DeepMind

Algorithmic Fairness and the Queer Community

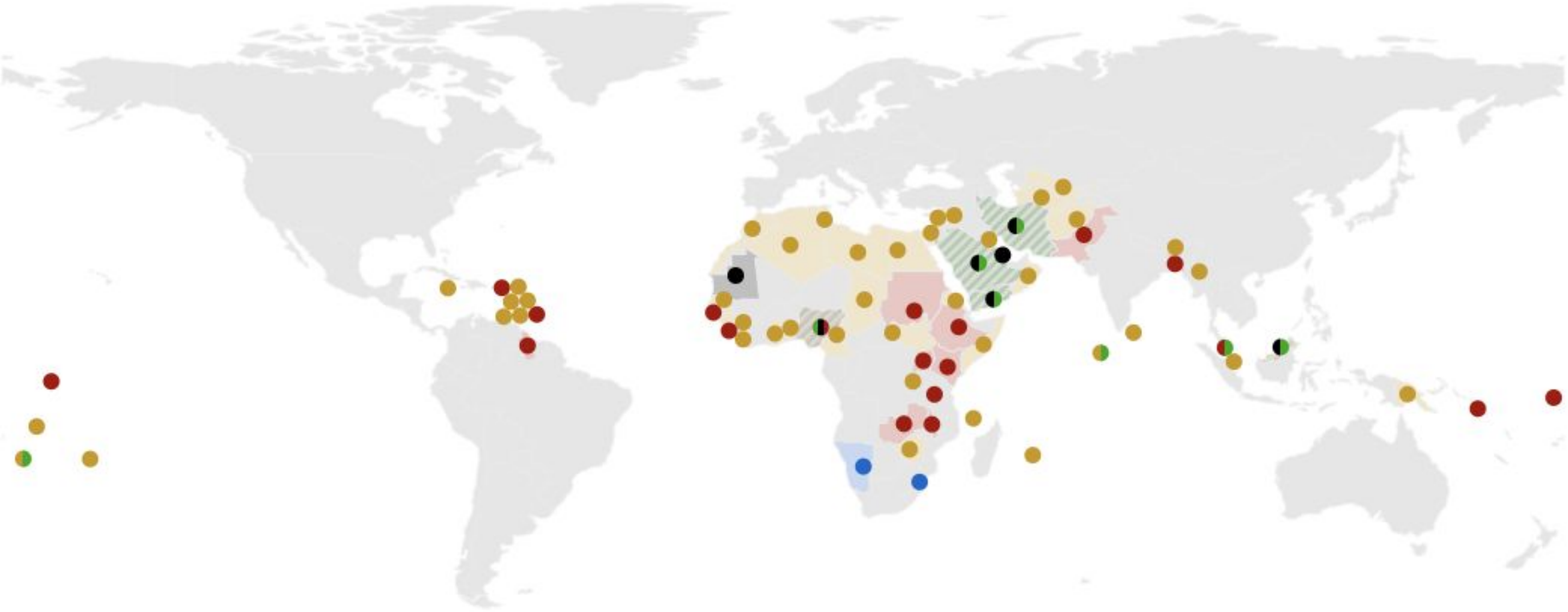
UCL AI Centre LGBTQ History Month
25/2/20
Speaker: Jackie Kay
Co-authors: Nenad Tomasev, Kevin McKee, & Shakir Mohamed

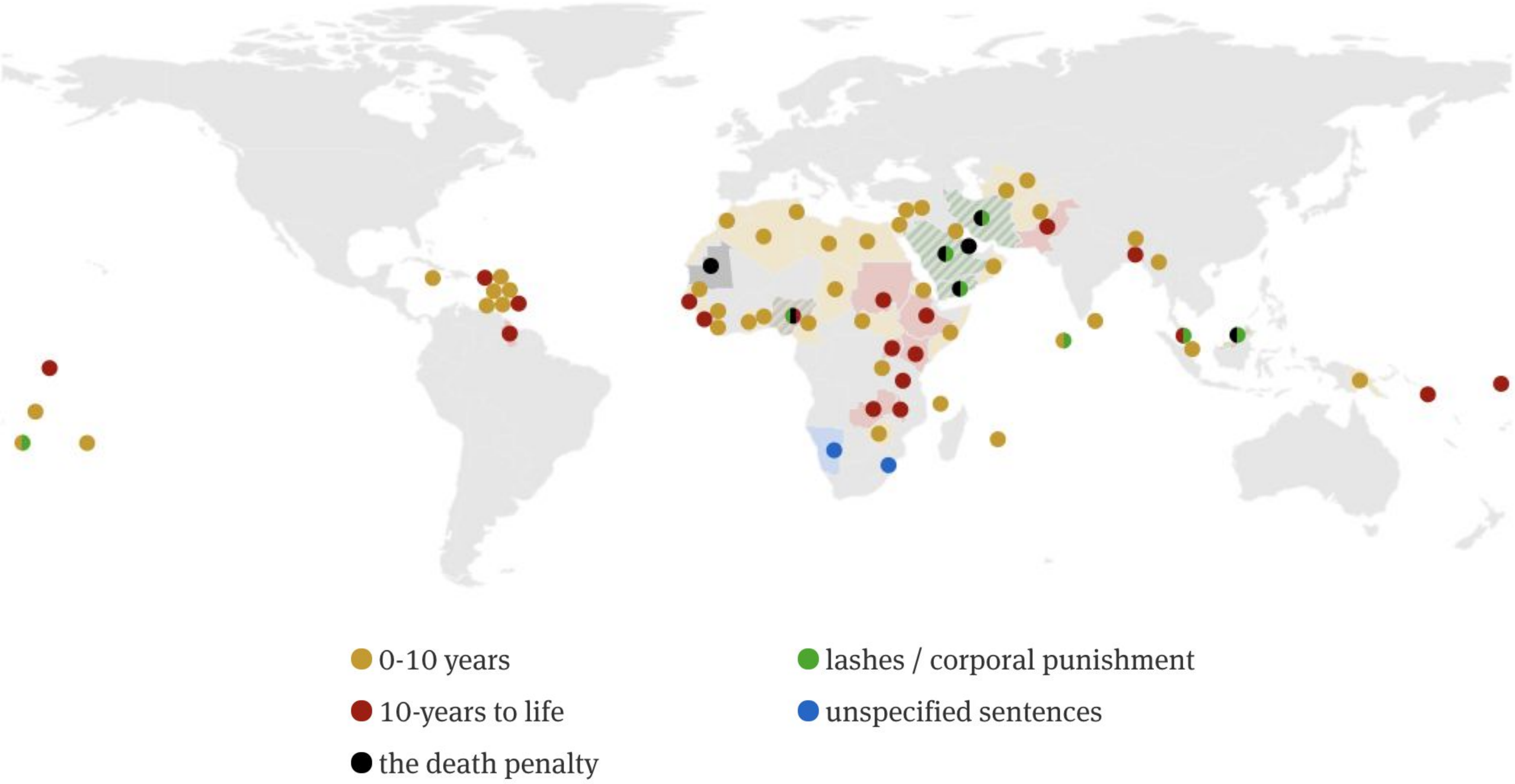
About Me

- 1st year PhD student studying machine learning for robotics
- Research engineer at DeepMind
- Nonbinary (pronouns: they/them)
- Co-author of the topic of today’s talk:
  - “Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities” (https://arxiv.org/pdf/2102.04257.pdf)

Queer AI/Queer in Tech Efforts

- Queer in AI
- Out in Tech
- Lesbians Who Tech
- Intertech LGBT+ Diversity Forum
- LGBT Technology Institute




● 0-10 years
● 10-years to life
● the death penalty
● lashes / corporal punishment
● unspecified sentences

Gathering Data on Queer Communities: Challenges

- Many countries have legislation that actively prohibits the collection of data on sexual orientation or gender identity
- Collecting data may threaten the privacy or safety of queer individuals
- Many existing datasets do not capture sexual orientation and gender identity
- Is it even possible to put an immutable, discrete label on everybody’s queerness?

Unobserved Characteristics

Fairness research has focused on characteristics that are widely available in data, or observable or identifiable by humans.

What about protected characteristics are unobserved or immeasurable?

- Many disability statuses
- Religion (often legally protected)
- Queerness

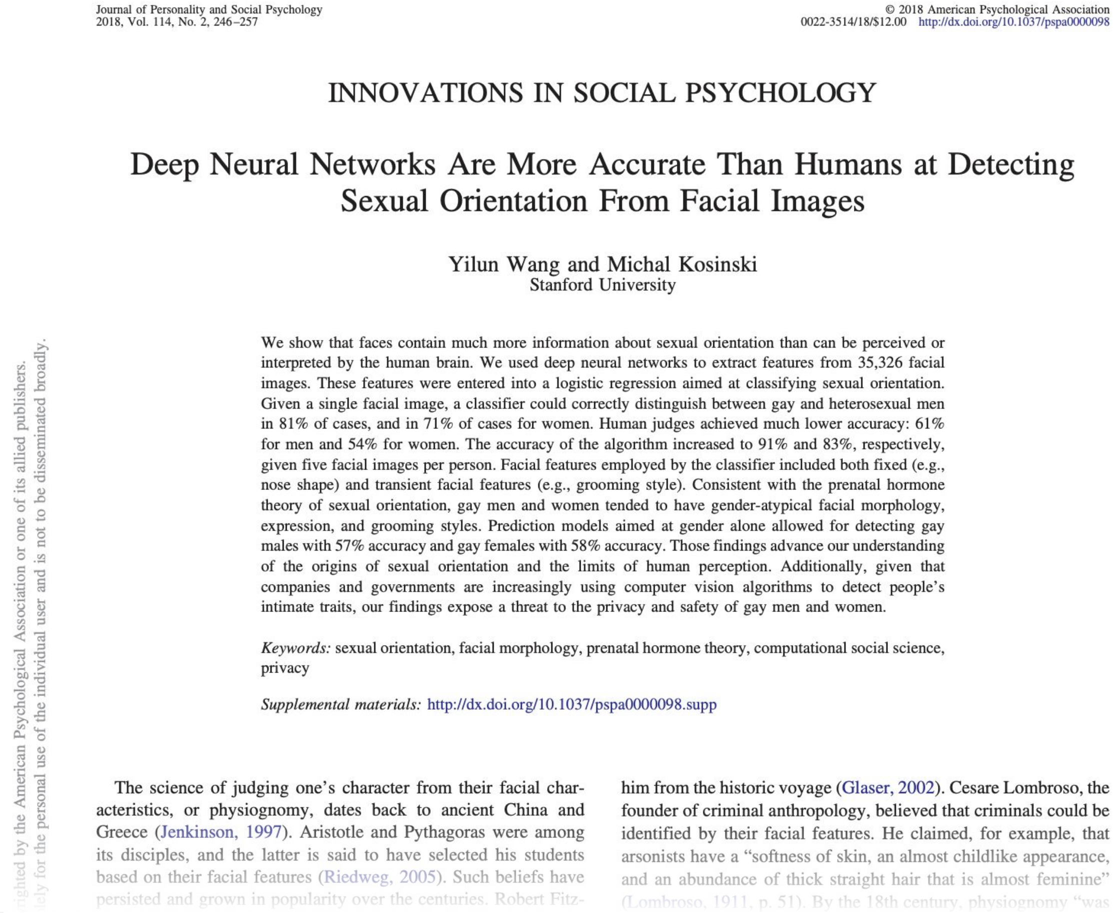
Race, ethnicity, and gender can also be unobserved, but race and (often binary) gender are more widely available in datasets.

Considerations for Queer Fairness

- How does AI interact with various issues faced by the queer community?
    - What are the potential benefits and harms?
- Where do unobserved characteristics complicate designing fair systems?
- Are privacy and fairness in tension with each other?

DeepMind

Privacy



Journal of Personality and Social Psychology
2018, Vol. 114, No. 2, 246–257

© 2018 American Psychological Association
0022-3514/18/$12.00 http://dx.doi.org/10.1037/pspa0000098

INNOVATIONS IN SOCIAL PSYCHOLOGY

Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images

Yilun Wang and Michal Kosinski
Stanford University

We show that faces contain much more information about sexual orientation than can be perceived or interpreted by the human brain. We used deep neural networks to extract features from 35,326 facial images. These features were entered into a logistic regression aimed at classifying sexual orientation. Given a single facial image, a classifier could correctly distinguish between gay and heterosexual men in 81% of cases, and in 71% of cases for women. Human judges achieved much lower accuracy: 61% for men and 54% for women. The accuracy of the algorithm increased to 91% and 83%, respectively, given five facial images per person. Facial features employed by the classifier included both fixed (e.g., nose shape) and transient facial features (e.g., grooming style). Consistent with the prenatal hormone theory of sexual orientation, gay men and women tended to have gender-atypical facial morphology, expression, and grooming styles. Prediction models aimed at gender alone allowed for detecting gay males with 57% accuracy and gay females with 58% accuracy. Those findings advance our understanding of the origins of sexual orientation and the limits of human perception. Additionally, given that companies and governments are increasingly using computer vision algorithms to detect people's intimate traits, our findings expose a threat to the privacy and safety of gay men and women.

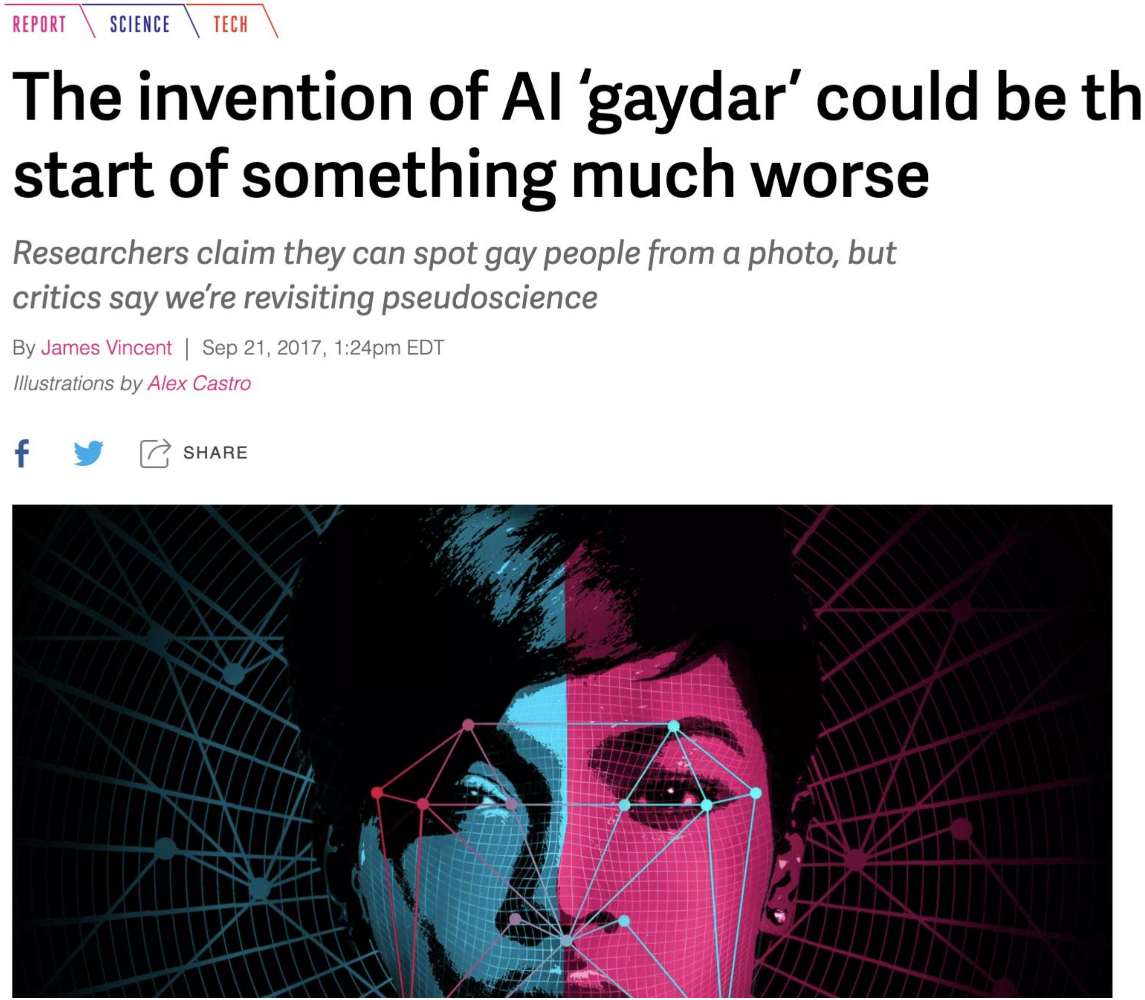
Keywords: sexual orientation, facial morphology, prenatal hormone theory, computational social science, privacy

Supplemental materials: http://dx.doi.org/10.1037/pspa0000098.supp

The science of judging one's character from their facial characteristics, or physiognomy, dates back to ancient China and Greece (Jenkinson, 1997). Aristotle and Pythagoras were among its disciples, and the latter is said to have selected his students based on their facial features (Riedweg, 2005). Such beliefs have persisted and grown in popularity over the centuries. Robert Fitz-

him from the historic voyage (Glaser, 2002). Cesare Lombroso, the founder of criminal anthropology, believed that criminals could be identified by their facial features. He claimed, for example, that arsonists have a "softness of skin, an almost childlike appearance, and an abundance of thick straight hair that is almost feminine" (Lombroso, 1911, p. 51). By the 18th century, physiognomy "was

Private & Confidential



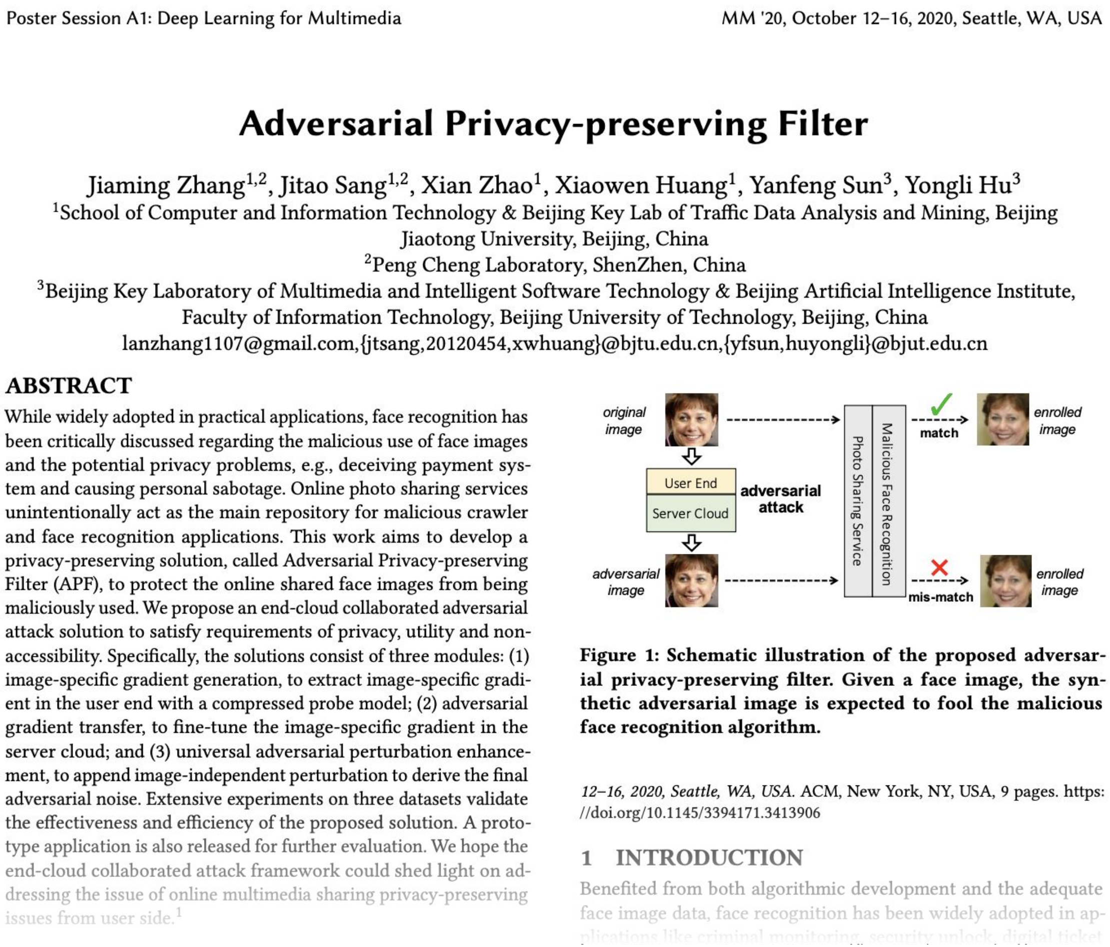
REPORT SCIENCE TECH

The invention of AI ‘gaydar’ could be the start of something much worse

Researchers claim they can spot gay people from a photo, but critics say we’re revisiting pseudoscience

By James Vincent | Sep 21, 2017, 1:24pm EDT

Illustrations by Alex Castro



Adversarial Privacy-preserving Filter

Jiaming Zhang[1,2], Jitao Sang[1,2], Xian Zhao[1], Xiaowen Huang[1], Yanfeng Sun[3], Yongli Hu[3]
[1]School of Computer and Information Technology & Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, China
[2]Peng Cheng Laboratory, ShenZhen, China
[3]Beijing Key Laboratory of Multimedia and Intelligent Software Technology & Beijing Artificial Intelligence Institute, Faculty of Information Technology, Beijing University of Technology, Beijing, China
lanzhang1107@gmail.com,{jtsang,20120454,xwhuang}@bjtu.edu.cn,{yfsun,huyongli}@bjut.edu.cn

ABSTRACT

While widely adopted in practical applications, face recognition has been critically discussed regarding the malicious use of face images and the potential privacy problems, e.g., deceiving payment system and causing personal sabotage. Online photo sharing services unintentionally act as the main repository for malicious crawler and face recognition applications. This work aims to develop a privacy-preserving solution, called Adversarial Privacy-preserving Filter (APF), to protect the online shared face images from being maliciously used. We propose an end-cloud collaborated adversarial attack solution to satisfy requirements of privacy, utility and non-accessibility. Specifically, the solutions consist of three modules: (1) image-specific gradient generation, to extract image-specific gradient in the user end with a compressed probe model; (2) adversarial gradient transfer, to fine-tune the image-specific gradient in the server cloud; and (3) universal adversarial perturbation enhancement, to append image-independent perturbation to derive the final adversarial noise. Extensive experiments on three datasets validate the effectiveness and efficiency of the proposed solution. A prototype application is also released for further evaluation. We hope the end-cloud collaborated attack framework could shed light on addressing the issue of online multimedia sharing privacy-preserving issues from user side.[1]

Figure 1: Schematic illustration of the proposed adversarial privacy-preserving filter. Given a face image, the synthetic adversarial image is expected to fool the malicious face recognition algorithm.

12–16, 2020, Seattle, WA, USA. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413906

1 INTRODUCTION

Benefited from both algorithmic development and the adequate face image data, face recognition has been widely adopted in applications like criminal monitoring, security unlock, digital ticket

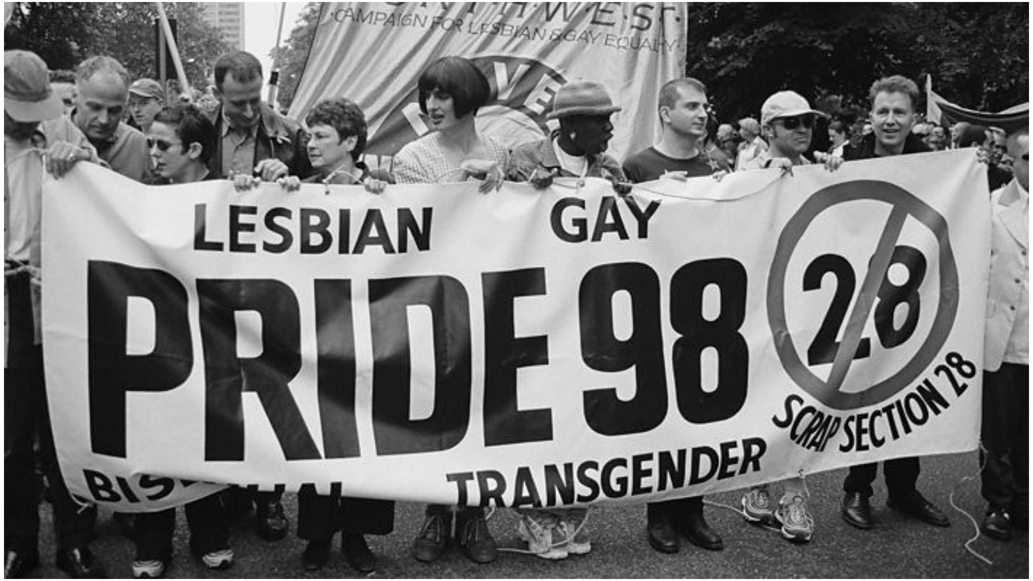
Privacy: summary points

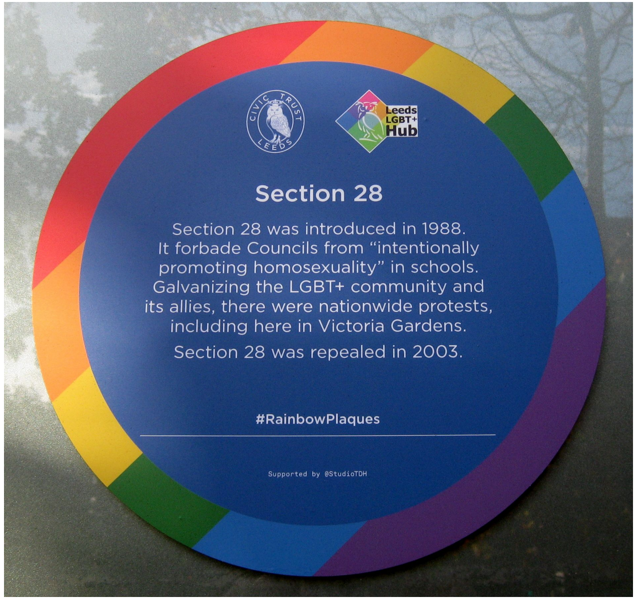
- 2017 Stanford paper criticised for methodological shortcomings (and ethical issues around its messaging)
    - But future algorithms might circumvent these methodological issues
    - Large amounts of online data that can be scraped and used
- **Risks**
    - Risks to physical safety, social and psychological well-being
    - Human rights issue in nations where homosexuality is criminalized, surveillance risks
- **Promise**
    - Adversarial filters that can obfuscate sensitive information in images and speech shared online
    - Reduce the risks of re-identification

DeepMind

Censorship

London Pride, 1998, photo by Steve Eason, Getty Images





Section 28

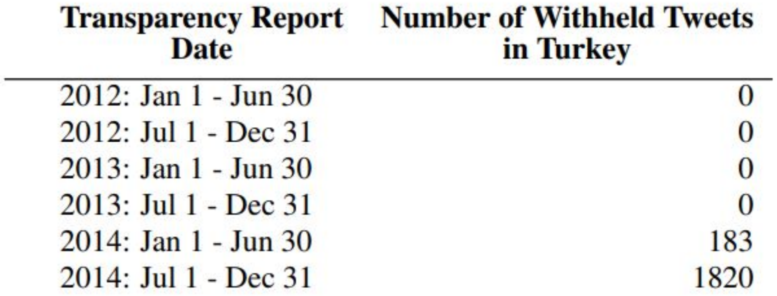
Section 28 was introduced in 1988.
It forbade Councils from "intentionally promoting homosexuality" in schools.
Galvanizing the LGBT+ community and its allies, there were nationwide protests, including here in Victoria Gardens.

Section 28 was repealed in 2003.

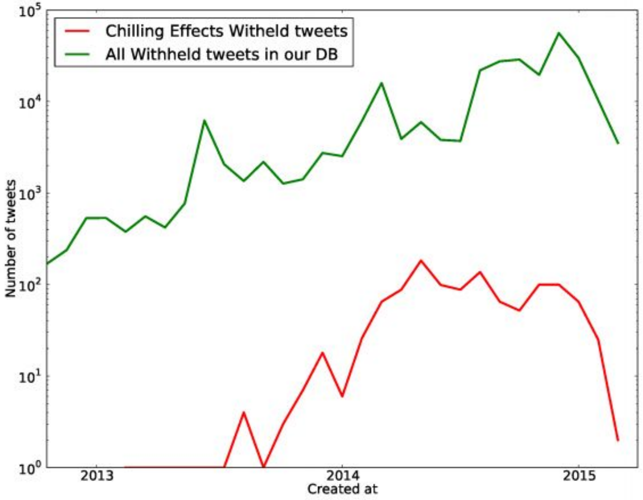
#RainbowPlaques

Supported by @StudioTDH

Leeds Rainbow Plaque Trail



Transparency Report Date	Number of Withheld Tweets in Turkey
2012: Jan 1 - Jun 30	0
2012: Jul 1 - Dec 31	0
2013: Jan 1 - Jun 30	0
2013: Jul 1 - Dec 31	0
2014: Jan 1 - Jun 30	183
2014: Jul 1 - Dec 31	1820



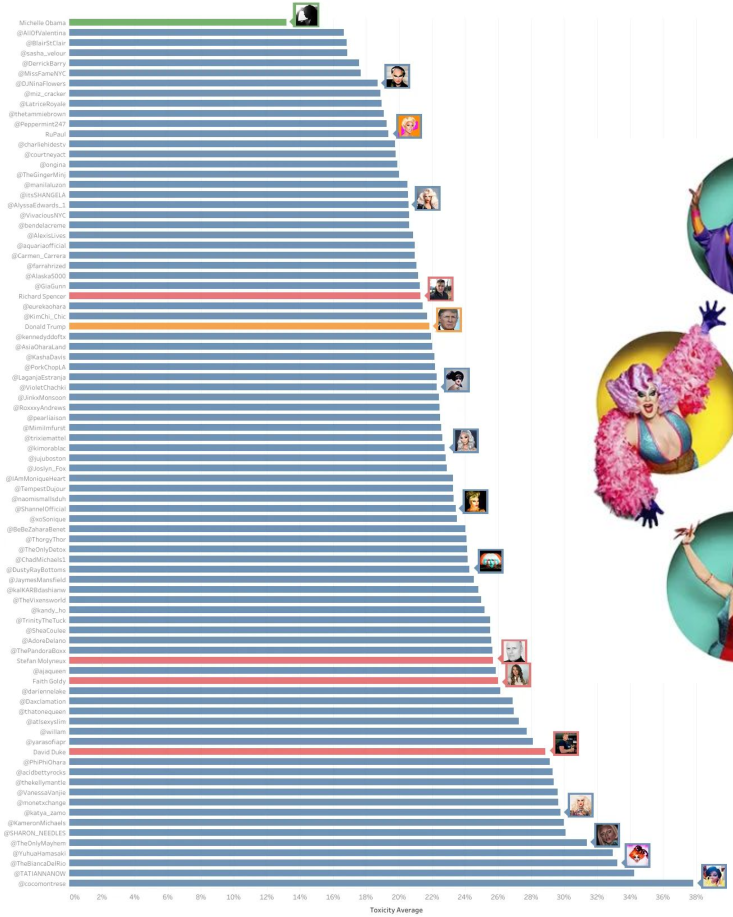
Tanash 2015, “Known Unknowns: An Analysis of Twitter Censorship in Turkey”

Censorship

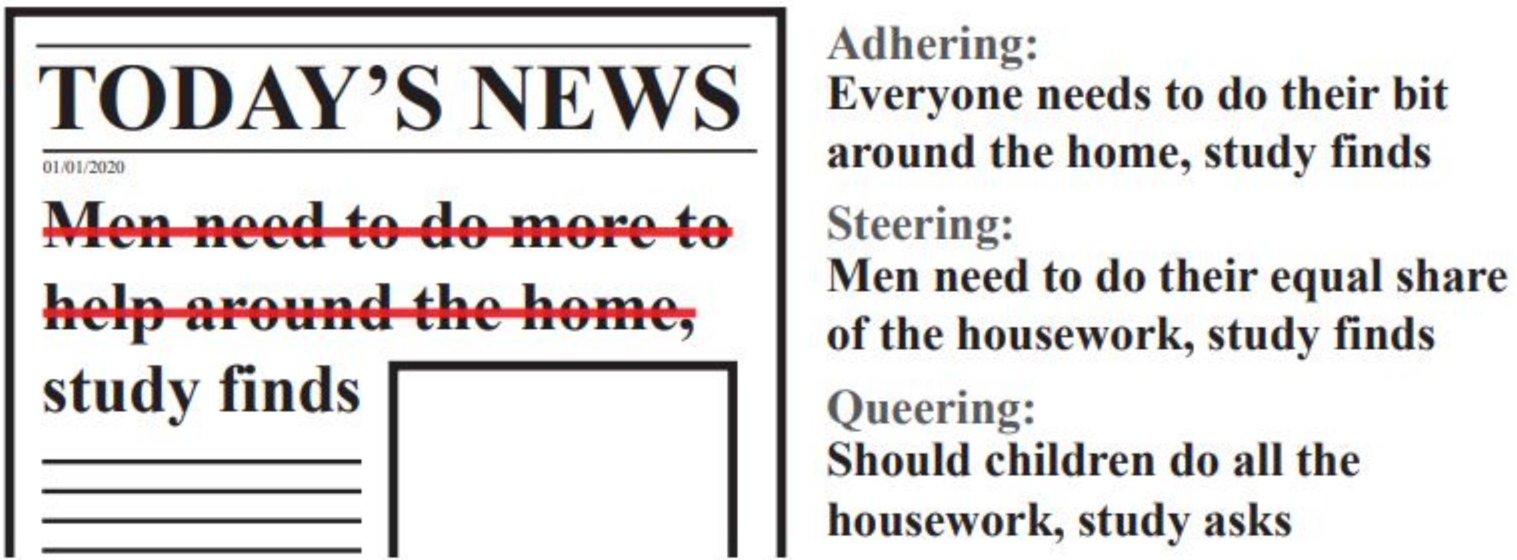
- Historically and in present day, queer media content has been censored by law in many nations
- Censorship implicitly shames queerness, isolates us from our communities, and leads to identity erasure
- **Risks**
  - Large-scale censorship enabled through automated classification of queer content through natural language and images
  - Deepfakes and generative language models could generate misinformation to demonise queer groups as justification for marginalisation
- **Promise**
  - Machine learning has been used to detect anomalous deletion of censored content, but has not been applied to queer censorship yet

DeepMind

Language & Content Moderation



Gomes, Antonialli, Oliva 2019,
“Drag Queens and Artificial Intelligence”



TODAY'S NEWS

01/01/2020

~~Men need to do more to help around the home,~~ study finds

Adhering:
Everyone needs to do their bit around the home, study finds

Steering:
Men need to do their equal share of the housework, study finds

Queering:
Should children do all the housework, study asks

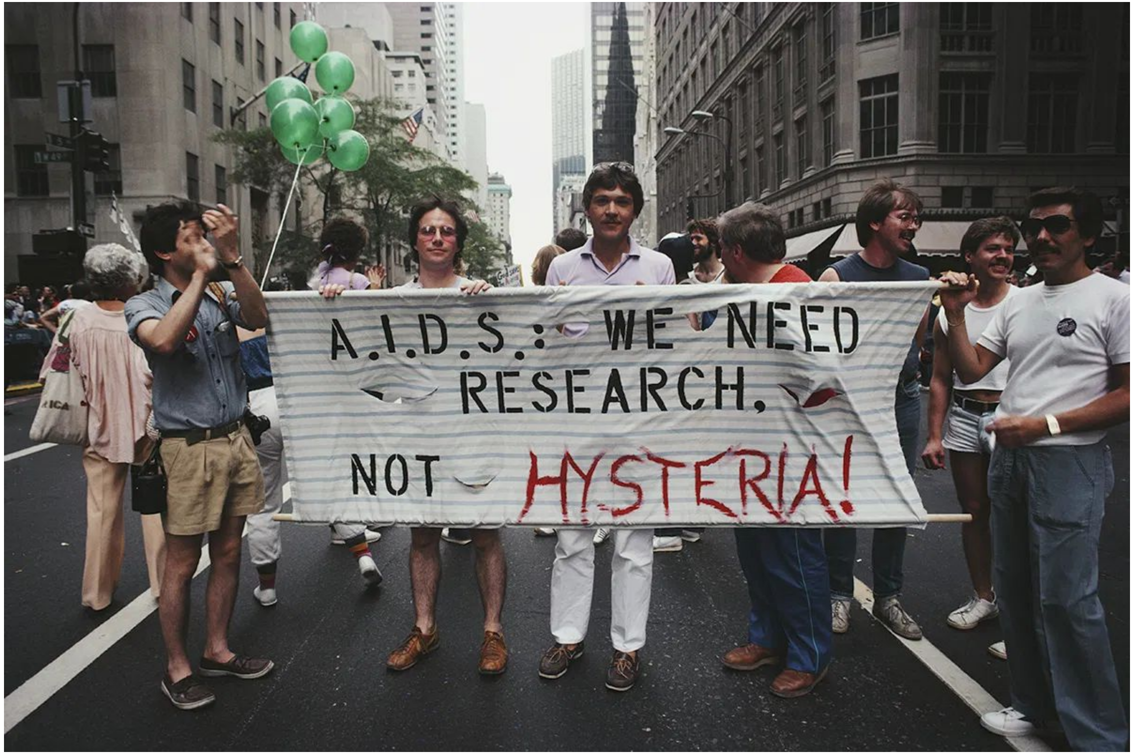
Strangers 2020,
“Adhering, Steering, and Queering: Treatment of Gender in Natural Language Generation”

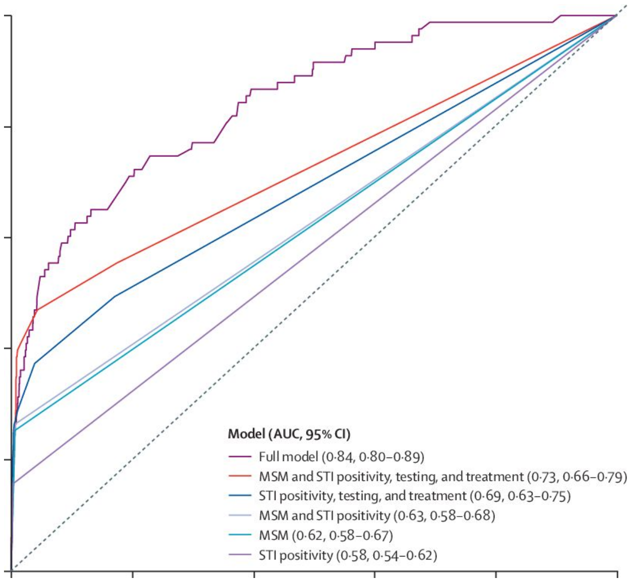
Language & Content Moderation

- **Promise**
  - Queering natural language generation e.g. by aligning with feminist HCI qualities (Strengers 2020)
  - Detecting deadnaming and misgendering
  - Content moderation: detecting homophobic and transphobic hate speech
- **Risks**
  - Accidental censorship of queer content thought to be provocative ()
    - Words such as “queer”, “gay”, “butch” etc. were/have been considered offensive
  - Overarching challenge: how to detect the toxicity of language in context, whether the usage is insulting or empowering?
  - Intersectionality: how to ensure fairness when homophobic stereotypes or language are also racialized or associated with other elements of identity?

DeepMind

Health & Mental Health





Left: New York Pride Parade, 1983, photo by Barbara Alper. Right: Marcus 2019, “Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis”

THE TREVOR PROJECT
Saving Young LGBTQ Lives

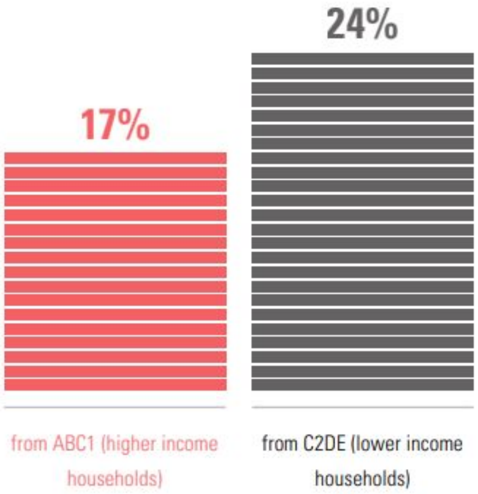
Health & Mental Health

- Queer communities disproportionately affected by substance abuse, higher transmission of STIs and HIV, depression, anxiety, and suicide ideation
- Barriers to care: isolation from family and support networks, discrimination in the medical system…
- **Promise**
  - AI for medicine: develop new treatments
  - Predict when patients will benefit from treatment via medical data and background
  - Identify patient’s affect and suicide risks before connecting to counselor (Trevor Project)
- **Risks**
  - Automated medical/mental health interventions may exacerbate existing health care biases
  - Sensitive data (HIV status, sexual orientation, mental health) often omitted from datasets, attempting to identify is often a privacy risk
  - Testing interventions in a reinforcement learning setting requires safe exploration guarantees

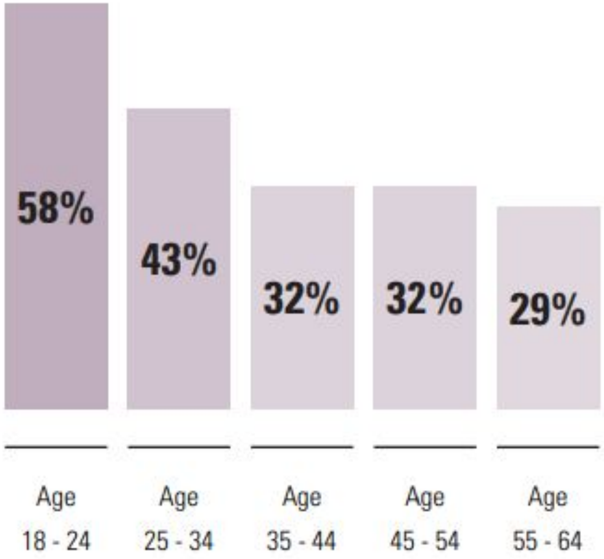
DeepMind

Employment

LGBT staff who have been the target of negative comments or conduct from work colleagues because they are LGBT



LGBT staff who have hidden or disguised that they're LGBT at work because they were afraid of discrimination



Stonewall 2018,
"LGBT in Britain: Work Report"

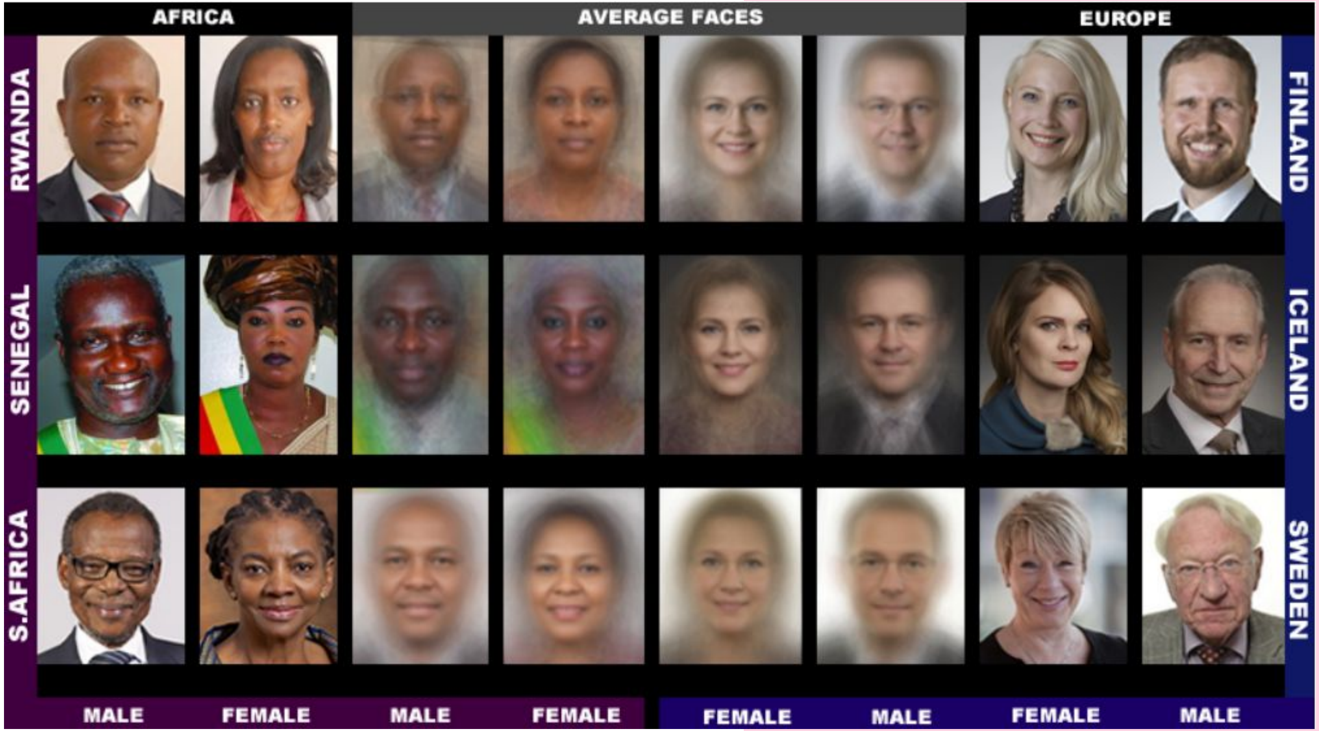
Employment

- Coming out often results in hiring bias and workplace discrimination
- Human graders more likely to give lower scores to resumes with LGBT associations in LeCroy 2019
- **Risks**
  - If AI fairness not implemented correctly, could exacerbate existing bias
  - Social media scrapers used by employers for personality assessment may risk outing candidates or employees
- **Promise**
  - Fair AI systems could identify discrimination in the hiring process and even suggest interventions for changing the employment pipeline

DeepMind

Where to go from here?

Algorithmic Fairness and Machine Learning



Buolamwini 2018, Gender Shades

Expanding AI Fairness Techniques

- Demographic parity
  - Requires group membership information
- Frameworks for fairness based on different metrics
  - Individual fairness: treat similar individuals similarly

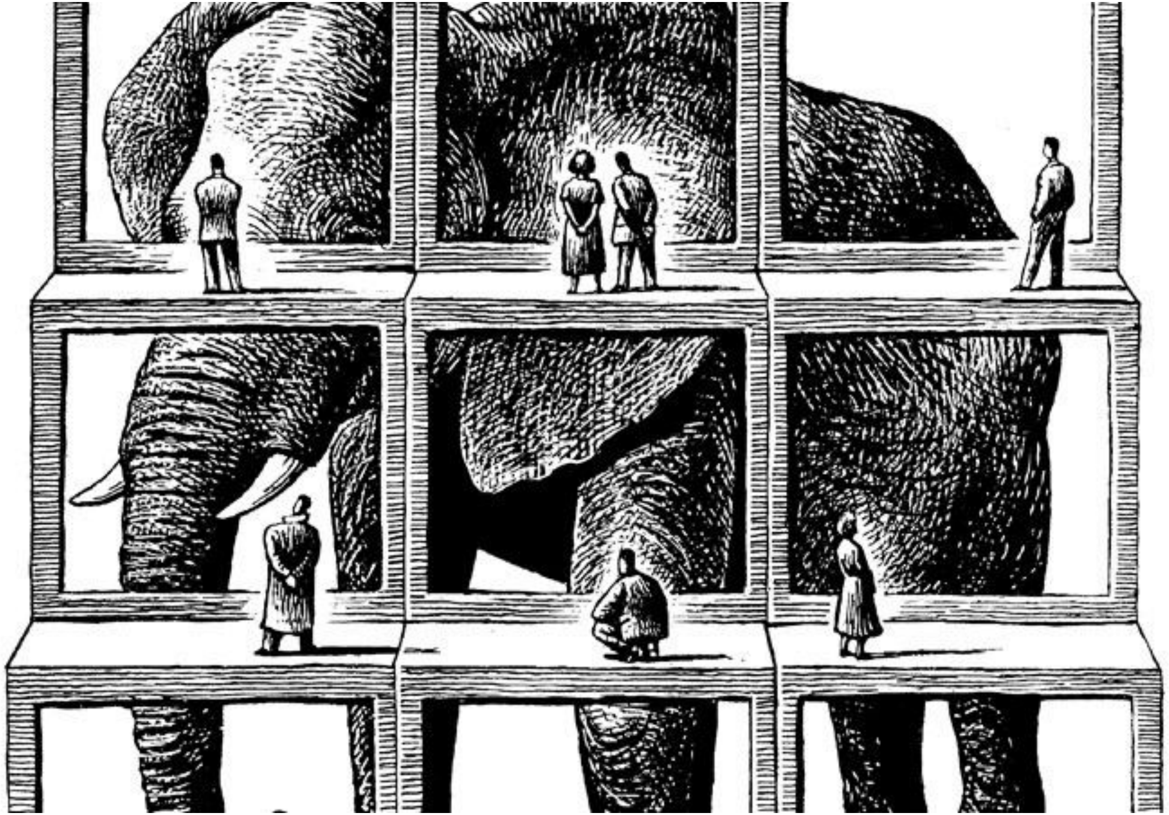
$$P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a)$$

  - Counterfactual fairness: infer latent causes and enforce fairness across alternative conditions
  - Contrastive fairness: extends counterfactual fairness, asks if it is fair to select different decisions for different individuals

$$P(\hat{Y}_{A_j\leftarrow a_i}(U_j) = d' \mid X_j = x_j, A_j = a_j)P(\hat{Y}_{A_j\leftarrow a_i}(U_j) = d \mid X_j = x_j, A_j = a_j)$$

- Adversarially reweighted learning
  - Addresses fairness for protected attributes using measurable covariates for those attributes
  - Maximize the minimum expected utility across groups with different attributes

It all depends on context



Conclusion

- Queer communities have surmounted enormous historical oppression and continue to do so today
- AI has enormous potential
- New technical frameworks in AI fairness are needed to mitigate risks and explore promise
- AI researchers responsible dignity and well-being of all people who interact with the systems we design
- The field needs a diversity of voices and backgrounds
- Want to explore these issues with us? Get in touch!
  - kayj@google.com

Image from The Met Office