Chapter 8: Computer Reliability and Ethics of AI

COMP422 Ethics and Professional Issues in Computing
Dr. Patrick Pang

Based on
Ethics for the Information Age (5th Ed.)
by
Michael J. Quinn

澳門理工學院
Instituto Politécnico de Macau
Macao Polytechnic Institute

Chapter Overview

- Introduction
- Data-entry or data-retrieval errors
- Software and billing errors
- AI Ethics
- Computer simulations
- Software engineering
- Software warranties

8.1 Introduction

- Computer systems are sometimes unreliable
  - Erroneous information in databases
  - Misinterpretation of database information
  - Malfunction of embedded systems
- Effects of computer errors
  - Inconvenience
  - Bad business decisions
  - Fatalities

8.2 Data-Entry or Data-Retrieval Errors

Two Kinds of Data-related Failure

- A computerized system may fail because wrong data entered into it
- A computerized system may fail because people incorrectly interpret data they retrieve

Disfranchised Voters

- November 2000 general election
- Florida disqualified thousands of voters
- Reason: People identified as felons
- Cause: Incorrect records in voter database
- Consequence: May have affected election’s outcome

False Arrests

- Three cases of false arrests due to incorrect information retrieved from the NCIC
- Sheila Jackson Stossier mistaken for Shirley Jackson
  - Arrested and spent five days in detention
- Roberto Hernandez mistaken for another Roberto Hernandez
  - Arrested twice and spent 12 days in jail
- Terry Dean Rogan arrested after someone stole his identity
  - Arrested five times, three times at gun point

Accuracy of NCIC Records

- March 2003: Justice Dept. announces FBI not responsible for accuracy of NCIC information
- Exempts NCIC from some provisions of Privacy Act of 1974
- Should government take responsibility for data correctness?

Dept. of Justice Position

- Impractical for FBI to be responsible for data’s accuracy
- Much information provided by other law enforcement and intelligence agencies
- Agents should be able to use discretion
- If provisions of Privacy Act strictly followed, much less information would be in NCIC
- Result: fewer arrests

Position of Privacy Advocates

- Number of records is increasing
- More erroneous records → more false arrests
- Accuracy of NCIC records more important than ever

Analysis: Database of Stolen Vehicles

- > 1 million cars stolen every year
  - Owners suffer emotional, financial harm
  - Raises insurance rates for all
- Transporting stolen car across a state line
  - Before NCIC, greatly reduced chance of recovery
  - After NCIC, nationwide stolen car retrieval
- At least 50,000 recoveries annually due to NCIC
- Few stories of faulty information causing false arrests
- Benefit > harm → Creating database the right action

8.3 Software and Billing Errors

Errors When Data Are Correct

- Assume data correctly fed into computerized system
- System may still fail if there is an error in its programming

Analysis: E-Retailer Posts Wrong Price, Refuses to Deliver

- Amazon.com in Britain offered iPaq for £7 instead of £275
- Orders flooded in
- Amazon.com shut down site, refused to deliver unless customers paid true price
- Was Amazon.com wrong to refuse to fill the orders?

Rule Utilitarian Analysis

- Imagine rule: A company must always honor the advertised price
- Consequences
  - More time spent proofreading advertisements
  - Companies would take out insurance policies
  - Higher costs → higher prices
  - All consumers would pay higher prices
  - Few customers would benefit from errors
- Conclusion
  - Rule has more harms than benefits
  - Amazon.com did the right thing

Kantian Analysis

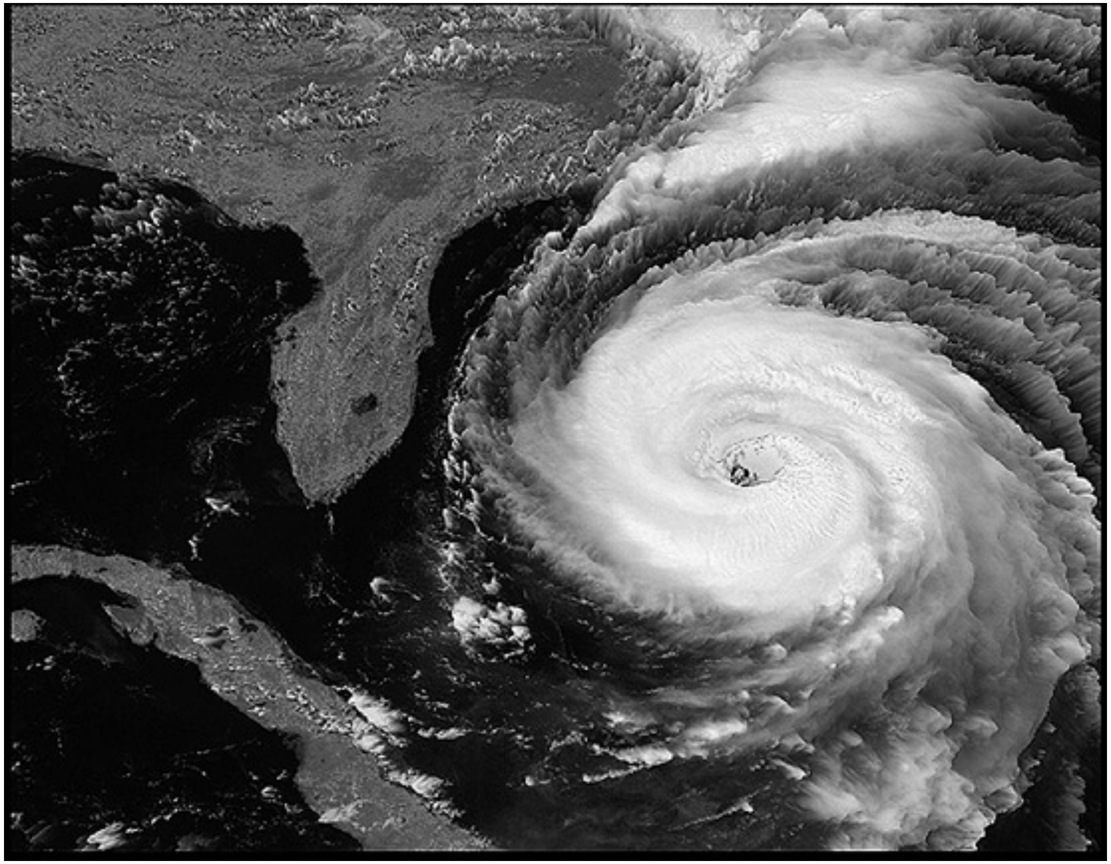
- Buyers knew 97.5% markdown was an error
- They attempted to take advantage of Amazon.com’s stockholders
- They were not acting in “good faith”
- Buyers did something wrong

8.4 Computer Simulations

Uses of Simulations

- Simulations replace physical experiments
  - Experiment too expensive or time-consuming
  - Experiment unethical
  - Experiment impossible
- Model past events
- Understand world around us
- Predict the future

Simulations Predict Path and Speed of Hurricanes

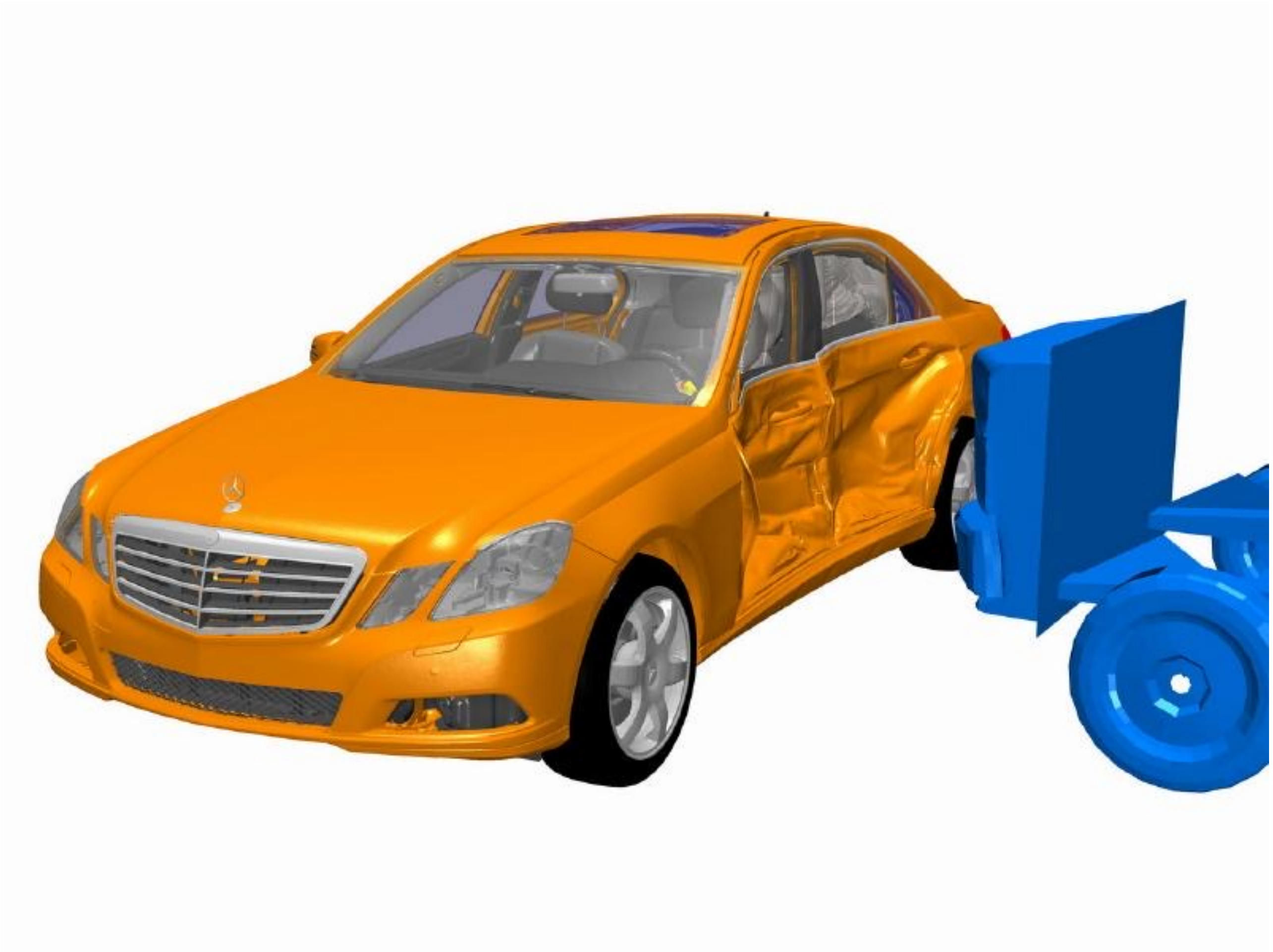


Courtesy of NASA

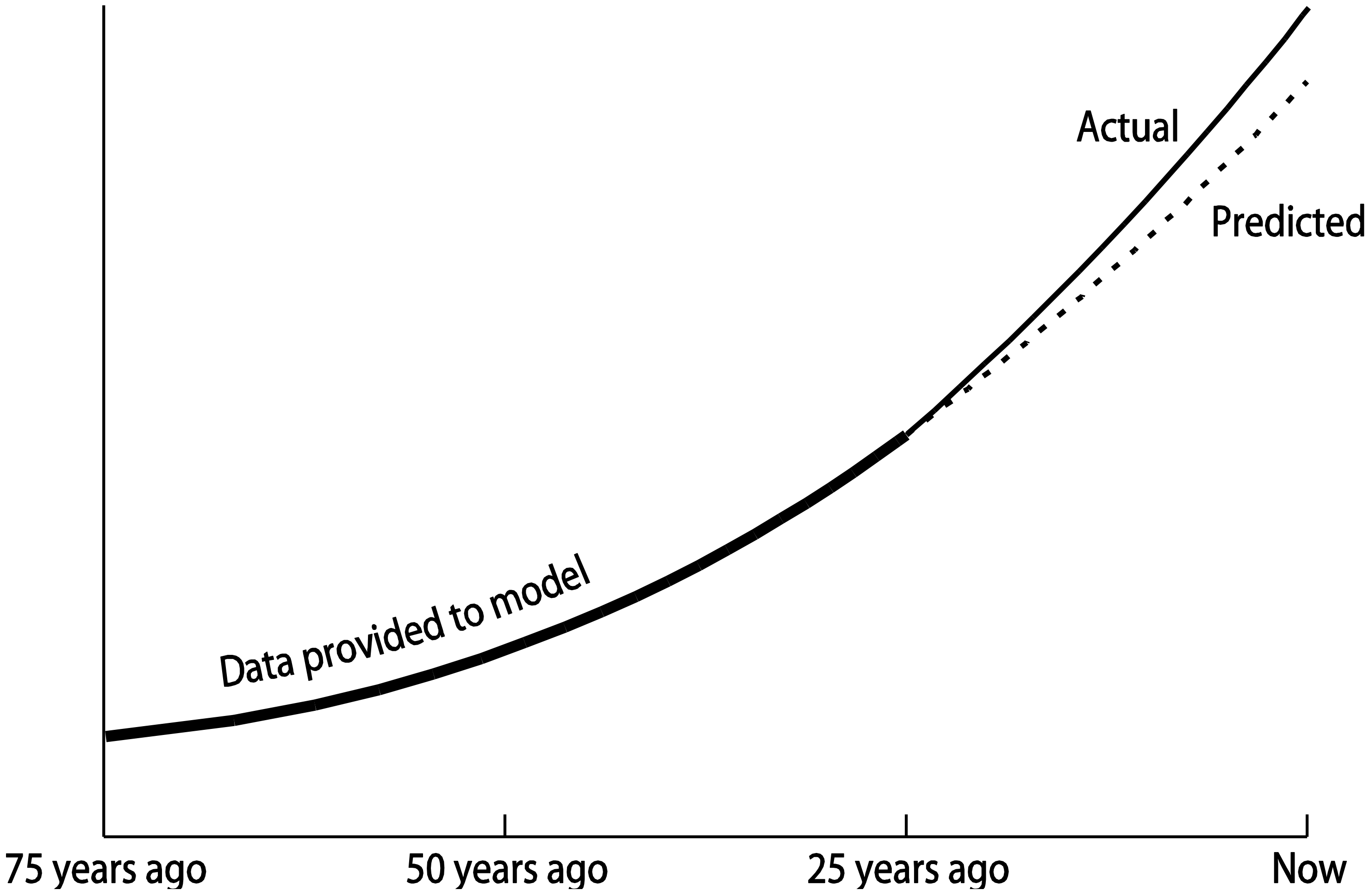
Validating Simulations

- Verification: Does program correctly implement model?
- Validation: Does the model accurately represent the real system?
- Validation methods
  - Make prediction, wait to see if it comes true
  - Predict the present from old data
  - Test credibility with experts and decision makers

Validation by Comparing Predicted and Actual Outcomes



Validation by "Predicting the Present"



8.5 Software Engineering

Four-step Process to Develop a Software Product

- Specification
  - defining the functions to be performed by the software
- Development
  - producing the software that meets the specification
- Validation
  - testing the software
- Evolution
  - modifying the software to meet the changing needs of the customer

Specification

- Determine system requirements
- Understand constraints
- Determine feasibility
- End products
  - High-level statement of requirements
  - Mock-up of user interface
  - Low-level requirements statement

Development

- Create high-level design
- Discover and resolve mistakes, omissions in specification
- CASE tools to support design process
- Object-oriented systems have advantages
- After detailed design, actual programs written
- Result: working software system

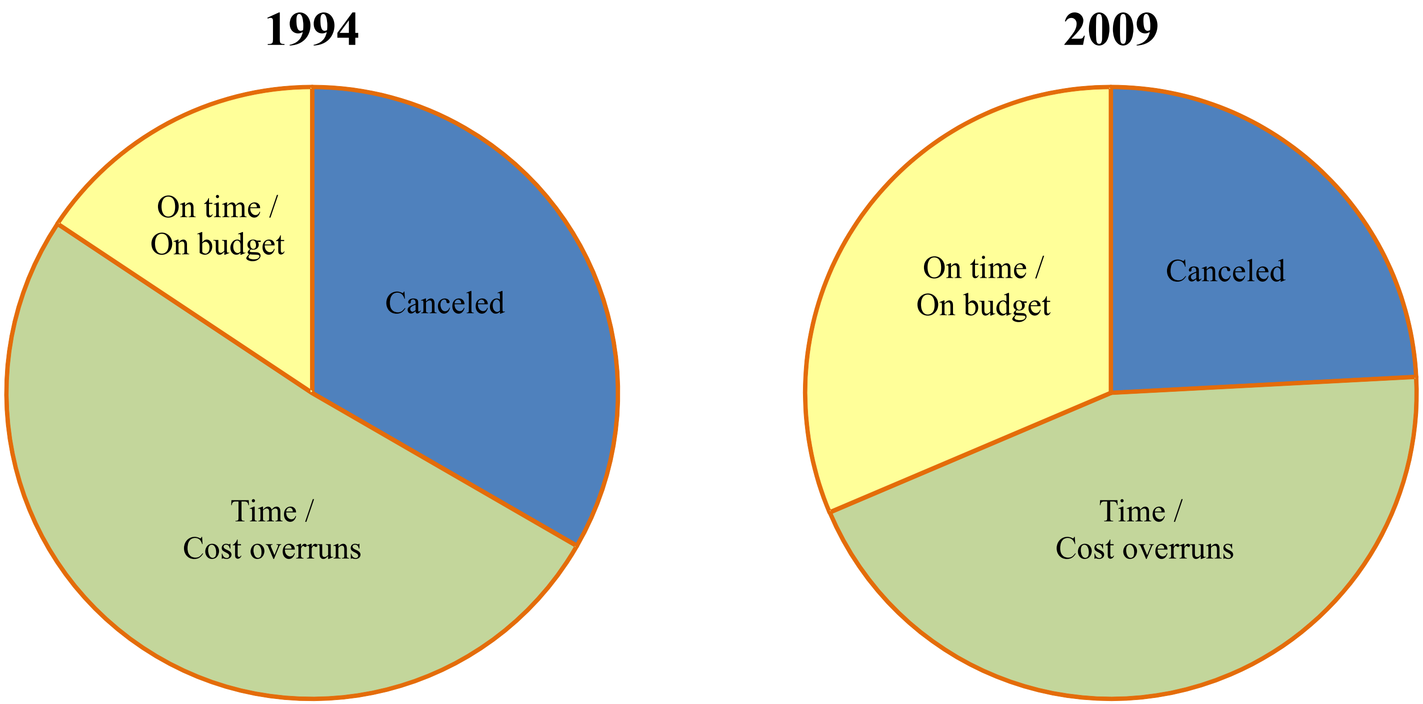
Validation (Testing)

- Ensure software satisfies specification
- Ensure software meets user’s needs
- Challenges to testing software
  - Noncontinuous responses to changes in input
  - Exhaustive testing impossible
  - Testing reveals bugs, but cannot prove none exist
- Test modules, then subsystems, then system

Software Quality Is Improving

- Standish Group tracks IT projects
- Situation in 1994
  - 1/3 projects cancelled before completion
  - 1/2 projects had time and/or cost overruns
  - 1/6 projects completed on time / on budget
- Situation in 2009
  - 1/6 projects cancelled
  - 1/2 projects had time and/or cost overruns
  - 1/3 projects completed on time / on budget

Success of IT Projects Over Time



8.6 Software Warranties

Shrinkwrap Warranties

- Some say you accept software “as is”
- Some offer 90-day replacement or money-back guarantee
- None accept liability for harm caused by use of software

Moral Responsibility of Software Manufacturers

- If vendors were responsible for harmful consequences of defects
  - Companies would test software more
  - They would purchase liability insurance
  - Software would cost more
  - Start-ups would be affected more than big companies
  - Less innovation in software industry
  - Software would be more reliable
- Making vendors responsible for harmful consequences of defects may be wrong, but…
- Consumers should not have to pay for bug fixes

8.7 Ethics of AI

Parts of this section is based on materials by:

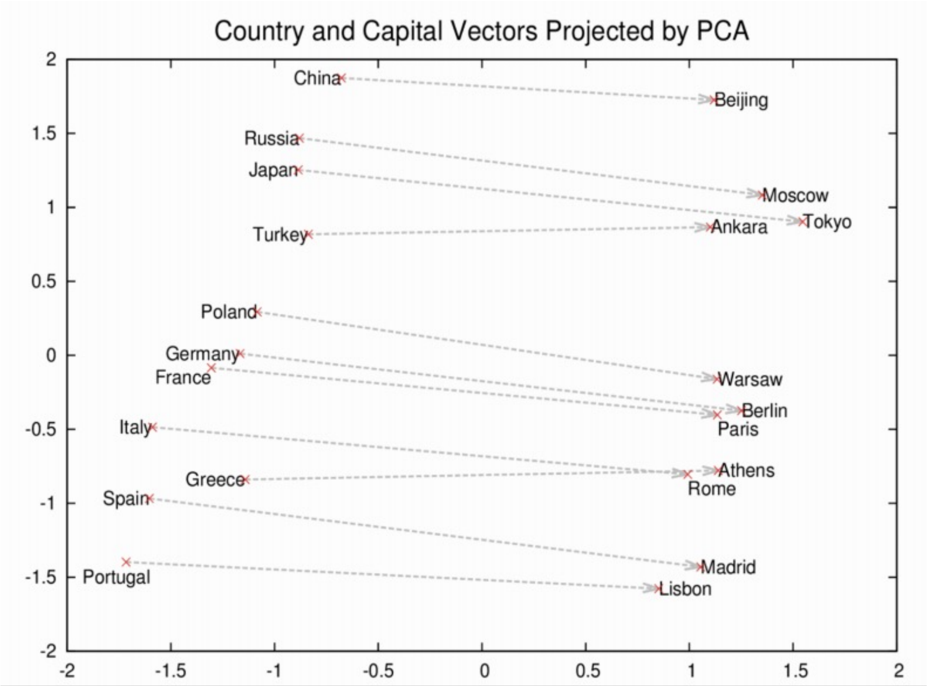
Weber, *“Computational Social Science: Theories, Methods and Data”*

Some questions

- Are algorithms “neutral”? Should they be neutral?
- Popularity-driven algorithms = Cruelty of the majority?
- How do algorithms change what we think, whom we befriend?

Example: word2vec

- Word2vec by Google: word embedding based on a two-layer neural network, summarizing its typical context
- Trained on large text corpora (Wikipedia, WSJ, …)



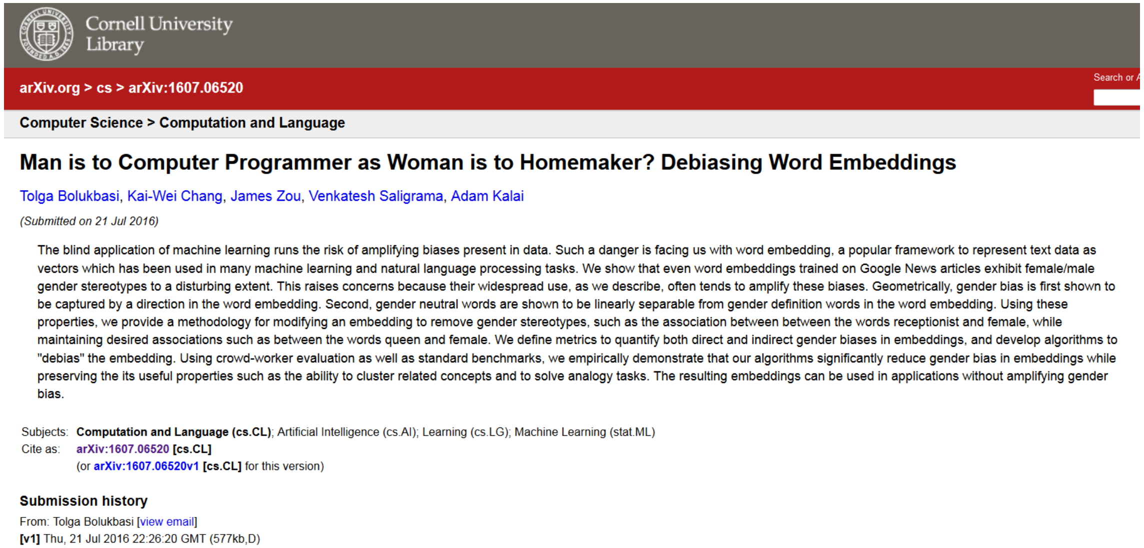
vec(“China”) – vec(“Beijing”) = vec(“Russia”) - vec(“Moscow”)
China is to Bejing as Russia is to Moscow.   Cool!

Try online: http://turbomaze.github.io/word2vecjson/

Word2vec and sexism

- Man is to king as woman is to … queen. ✓
- Sister is to woman as brother is to … man. ✓

- But:
- Father is to doctor as mother is to … nurse. ✗
- Man is to programmer as woman is to … homemaker. ✗



arXiv.org > cs > arXiv:1607.06520

Computer Science > Computation and Language

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai

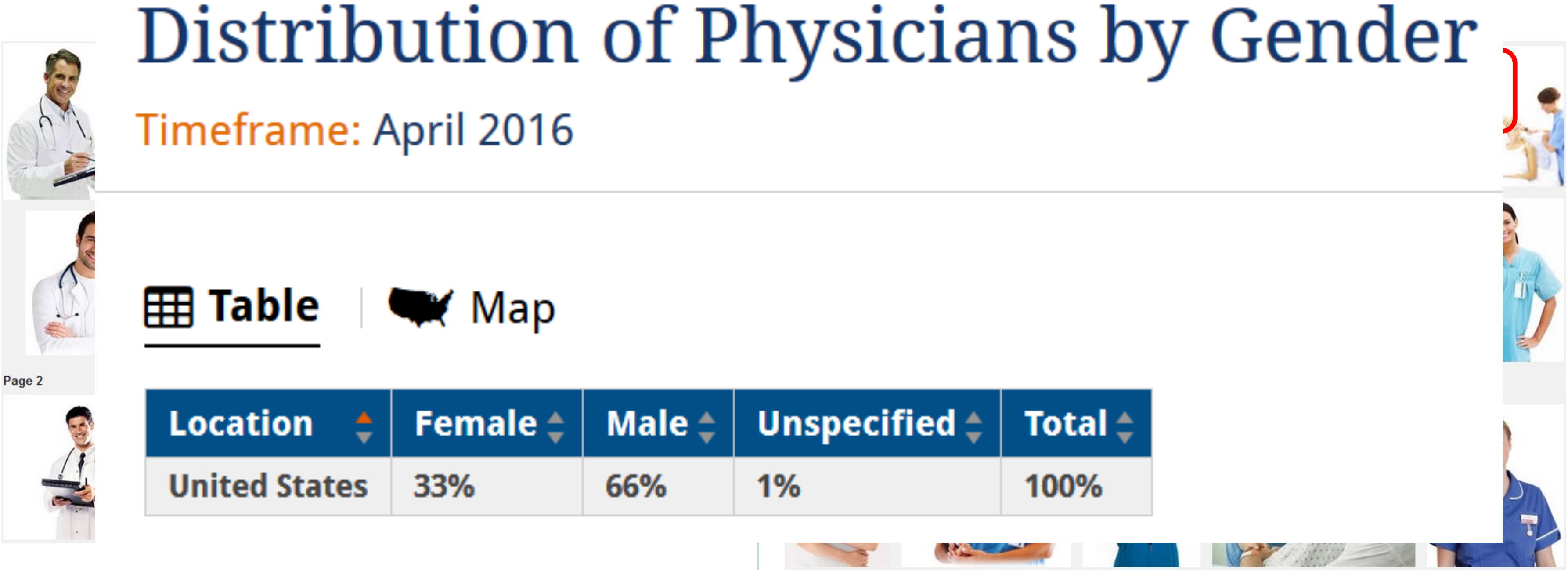
(Submitted on 21 Jul 2016)

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with word embedding, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between between the words receptionist and female, while maintaining desired associations such as between the words queen and female. We define metrics to quantify both direct and indirect gender biases in embeddings, and develop algorithms to "debias" the embedding. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.

Subjects: **Computation and Language (cs.CL)**; Artificial Intelligence (cs.AI); Learning (cs.LG); Machine Learning (stat.ML)
Cite as: **arXiv:1607.06520 [cs.CL]**
(or **arXiv:1607.06520v1 [cs.CL]** for this version)

Submission history
From: Tolga Bolukbasi [view email]
[v1] Thu, 21 Jul 2016 22:26:20 GMT (577kb,D)

Stereotypes



Distribution of Physicians by Gender

Timeframe: April 2016

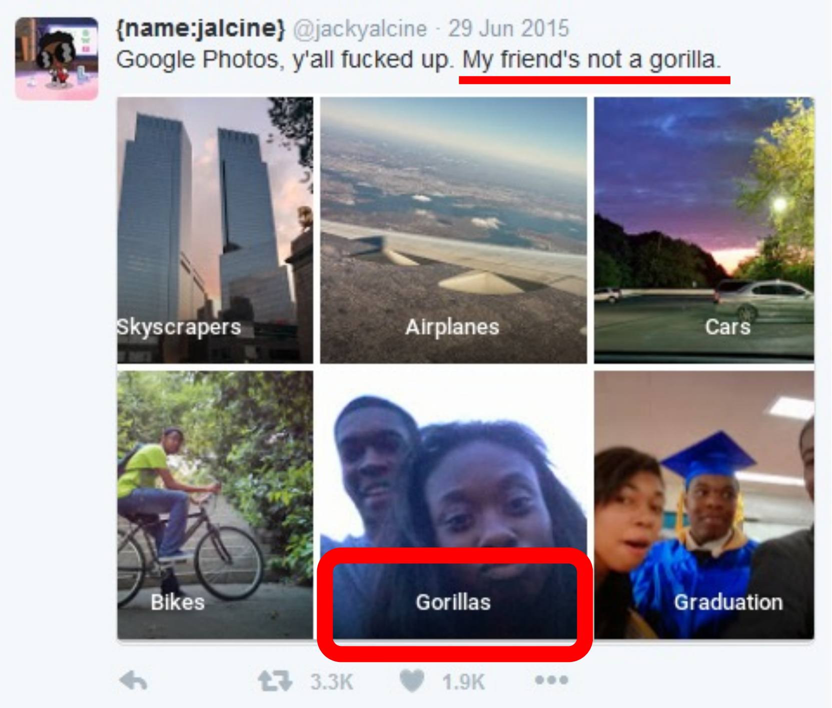
Location	Female	Male	Unspecified	Total
United States	33%	66%	1%	100%

„evidence for ***stereotype exaggeration*** *and systematic* ***underrepresentation of women"***

Kay, Matthew, Cynthia Matuszek, and Sean A. Munson. "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations." *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015.

Racism

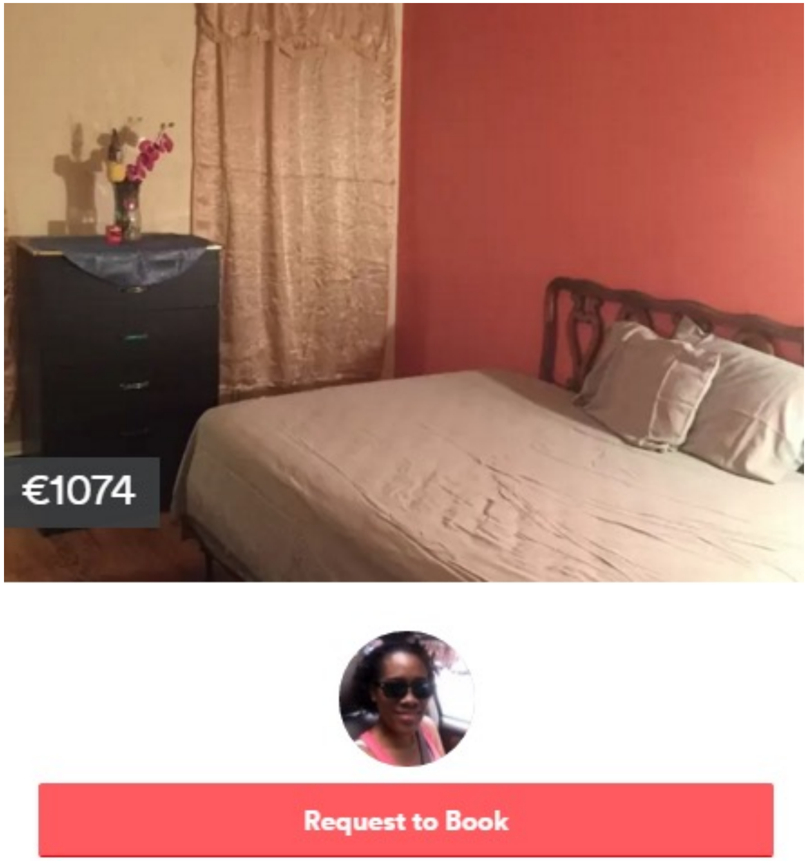
Photo app **tagging** black people **as "gorillas"**



{name:jalcine} @jackyalcine · 29 Jun 2015
Google Photos, y'all fucked up. My friend's not a gorilla.

Discrimination

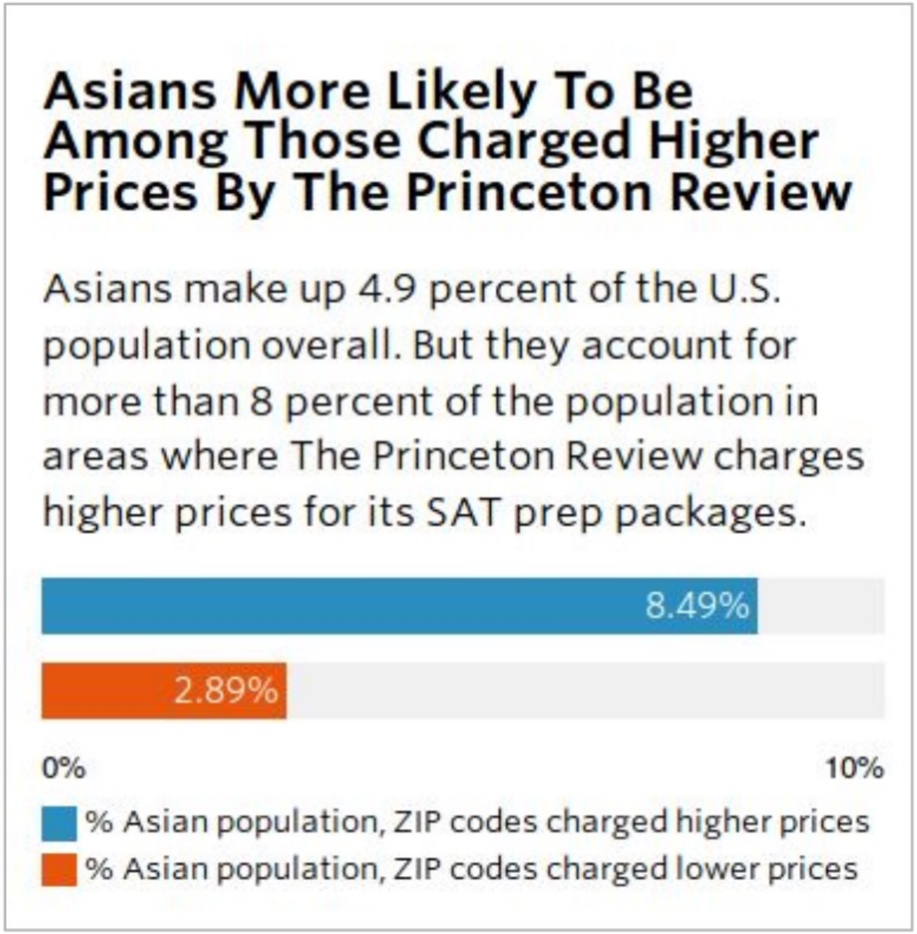
*“**non-black hosts are able to charge approximately 12% more than black hosts**, holding location, rental characteristics, and quality constant.“*



Edelman, Benjamin G. and Luca, Michael, Digital Discrimination: The Case of Airbnb.com (January 10, 2014). Harvard Business School NOM Unit Working Paper No. 14-054. http://dx.doi.org/10.2139/ssrn.2377353

Price discrimination

- we find **ev** **price steeri** **discriminat** general reta travel sites.

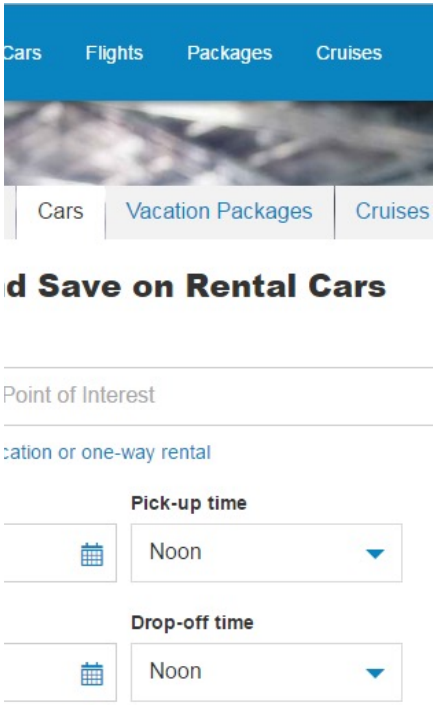


Asians More Likely To Be Among Those Charged Higher Prices By The Princeton Review

Asians make up 4.9 percent of the U.S. population overall. But they account for more than 8 percent of the population in areas where The Princeton Review charges higher prices for its SAT prep packages.

🟦 % Asian population, ZIP codes charged higher prices

🟧 % Asian population, ZIP codes charged lower prices



d Save on Rental Cars

Point of Interest

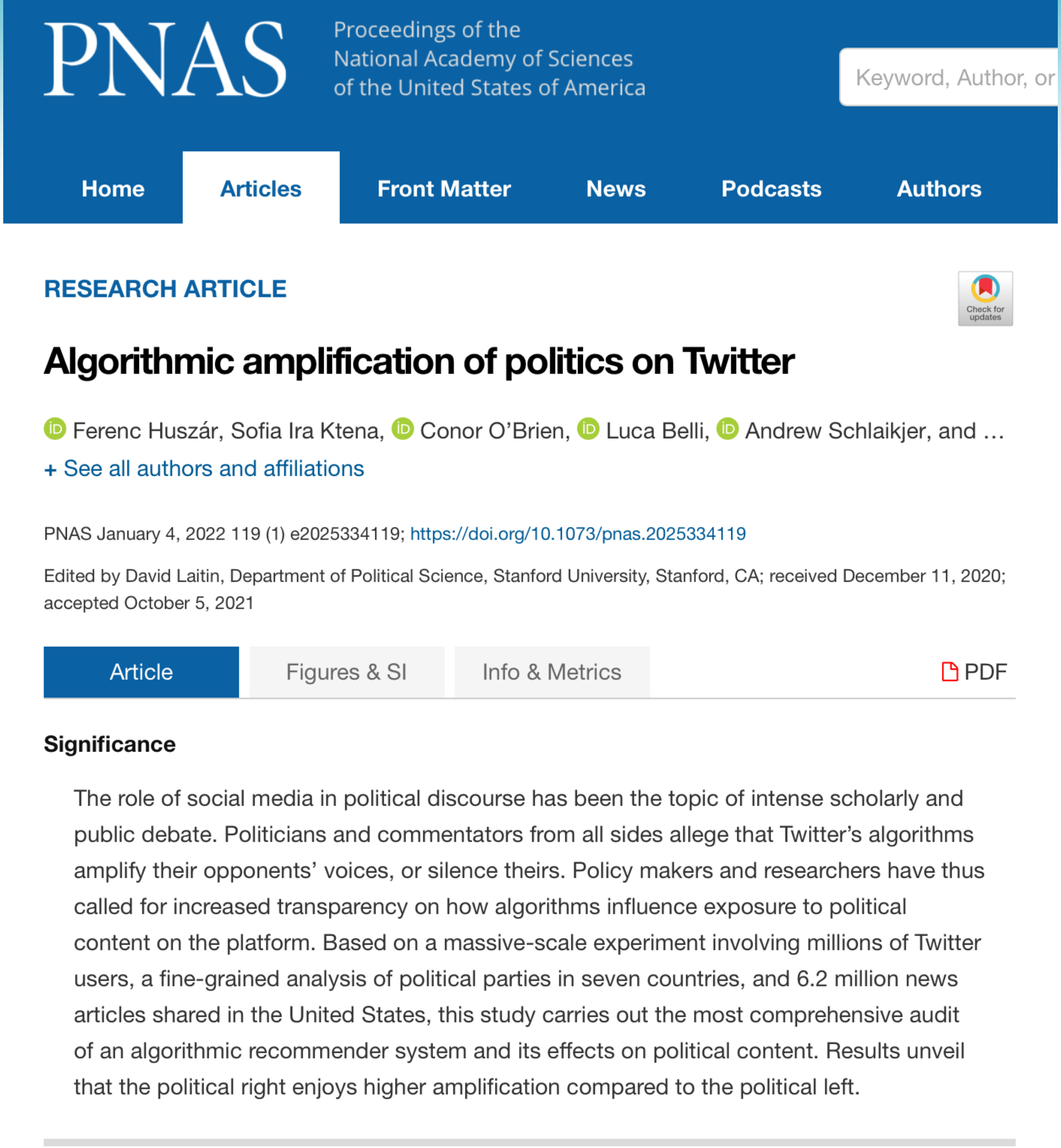
:ation or one-way rental

Pick-up time: Noon

Drop-off time: Noon

Measuring Price Discrimination and Steering on E-commerce Web Sites, Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson, In *Proceedings of the 14th ACM/USENIX Internet Measurement Conference (IMC'14)*, Vancouver, Canada, November 2014.

http://francescobonchi.com/algorithmic_bias_tutorial.html#slides



RESEARCH ARTICLE

Algorithmic amplification of politics on Twitter

Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and …

+ See all authors and affiliations

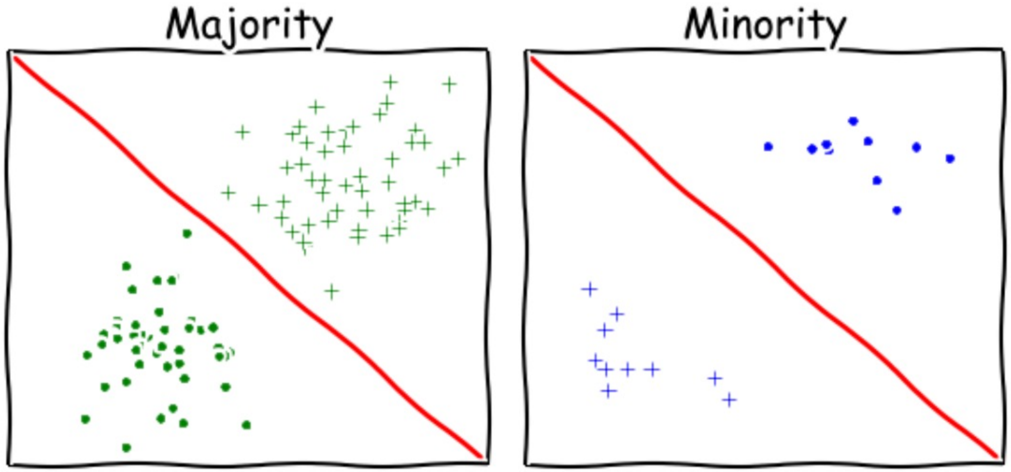
PNAS January 4, 2022 119 (1) e2025334119; https://doi.org/10.1073/pnas.2025334119

Edited by David Laitin, Department of Political Science, Stanford University, Stanford, CA; received December 11, 2020; accepted October 5, 2021

Significance

The role of social media in political discourse has been the topic of intense scholarly and public debate. Politicians and commentators from all sides allege that Twitter's algorithms amplify their opponents' voices, or silence theirs. Policy makers and researchers have thus called for increased transparency on how algorithms influence exposure to political content on the platform. Based on a massive-scale experiment involving millions of Twitter users, a fine-grained analysis of political parties in seven countries, and 6.2 million news articles shared in the United States, this study carries out the most comprehensive audit of an algorithmic recommender system and its effects on political content. Results unveil that the political right enjoys higher amplification compared to the political left.

Some Reasons for Algorithmic Bias



Positively labeled examples are on opposite sides of the classifier for the two groups.

Data tyranny of the majority class

Data selection bias

- Only cars, no bicycle trips
- Track smartphone users, not others

Algorithmic issues

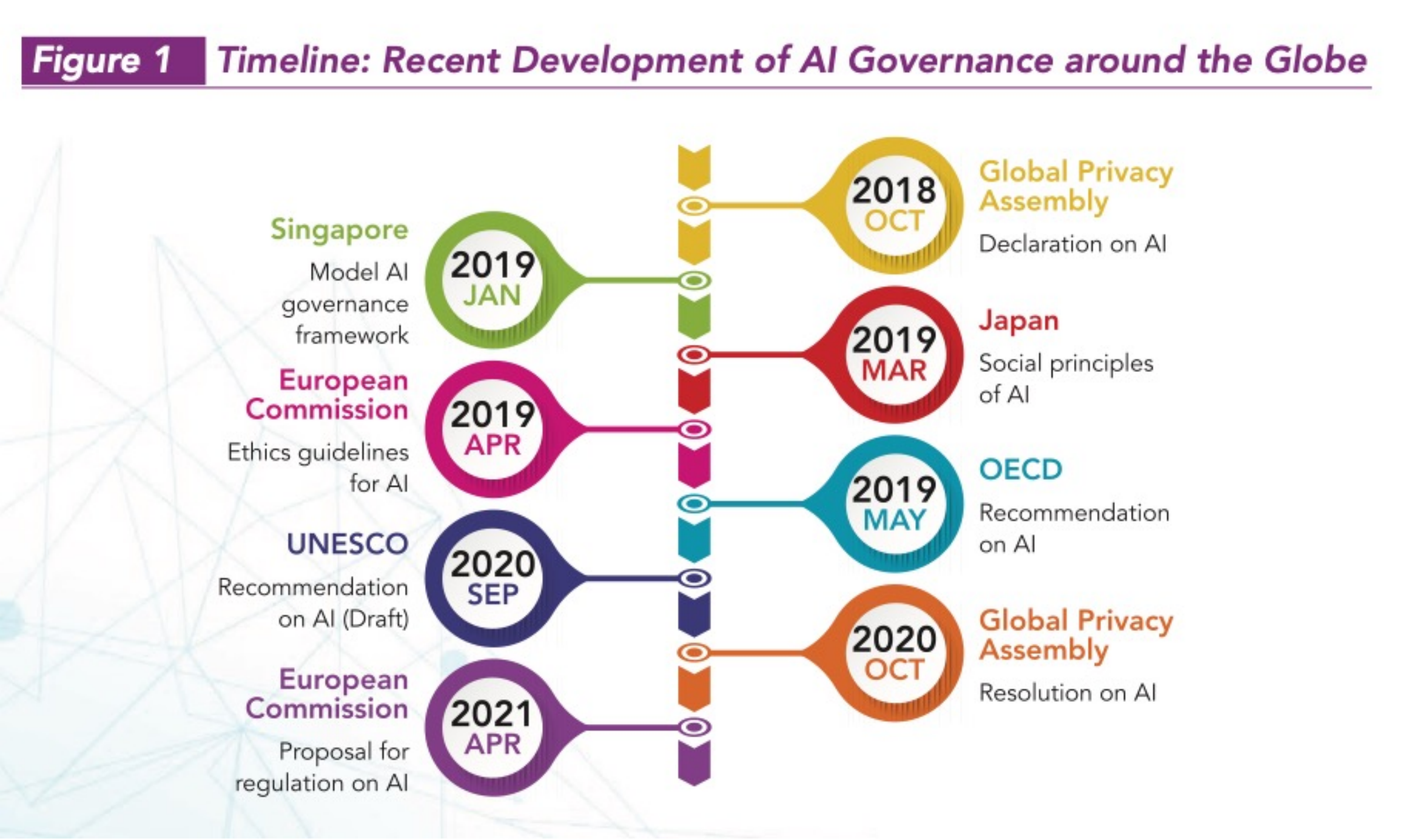
- Recommender systems that narrow, instead of broaden
- Not compensating for selection bias

Other Potential Sources for Bias

- PageRank
  - If there's within-group linkage preference, then the top 1% of the majority group will have more in-links than the top 1% of the minority group
- Click logs
  - If there's within-group click preference, then the majority group pages will get more clicks
- Language models
  - If language models reflect bias in the data then a feedback loop can be created amplifying the bias

Can AI see everything?

- Black Swan Problem: An unpredictable event that is beyond what is normally expected of a situation
- Sample 1,000 people
  - Collect their weight, age, gender, nutrition, race, …
  - Build model to predict their height
  - No major out-of-sample outliers
- Again sample 1,000
  - Collect their education, age, marital information, …
  - Build model to predict their income
  - No way to predict **Bill Gates**!



Source: https://www.pcpd.org.hk/english/resources_centre/publications/files/guidance_ethical_e.pdf

Ethics Guidelines for Trustworthy AI (EU)

1. **Lawful** - respecting all applicable laws and regulations
2. **Ethical** - respecting ethical principles and values
3. **Robust** - both from a technical perspective while taking into account its social environment

European Commission

Independent High-Level Expert Group on Artificial Intelligence

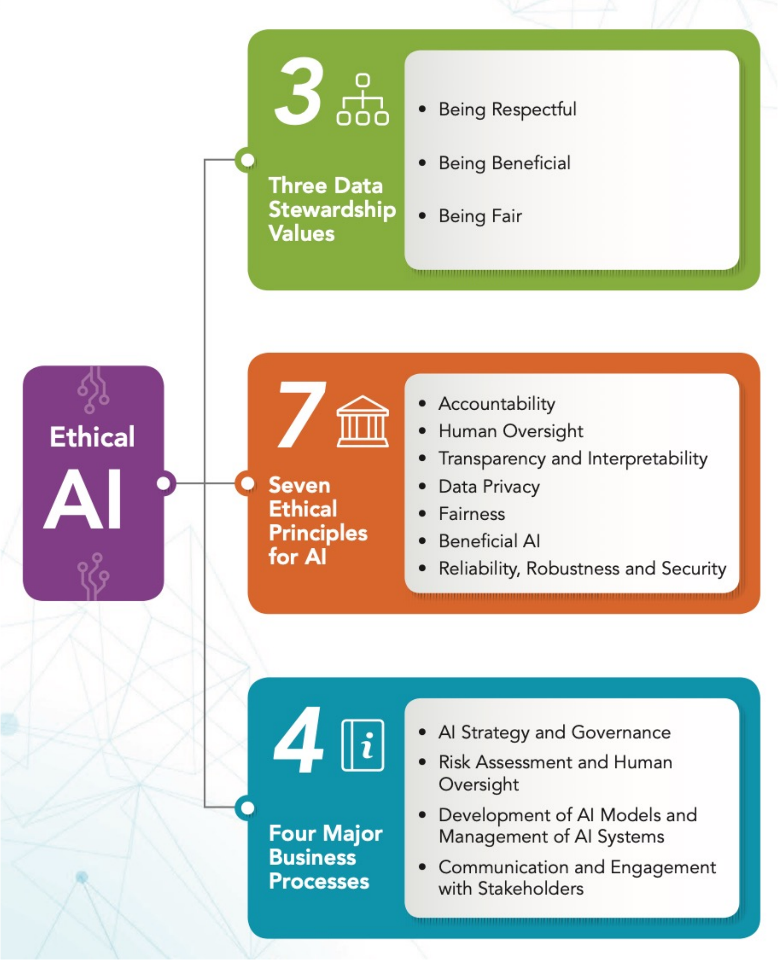
set up by the European Commission

AI

Ethics Guidelines for Trustworthy AI

Hong Kong’s Guidelines

Guidance on the Ethical Development and Use of
Artificial Intelligence
AI
PCPD
HK
PCPD.org.hk
香港個人資料私隱專員公署
Office of the Privacy Commissioner for Personal Data, Hong Kong
1996-2021
守護 · 私隱 · 廿五載
GUARDIAN · PRIVACY · 25 YEARS



Ethical AI

3 Three Data Stewardship Values

- Being Respectful
- Being Beneficial
- Being Fair

7 Seven Ethical Principles for AI

- Accountability
- Human Oversight
- Transparency and Interpretability
- Data Privacy
- Fairness
- Beneficial AI
- Reliability, Robustness and Security

4 Four Major Business Processes

- AI Strategy and Governance
- Risk Assessment and Human Oversight
- Development of AI Models and Management of AI Systems
- Communication and Engagement with Stakeholders

https://www.pcpd.org.hk/english/resources_centre/publications/files/guidance_ethical_e.pdf

3 Values

1. **Respectful**
   - every individual should be treated ethically, instead of as an object or a piece of data
2. **Beneficial**
   - need to provide benefits to stakeholders, which include individuals affected by the use of AI and the wider community
   - any harm should be prevented or minimised
3. **Fair**
   - decisions are made reasonably without unjust bias or unlawful discrimination
   - people should be treated alike
   - differential treatments between different individuals should be justifiable with sound reasons

7 Ethical Principles for AI

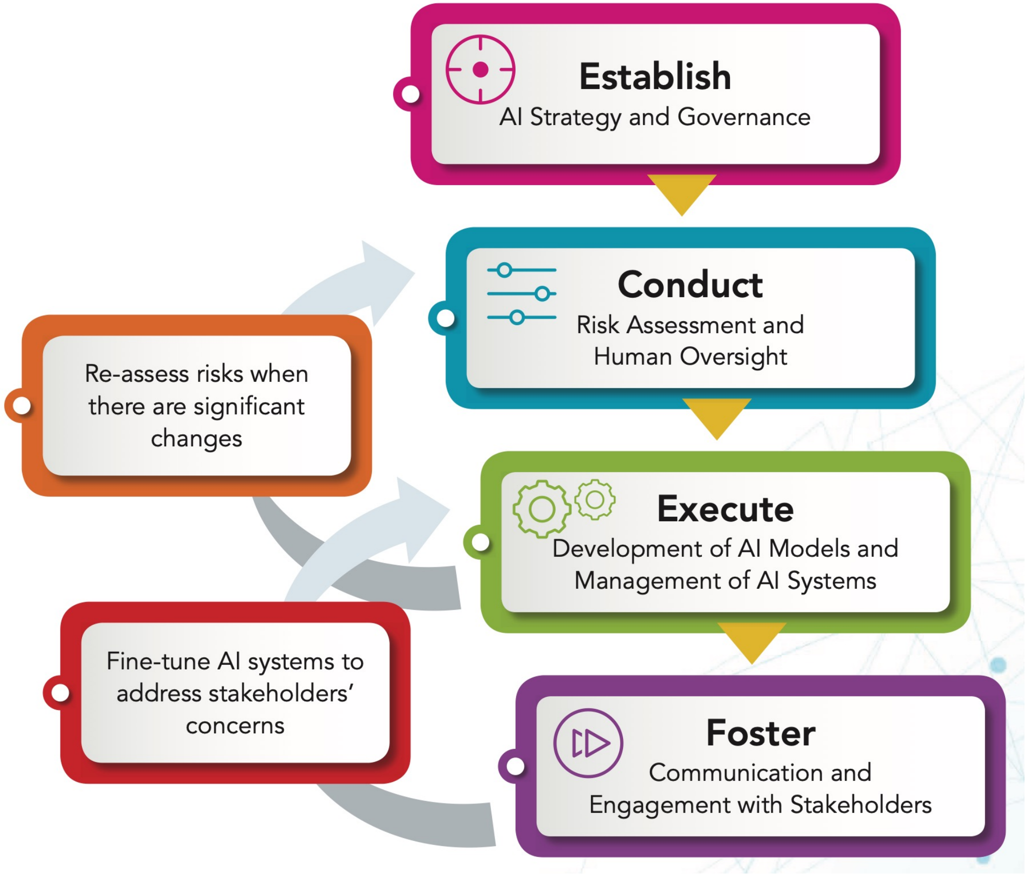
1. **Accountability**
   - Organisations should be responsible for what they do
2. **Human Oversight**
   - Users should be able to informed and autonomous actions regarding the recommendations or decisions of AI systems
   - Human intervention should always exist if the use of AI is assessed to be of high risk
3. **Transparency and Interpretability**
   - Organisations should clearly and prominently disclose their use of AI and the relevant data privacy practices while striving to improve the interpretability of automated and AI-assisted decisions

7 Ethical Principles for AI

4. **Data Privacy**
   - Protect individuals’ privacy in the development and use of AI
5. **Fairness**
   - Individuals are entitled to be treated in a reasonably equal manner, without unjust bias or unlawful discrimination
6. **Beneficial AI**
   - AI should provide benefits to human beings, businesses and the wider community
   - Provision of benefits encompasses prevention of harm
7. **Reliability, Robustness and Security**
   - Organisations should ensure that AI systems operate reliably as intended over their expected lifetime
   - AI systems should also be protected against attacks, such as hacking and data poisoning

Business Processes

- Defines what should do when using AI


Establish
AI Strategy and Governance
Conduct
Risk Assessment and Human Oversight
Execute
Development of AI Models and Management of AI Systems
Foster
Communication and Engagement with Stakeholders
Re-assess risks when there are significant changes
Fine-tune AI systems to address stakeholders' concerns