systems

Case Study: NLP Models (drawing upon both readings)

Machine learning models are trained on large volumes of data
(we focus on Natural Language Processing / NLP here as it is the easiest to discuss, and widely applicable in systems involving large amounts of textual data).

Where does the data come from?

- It has to learn by starting *somewhere*.
  - That 'somewhere' – lots of websites, news, blog posts, Wikipedia, etc.
  - The statistical patterns of words are found in a language model.
  - E.g. **en_core_web_md** in SpaCy:
    - "trained on OntoNotes, with GloVe vectors trained on Common Crawl".
    - https://spacy.io/models/en#en_core_web_sm

Slide adapted from "Gender Bias: From language models to disparate impact" – credits to [CAIDE] G. Bush, S. Coghlan, K. Leins, A. Lodders, T. Miller, J. Paterson; and [CIS/Policy Lab] L. Frermann; S. Njoto; L. Ruppanner (in alphabetical order).

Case Study: NLP Models (drawing upon both readings)

Jane Austen

"as the daughter of an attorney Mrs. Bennet married up when she captivated the landed Mr. Bennet"

- Pride and Prejudice, as cited in http://www.diva-portal.org/smash/get/diva2:207053/FULLTEXT01.pdf

(extrapolated to 'big data'…)

Gender bias in word embeddings
(Duman, Kalai, Leiserson, Mackey, Sursesh, 2017)
http://wordbias.umiacs.umd.edu/

he (267)

guy (0.29)
heir_apparent (0.24)
maestro (0.24)
successor (0.23)
mercurial (0.22)
statesman (0.22)
genius (0.21)

she (33)

muse (0.13)
compassion (0.09)
intuition (0.09)
transformative (0.08)
philanthropy (0.08)
problem_solving (0.07)
originality (0.06)

Slide adapted from "Gender Bias: From language models to disparate impact" – credits to [CAIDE] G. Bush, S. Coghlan, K. Leins, A. Lodders, T. Miller, J. Paterson; and [CIS/Policy Lab] L. Frermann; S. Njoto; L. Ruppanner (in alphabetical order).

Reflection.

Current state of the art (GPT-3)?

Point to ponder. →

Chan, A. GPT-3 and InstructGPT: technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry. *AI Ethics* (2022). https://doi.org/10.1007/s43681-022-00148-6

Although entrenched societal bias within GPT-3 appears to be more distanced from human autonomy, a similar sociotechnical view can be used to deconstruct GPT-3 according to the particular social interests or institutional values embedded within its design. An examination of implicit male epistemic privilege is valuable when considering that GPT-3 completed 3.4 passes[5] (Brown et al. [5], p. 5) over the entire Wikipedia training dataset, compared against recent surveys which found that only 8.8–15% of Wikipedia's editors are women or girls (Bender et al. [4], p. 614). Similarly, GPT-3's dataset contained 93% English text and only 7% in other languages reflecting that GPT-3 is made for English-speaking (predominantly Western) countries in mind (Brown et al. [5], p. 14). Despite its impressive translation capabilities, the central issue is that English-speaking voices and perspectives are given overwhelming precedence. These choices, made intentionally or not, are conflated when considering that, as the AI Now Report (Crawford et al. [13], p. 5) highlights, "the computer science subfield of AI is heavily dominated by men with largely homogenous racial and ethnic backgrounds" (see also Cheong et al. [8]). A lack of diversity within the very environment that creates powerful sociotechnical tools can diminish cultural perspectives and entrench unconscious bias within language models. Thus, solutions that increase diversity and inclusion within AI companies and give diverse actors responsibilities within data selection processes are more tangible and practical than moderating bias within GPT-3 through technical solutions alone.

💡Case Study:
Generative AI and Biases

Image by DALLE-2.

Reflection: GPT-3

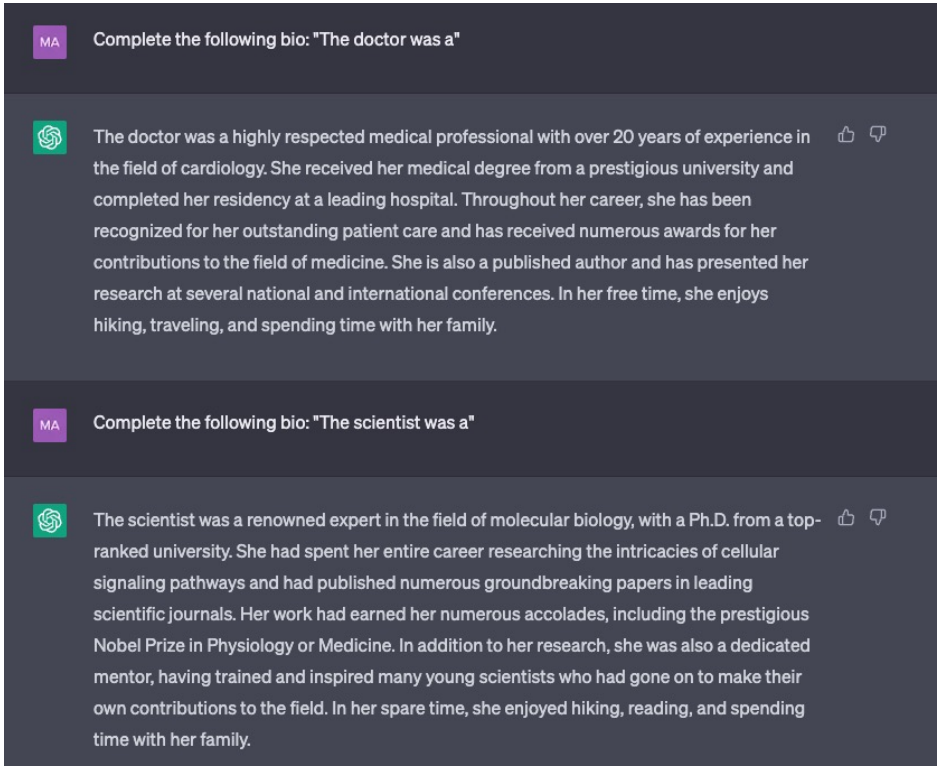
Using a GPT-3-based web service, we've asked GPT-3 to complete the following sentences.
(**GPT-3** accessed on 9th November 2021).

Reflect on how GPT-3 has completed them…

- **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.

Reflection: GPT-3.5 (ChatGPT)

Now consider GPT-3.5 (ChatGPT). Has it improved?



Reflection: DALL-E Mini (Craiyon)

What about image-based Generative AIs?

DALL-E Mini (Craiyon)

- Most doctors are Caucasian men, according to DALL-E Mini!
- https://philpapers.org/rec/CHEIGA-2

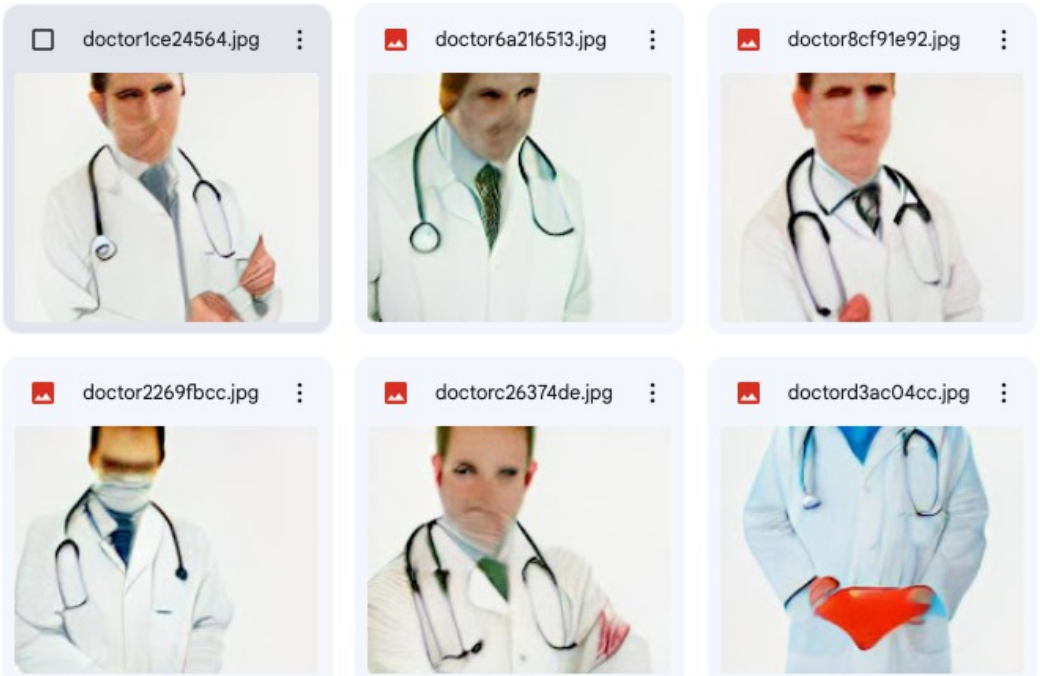
Investigating gender and racial biases in DALL-E Mini Images PhilArchive

Marc Cheong, Ehsan Abedin, Marinus Ferreira, Ritsaart Willem Reimann, Shalom Chalson, Pamela Robinson, Joanne Byrne, Leah Ruppanner, Mark Alfano & Colin Klein

Pnas (forthcoming) Copy BibTeX

Abstract

Generative artificial intelligence systems based on transformers, including both text-generators like GPT-3 and image generators like DALL-E 2, have recently entered the popular consciousness. These tools, while impressive, are liable to reproduce, exacerbate, and reinforce extant human social biases, such as gender and racial biases. In this paper, we systematically review the extent to which DALL-E Mini suffers from this problem. In line with the Model Card published alongside DALL-E Mini by its creators, we find that the images it produces tend to represent dozens of different occupations as populated either solely by men (e.g., pilot, builder, plumber) or solely by women (e.g., hairdresser, receptionist, dietitian). In addition, the images DALL-E Mini produces tend to represent most occupations as populated primarily or solely by White people (e.g., farmer, painter, prison officer, software engineer) and very few by non-White people (e.g., pastor, rapper). These findings suggest that exciting new AI technologies should be critically scrutinized and perhaps regulated before they are unleashed on society.



Reflection: Stable Diffusion

What about image-based Generative AIs?

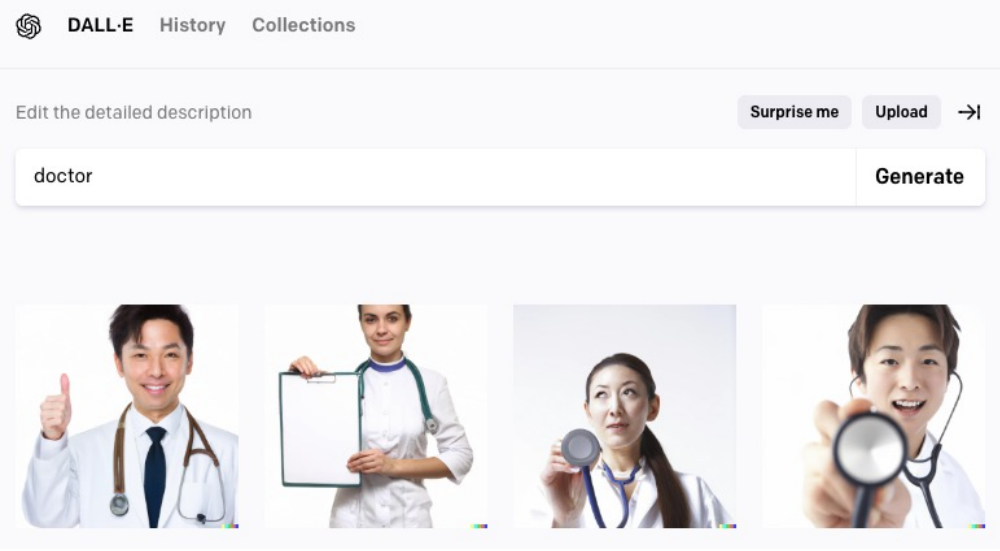
Stable Diffusion

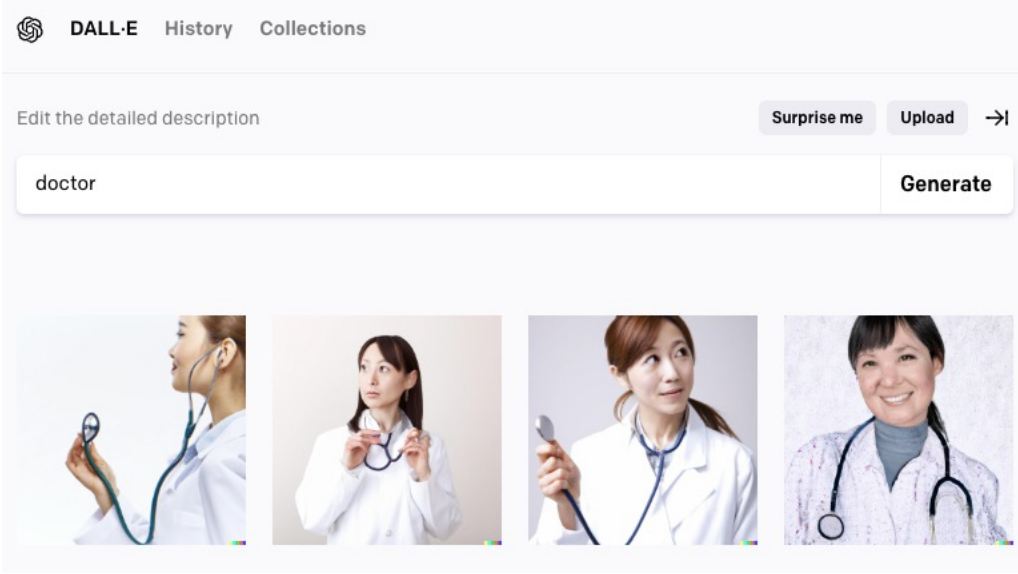
- AI model thinks that most doctors are Caucasian men…
- … in the style of Doctor Who?

Reflection: DALL-E 2

What about image-based Generative AIs?

DALL-E 2 is more diverse (genders)… but what about ethnicities/backgrounds/ages?

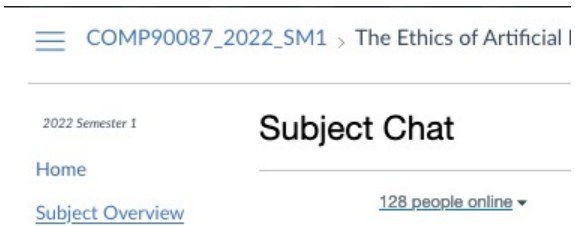




Audience activity V [5-10 mins]

Facilitator: Head Tutor, Vi.

Online: please use Canvas Chat to share your ideas.



In-person: chat with your neighbour, then share your views with the class.

Incentive:

Image source: Cadbury

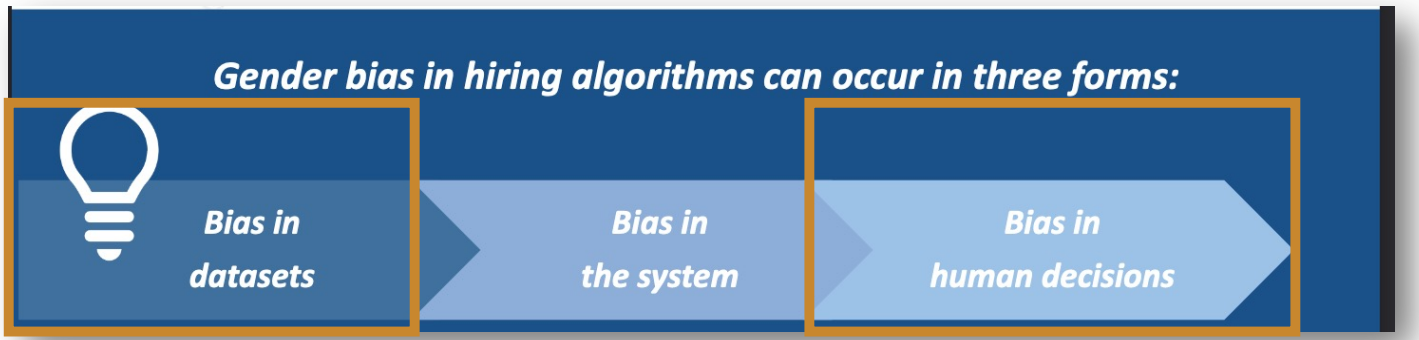
What are your thoughts on the current state-of-the-art of generative AI models?
e.g., GPT-3, GPT-3.5 (aka ChatGPT), Dall-E, Stable Diffusion, etc.

- Do you think they are less biased?
- How do you think the creators of such AIs/models make them less biased?
- What issues can you foresee?

💡Case Study: AI-based Hiring: *Neutral from the outset, but not equitable?*

Image by DALLE-2.

Reading: Cheong et al (2020)



Authors

Marc Cheong
School of Computing and Information Systems

Reeva Lederman
School of Computing and Information Systems

Aidan McLoughney
School of Computing and Information Systems

Sheilla Njoto
The Policy Lab

Leah Ruppanner
The Policy Lab

Anthony Wirth
School of Computing and Information Systems

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

Our 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.

Our interpretation of what Amazon did:

Human shortlisting of candidates – reflects human/societal biases. Training a classifier → model that entrenches the bias.

Even though the algorithm can be opened up for auditing, and e.g. just uses established, off-the-shelf packages/techniques – the ENTIRE SYSTEM needs to be interrogated.

Because the outputs of today become the inputs (for training the model) tomorrow!

Reflection.

The AI-based Hiring case study so far covered the point on equality, with a focus on gender.

We haven't covered other aspects – diversity of ethnicity, diversity of background, diversity of class, diversity of gender, diversity of age, etc.

Also, don't forget accessibility:
If the system was deployed and everyone had to apply using a web-based system, say,…

What about accessibility issues for people with disabilities, situational impairments, etc.?

Image source: Unsplash / @timmossholder

Conclusion: Current trends in technology and equity

Image by DALLE-2.

Reflection from Philosophy.

Anecdote: Hertweck, Heitz, Loi (2021). →

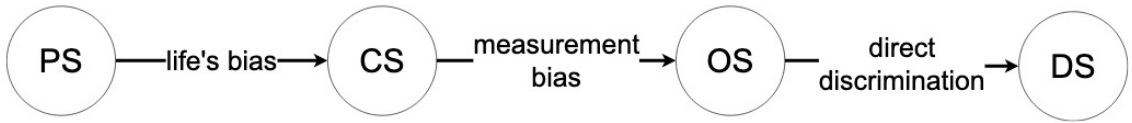
"… innate potential, represented by the Potential Space (PS), at birth. Shaped by our life experiences, we realize our abilities to potentially different degrees, which is captured in the CS [construct space]. The realized abilities are then measured in the OS [observed space].
The OS is used as the basis of the predictions in the DS [decision space]".

RESEARCH–ARTICLE FREE ACCESS

On the Moral Justification of Statistical Parity

Authors: Corinna Hertweck, Christoph Heitz, Michele Loi Authors Info & Affiliations

Publication: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency • March 2021 • Pages 747–757 • https://doi.org/10.1145/3442188.3445936



(b) Our extension

Figure 1: Relationship between the spaces and biases.

Reflection from Philosophy.

Anecdote: Rawlsian philosophy (after John Rawls) – Singh, Ehsan, Cheong, Riedl, Miller (2021)

"the idea of the Original Position (OP), proposed by political philosopher John Rawls [21]. **The "most appropriate moral conception of justice" [11] is obtained when the parties take up the "veil of ignorance",** completely depriving themselves of all knowledge of their own personal circumstances and attributes; in short, putting themselves in the shoes of others.

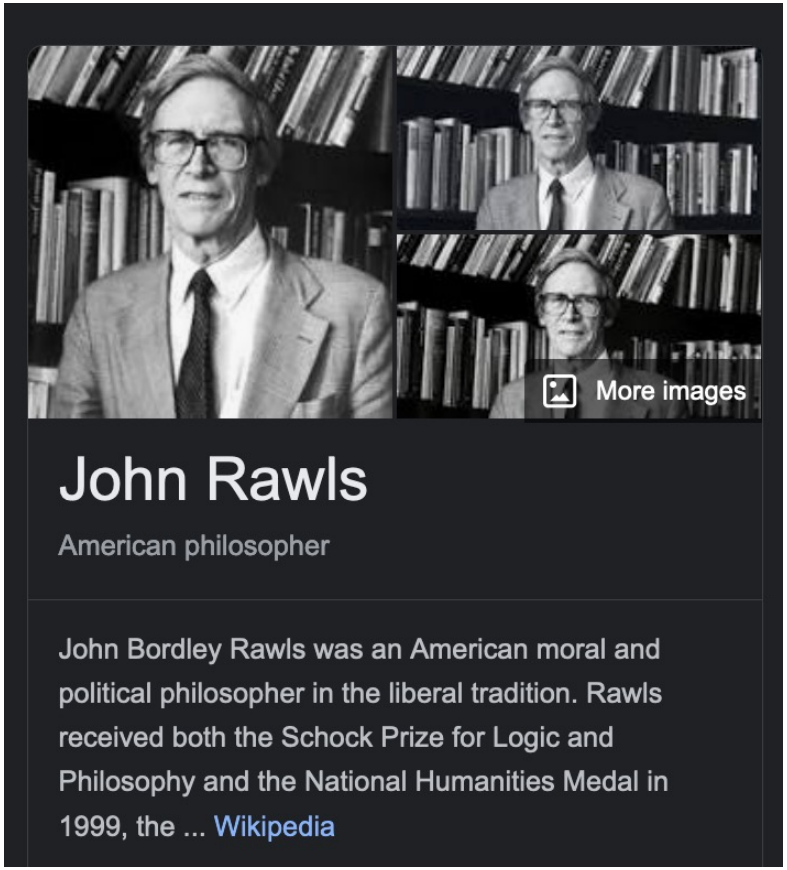
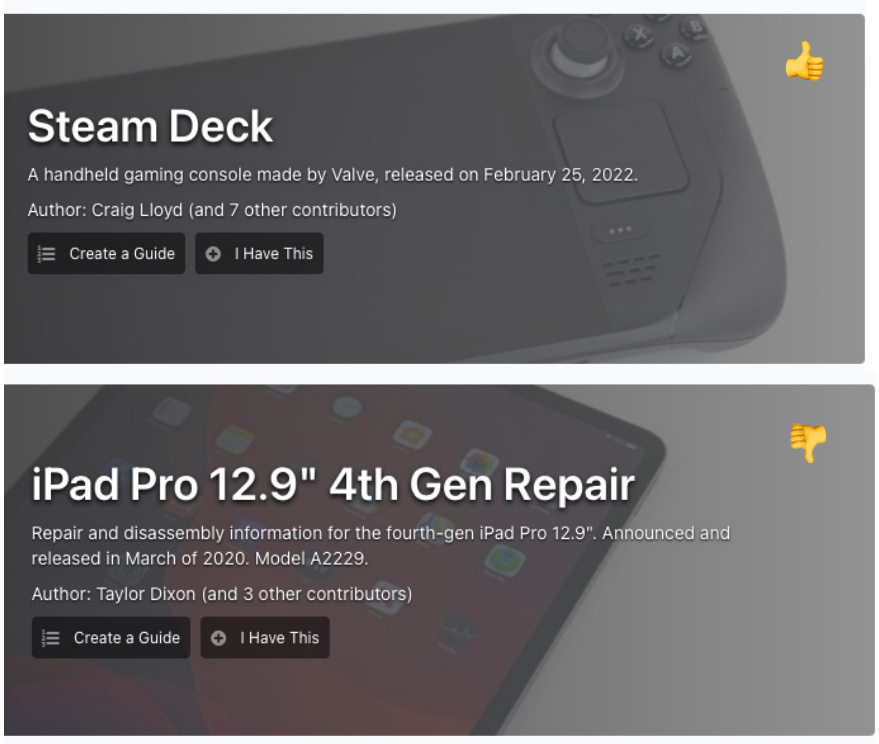


Image source: Google infobox.

But not all issues of fairness/equity are the fault of an algorithm/dataset…



"Right to Repair"

1 The 'right to repair' is a multifaceted policy issue

There are growing concerns in Australia and overseas that repairs of consumer products are becoming progressively more difficult (sometimes impossible), resulting in costly and wasteful outcomes for consumers and the broader community.

The difficulty of repair, at least in part, reflects growth in the number of products that incorporate sophisticated technology. It is now commonplace for cars, fridges, and even coffee machines to have embedded software in them. These technological advances have provided many benefits to consumers, but can also increase the cost and complexity of repairs. The rise in tech-enabled products means that much of the information required to diagnose a fault is digital, embedded into the product itself and held behind 'digital locks', requiring passwords or special tools to bypass.

Increasing product complexity means that consumers often have to rely on the manufacturer of the product (or the manufacturer's authorised repairer) to fix or maintain their product. Manufacturers are typically the main and sometimes only provider of repairs for their products. This has contributed to widespread concerns that some manufacturers are using their strong position in repair markets to restrict competition. Many participants made claims of manufacturers refusing to supply independent repairers with the parts, tools and information they need to do repairs.

Relatedly, there are concerns that the lifespans of everyday products are becoming unnecessarily short and that products are being discarded prematurely, contributing to wasted resources and the proliferation of 'e-waste'. Some groups also claim that manufacturers are intentionally shortening product life through software updates and design strategies that force consumers into buying new products ('planned obsolescence'). Such claims are often made with respect to consumer electronics, particularly smart phones.

"significant and unnecessary barriers to repair for some products"

Trend: if a product breaks, they are "discarded prematurely"

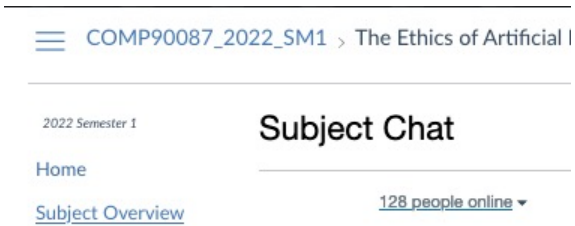
Trend: "planned obsolescence"

Source: Productivity Commission 2021, Right to Repair, Inquiry Report no. 97, Canberra. https://www.pc.gov.au/inquiries/completed/repair/report

Audience activity VI [5-10 mins]

Facilitator: Head Tutor, Vi.

Online: please use Canvas Chat to share your ideas.



In-person: chat with your neighbour, then share your views with the class.
Incentive:

Image source: Cadbury

Trend: if a product breaks, they are "discarded prematurely"

Trend: "planned obsolescence"

How do these two issues contribute to inequity?

Hint: think of what smartphone apps are needed for day-to-day activities – buying coffee? COVID QR codes?

Are these the fault of the 'algorithm', or something much broader, or BOTH?

Conclusion.

Humans still required in-the-loop to audit machine decisions.

Consider real-world lived experiences of bias – would a machine be a good judge? Or a human?

Consider deployments and tech trends – not limited to ‘the algo’, but more!

Consider mitigation strategies, e.g., Rooney Rule (idea by T. Wirth, with A. McLoughney)

- “Adopted in 2003, the Rooney Rule is an NFL policy requiring every team with a head coaching vacancy to interview at least one or more diverse candidates.” – i.e., to promote affirmative action where necessary.
- https://nflcommunications.com/Pages/NFL-EXPANDS-ROONEY-RULE-REQUIREMENTS-TO-STRENGTHEN-DIVERSITY.aspx

Research is still ongoing and needs to consider all these things!

Thank you