Wharton
University of Pennsylvania
ONLINE

AI Strategy and Governance

Risks with AI

Kartik Hosanagar, Professor of Operations, Information and Decisions

Risks with AI: Overfitting

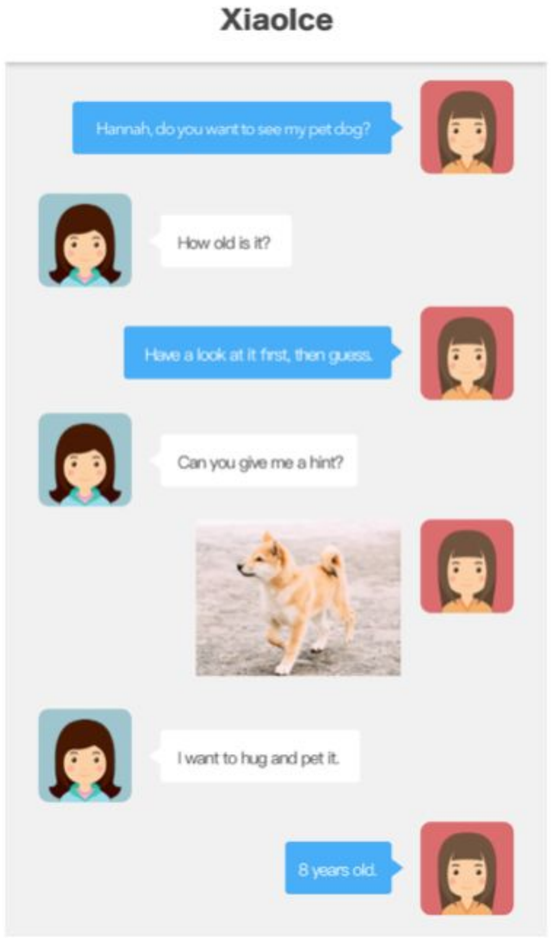
- Complex AI models such as neural nets can easily overfit (i.e. fit historical data too well but fail in realistic test conditions)
  - If we don’t understand what is helping the model perform well, there is a risk that the model will fail upon deployment
- Operational Risks
  - Direct financial risks (e.g. a trading algorithm)
  - Customer perception and reputation (e.g. poor personalization experience)
- ML Models need to go through thorough stress-testing (discussed later under audits)

Xiaoice: Darling of Chinese Social Media

XIAOBING

40M followers

China



YUAN ZHANG

22 years old

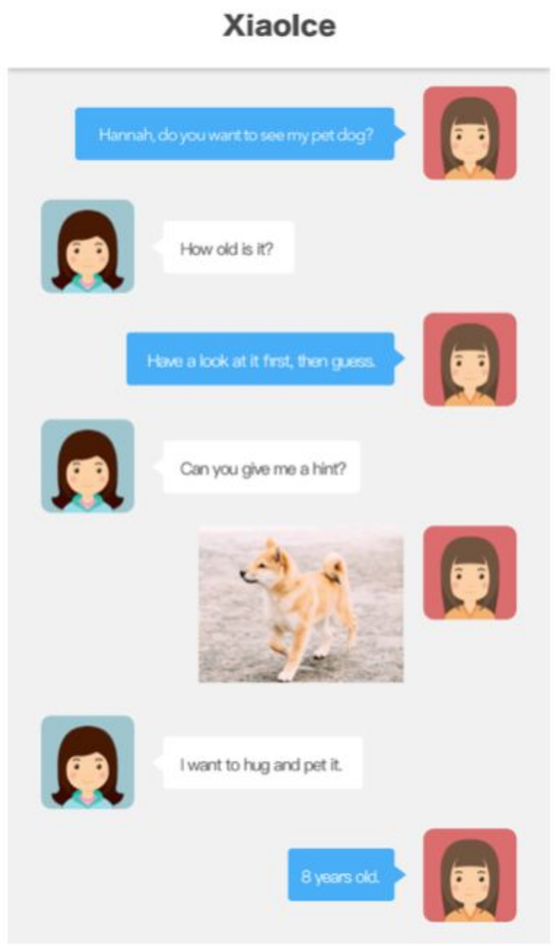
China

Xiaoice: Darling of Chinese Social Media

XIAOICE

40M followers

Chatbot



YUAN ZHANG

22 years old

China

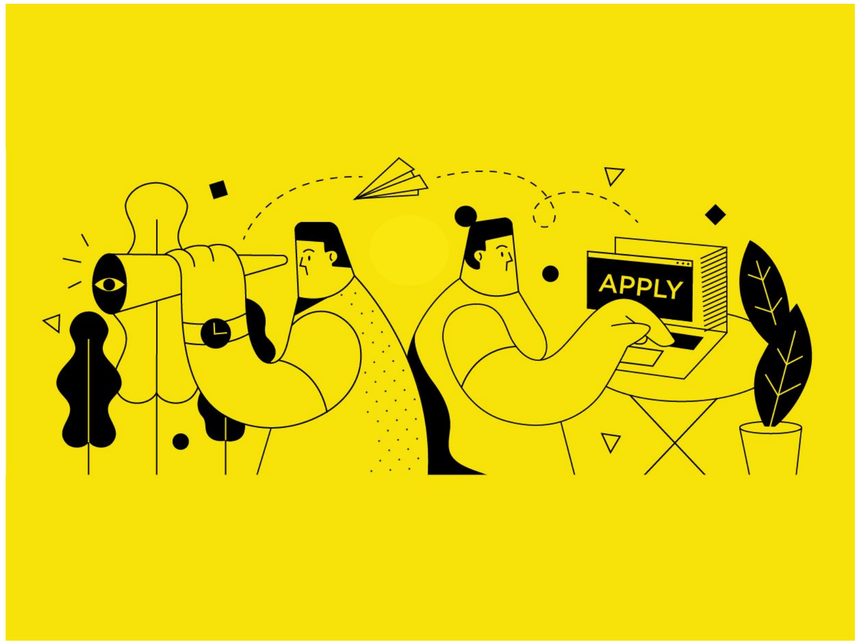
Tay.ai: Xiaoice's Evil Cousin

Microsoft's racist chatbot, Tay, makes MIT's annual worst-tech list

www.geekwire.com/2016/**microsoft**-chatbot-**tay**-mit-technology-fails/

Dec 27, 2016 - **Tay**, the **Microsoft** chatbot that pranksters trained to spew racist views, has resurfaced on MIT Technology Review's list of 2016's top technology ...

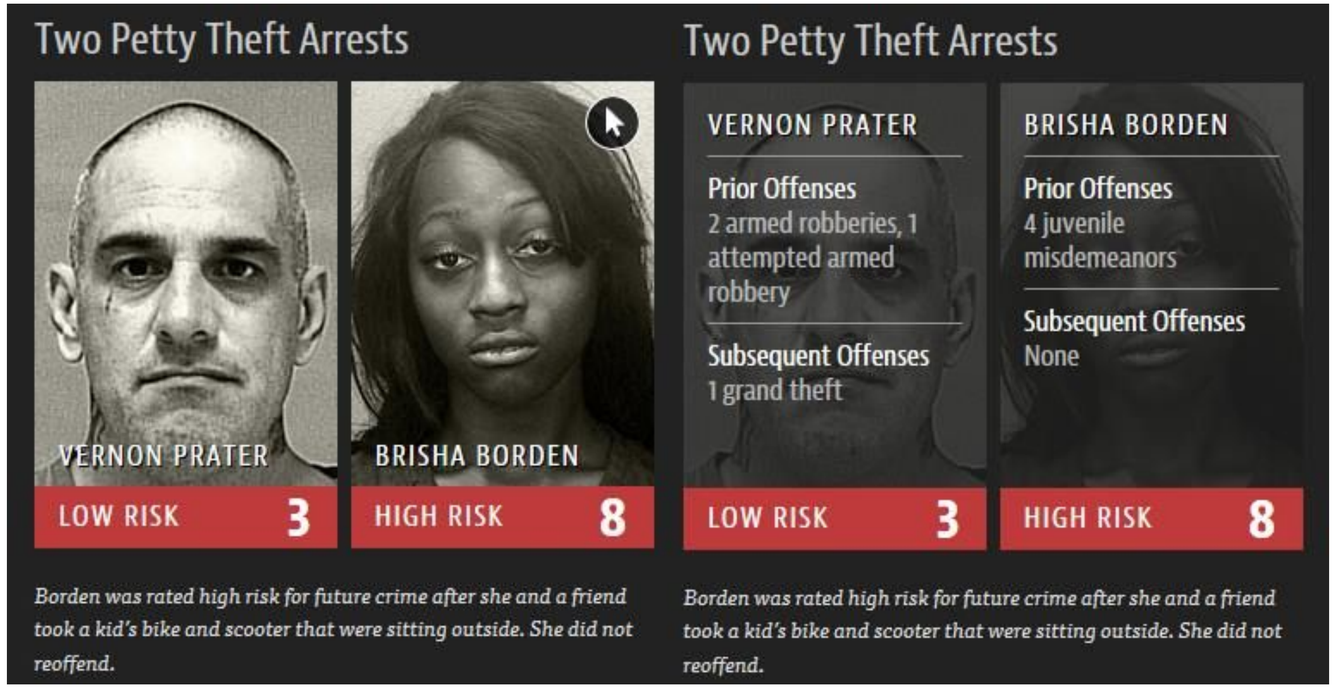
Algorithm Bias in Recruiting Software



Amazon’s machine learning specialists uncovered a big problem: their new recruiting engine **did not like women**

Algorithms Incorrectly Predict Recidivism

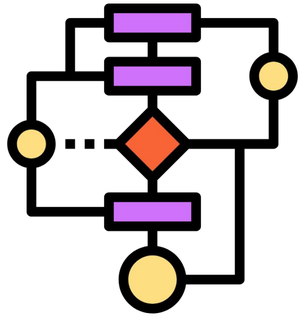
There are **higher false positive and false negative rates** among African Americans



Algorithms racial bias in predicting recidivism rates

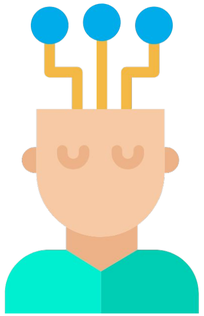
Why Might AI-Based Decisions be Unpredictable?

Algorithm Logic
Nature



Data
Nurture

AI Behavior



Analogous to human behavior, algorithms are driven by **nature and nurture.**

Risks to Society

- Social risks can result from automated decisions because these decisions may result in disadvantaged minorities continuing to be left behind
  - The AI Now Institute classifies theses risks into two groups:

Harms of Allocation

- About situations when a scarce resource has to be allocated to people
- E.g.: Unfair loan approval decisions, or job applicant decisions

Harms of Representation

- About situations when a system represents a group in an unfavorable way
- E.g.: Airport screening system being more likely to false alarm people of color (based on hairstyles)

Content/quotes from: https://medium.com/@jstanier/we-must-fix-ais-diversity-problem-6ad5fddc2f8c
Content from: https://gizmodo.com/microsoft-researcher-details-real-world-dangers-of-algo-1821129334

Risks to Firms

- These social risks then create additional risks for companies

Reputational Risk

- Perceived to be a biased, prejudiced company
- Firms may face PR issues and backlash as a result

Legal Risk

- Sued for unfair practices and discrimination against particular groups

Regulatory Risk

- Increased regulation & cost of compliance
- Upcoming interest in auditing and data protection (GDPR)

Wharton
University of Pennsylvania
ONLINE

AI Strategy and Governance

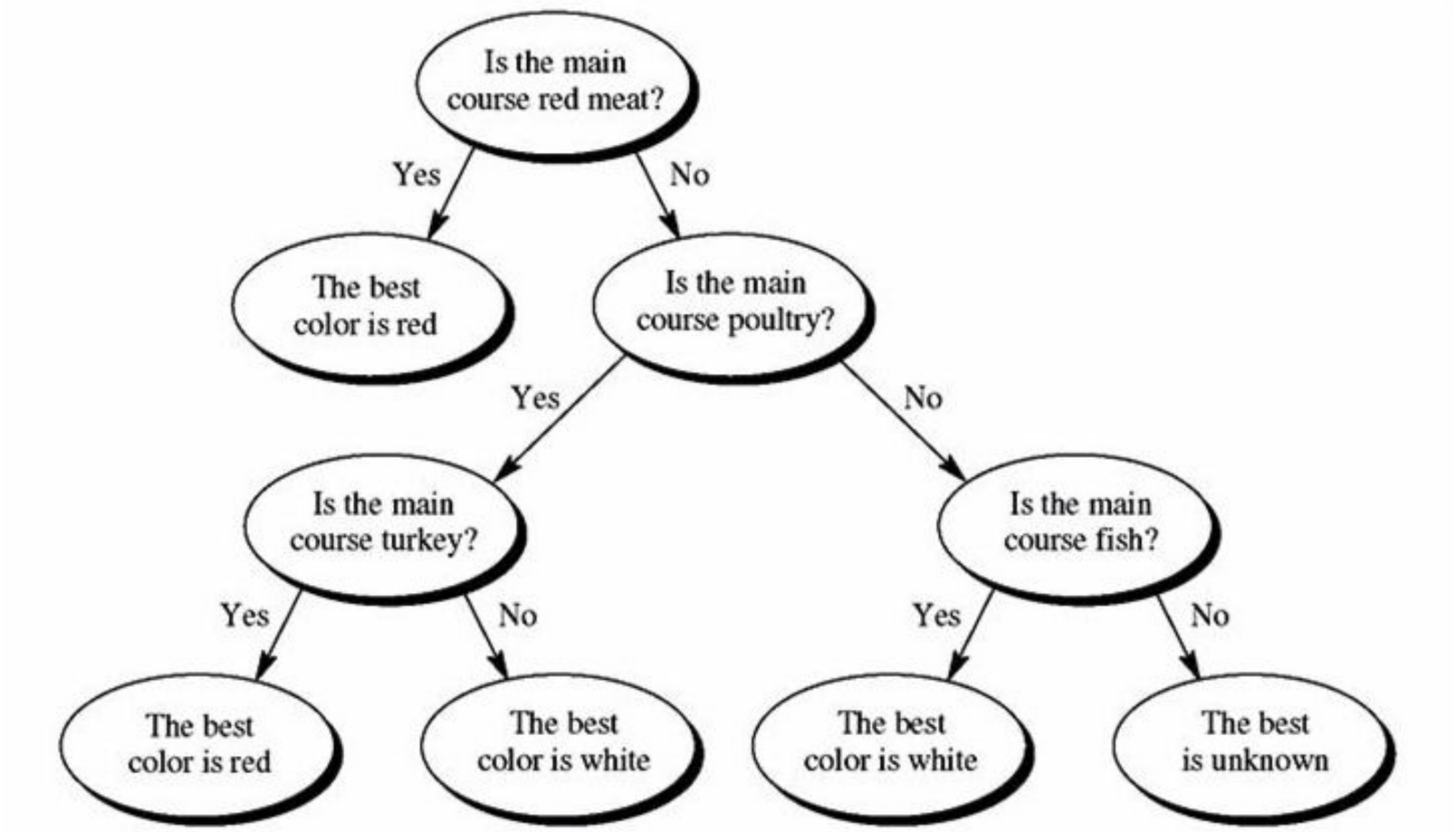
Explainable AI: What is Explainable AI?

Prasanna (Sonny) Tambe, Associate Professor of Operations, Information and Decisions

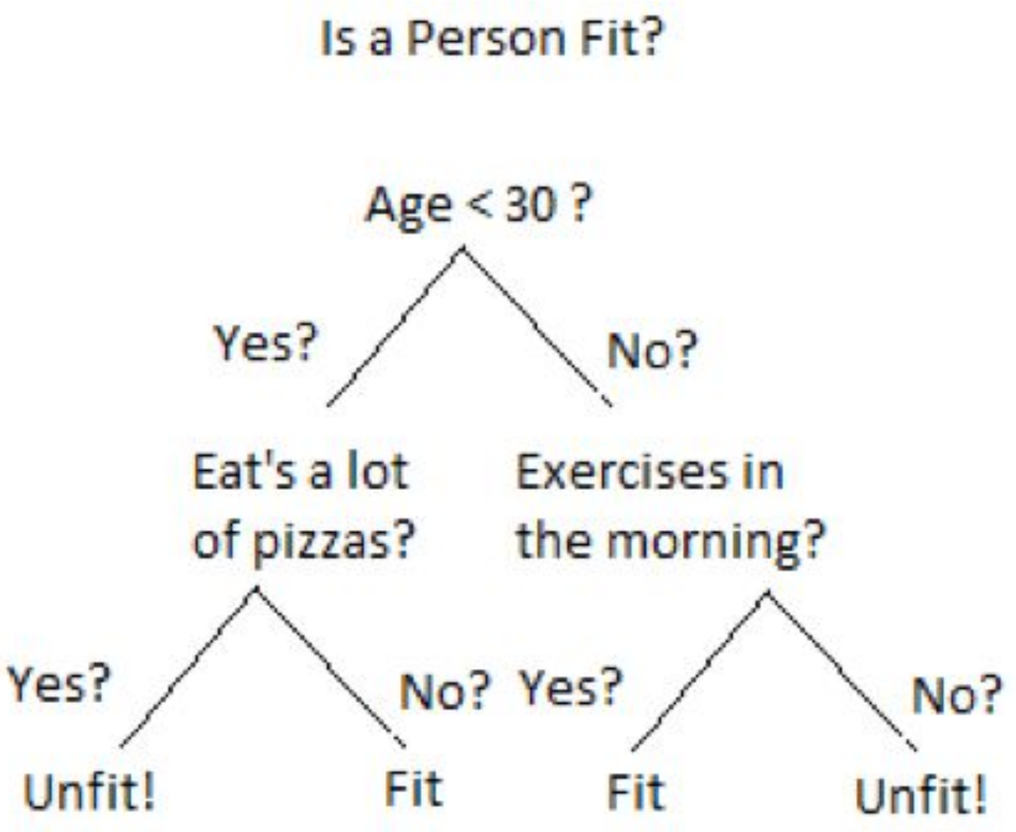
AI Explainability

- The use of methods in AI systems where why the algorithm arrived at a particular result can be easily understood by human experts
- Closely related to interpretability — understanding why a decision was arrived at by an algorithm, even if you can't necessarily explain that logic
- Contrasts with the "black box" approach normally associated with some types of complex machine learning (e.g. deep learning)

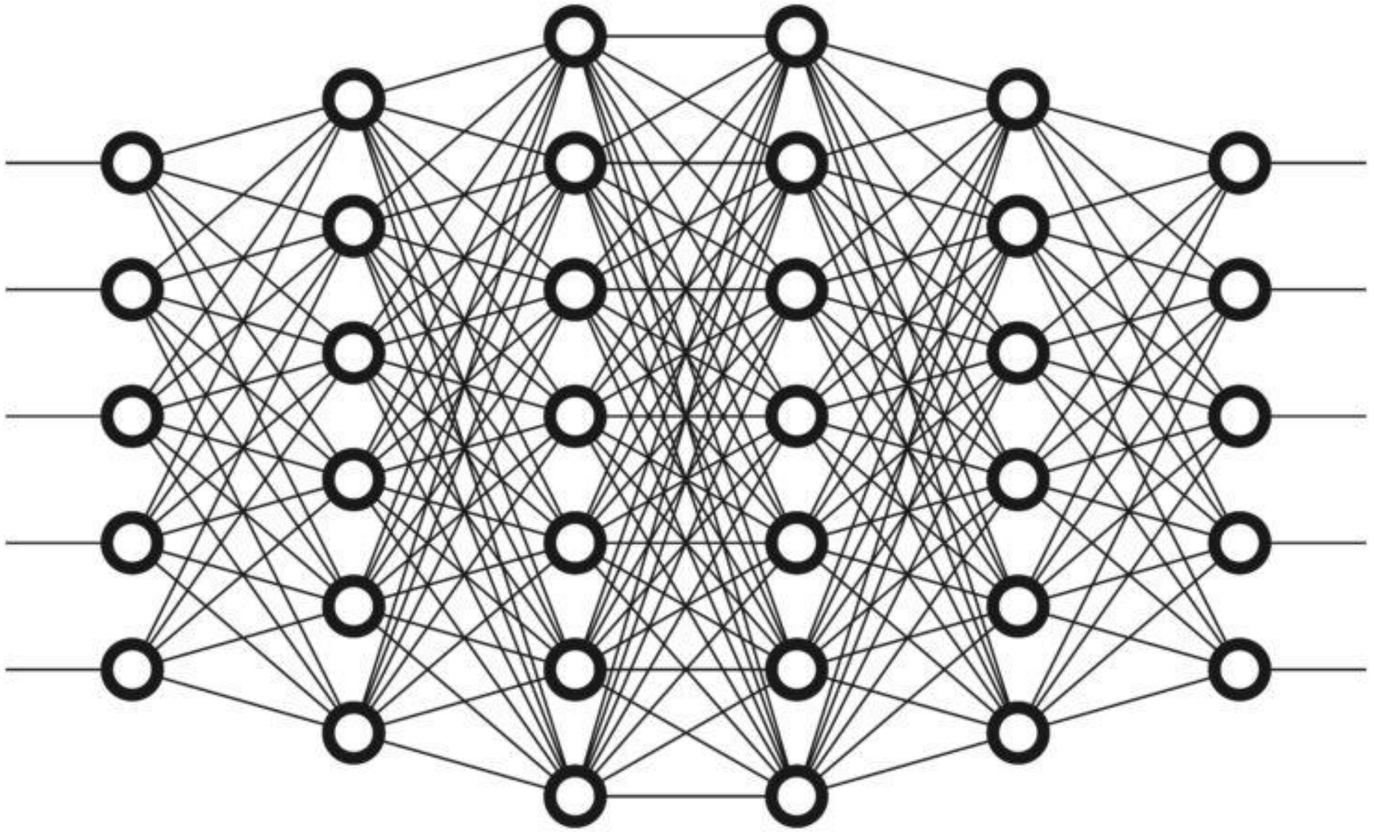
Decisions Based on Business Rules Are Easy to Explain



Decision Tree Models Are Relatively Easy to Interpret



Deep Learning Models Are Difficult to Interpret



AI Explainability

- Major tradeoff with more complex models
  - Able to handle enormous amounts of data and make very accurate predictions
  - Can be difficult to explain the logic
- Explainability is the key to adoption in many contexts

AI Explainability

- Big tech companies are currently heavily invested in this issue
- Efforts by the government that are funding programs to develop better explainable AI

Wharton
University of Pennsylvania
ONLINE

AI Strategy and Governance

Algorithmic Bias and Fairness

Kevin Werbach, Professor of Legal Studies and Business Ethics

Bias and Fairness

- Data can embed human prejudice
  - If women traditionally fail to get promoted and enjoy long careers at a company because of rampant sexism, an AI system will find that being female is associated with poor outcomes
- In other cases, the data itself can be biased
  - There may simply be fewer examples of minority populations in the training dataset, resulting in less-accurate models
- Sometimes even when classifications such as race and gender aren't even in the dataset, they influence models through proxies
  - A zipcode is just an address marker, but it can be strongly correlated with race or socio-economic status, for example

Technical Responses to Algorithmic Bias

- There are now a variety of tools to incorporate fairness criteria into the design of algorithmic systems directly, or to assess whether they produce discriminatory results
- However, they aren't foolproof
  - There isn't a single definition of fairness
  - Systems that are more fair might also be less accurate
  - In some cases, there may not be objective standards of fairness at all
- It's not an accident what data gets collected, how that data is evaluated, or what questions get asked in the design of algorithms
  - The same human factors that lead to marginalized groups being discriminated against in other contexts still apply here

Legal Responses of Algorithmic Bias

- There are some legal claims that can be brought against biased or unfair algorithms, but their scope is quite limited
- Disparate impact — when a policy or practice that appears to be neutral (doesn't explicitly treat minority populations different from other populations) still has an effect of different treatment
  - Only applies to a limited set of protected classes, (e.g. race and gender)
  - Generally applies only to certain activities specified under the law (e.g. employment and housing)
  - Requires a defined "policy or practice" affecting a protected class
  - US Supreme Court has held that statistical disparity alone is not enough

Legal Responses of Algorithmic Bias

- In Europe, the General Data Protection Regulariont (GDPR) has general anti-bias provision for ”fully-automated processing”
  - Limited context and not clear what counts as bias
- Proposals for new laws
  - European Union AI paper
  - In the US, the Algorithmic Accountability Act

How Organizations Should Respond to These Challenges

- Deep and diverse data
- Think about proxies for illegitimate factors (e.g. zip codes)
- Consider the appropriate fairness function
  - Test and evaluate system performance to assess tradeoffs
- Be aware of hidden historical biases
  - Having diverse teams is critically important

Wharton
University of Pennsylvania
ONLINE

AI Strategy and Governance

Manipulation

Kevin Werbach, Professor of Legal Studies and Business Ethics

Manipulation

- Manipulation falls between legitimate persuasion and illegitimate coercion
  - Getting someone to do something that you want by somehow short-circuiting their capacity for rational decision-making
- All manipulation isn't illegal or unethical
  - Advertising, political campaigning, and fund-raising, for example
- AI can manipulate people when it's not obvious that choices or decisions are being shaped by algorithms

Manipulation: Facebook

- Facebook collaborated with academic researchers to measure whether changes in its newsfeed algorithm could generate what psychologists call Emotional Contagion
  - It deliberately fed certain users happier content, and sure enough, they shared happier content with their friends
- In other research Facebook found it could increase voter participation by tweaking the newsfeed
- One of the challenges here is that everything you see on the Facebook newsfeed is the result of algorithms and they are changing all the time

Deception

- Some forms of manipulation are legally prohibited, such as false or subliminal advertising, but those are generally defined narrowly
- The major legal concept here is deception
  - It’s perfectly fine to market a product to customers that you think they want to buy, even when it’s personalized through AI, because users understand that advertising is about selling them things
  - The problem comes when the nature of the relationship isn’t obvious

Exploitation

- Exploitation is a more harmful form of manipulation that involves taking advantage of vulnerabilities to produce voluntary agreements that would not occur in a competitive market
  - UK airlines deliberately seating families apart from one another when they purchased cheaper tickets, in order to encourage them to upgrade to higher priced tickets so they could sit together
  - Leaked Facebook advertising presentation suggested it could identify when teenagers were feeling “worthless,” “insecure,” or “anxious”
- This is the point at which responsible AI practitioners need to draw a line
  - If you wouldn’t consciously design a business practice to exploit vulnerable people, you shouldn’t do it through algorithms

Market Manipulation

- Using algorithms to subtly undermine competitive markets
- Amazon sells both third party products and its own private-label offerings
  - Its search engine doesn’t directly prioritize its own products, but it does incorporate signals that use proxies for Amazon’s profitability, which could result in that kind of bias
- In some cases, algorithms can execute collusive strategies
  - 2018 DOJ action against poster sellers on Amazon
  - Researchers have shown that machine learning algorithms may self-discover collusive strategies

Manipulation Responses

- There is no bright line defining manipulation, except perhaps in the market manipulation cases where general principles of antitrust or competition policy can be applied
- The question you should ask yourself is whether the objectives of your AI system are creating mutually beneficial relationships with your stakeholders
  - Are people getting what they would likely choose on their own if they understood the nature of the relationship?
  - Or are you essentially tricking them?

Manipulation Responses

- In the context of academic research on human subjects, a standard set of principles were developed in something called the Belmont Report in the 1970s
- The four main elements are:
  - Informed consent — meaning users truly understand what they are getting into, unless it's something that involved no real risk to them
  - Beneficence — "do not harm"
  - Justice — non-exploitative, administered fairly
  - A dedicated review board

Manipulation Responses

- Many organizations find it helpful to have a dedicated center of excellence or committee to evaluate these and other ethical questions about major AI implementations
  - Shouldn’t come at the expense of diffusing responsible AI throughout the organization

Wharton
University of Pennsylvania
ONLINE

AI Strategy and Governance

Data Protection

Kevin Werbach, Professor of Legal Studies and Business Ethics

Data Protection Introduction

- Governments are increasingly concerned about what the scholar Shoshana Zuboff labeled “surveillance capitalism”
  - The business models that see personal data as a resource to be exploited through ever more sophisticated personalization and targeting
- Ever since databases were widely applied in business in the 1960s, there have been concerns about how that power might be abused, or violate fundamental rights

Novel Data Protection Issues

Big data and machine learning raise novel data protection issues

- Big data requires large datasets
- Inferential privacy violations
  - e.g. Researcher Mikal Kosinski's algorithm to predict sexual orientation based on Facebook profile photos or likes
  - Murky area, both legally and ethically
- Models often require significant data transfers in order to be updated (e.g. from phones to the cloud)
  - Federated privacy can be used to moderate this problem

5 Stages of the Privacy Lifecycle

1. Collection
2. Aggregation/analysis
3. Storage
4. Use
5. Distribution

Data Protection: Legal Frameworks

- US: Market based, user choice (notice and consent), sectoral
- Europe: Human-rights based, comprehensive
  - GDRP applies to both data controllers and processors, for any personally identifying information involving European citizens

US	Europe
• Opt-out • No requirement to specify purpose • Gray area regarding brokers	• Opt-in • Defined purpose/purpose limitation • Regulates brokers • Additional rules for fully automated processing • Substantive rights

Data Protection: Legal Frameworks

- Most other major jurisdictions adopting rules similar to the European approach
- US is moving in that direction
- Movement to require more explicit protections for AI systems when they involve high-risk data collection
- Seem to be seeing a race to the top, rather than to the bottom
  - Especially for multinational companies

Data Protection: Responses

- In building an AI project, it’s crucial to consider data protection at two levels: the technical level and the operational level

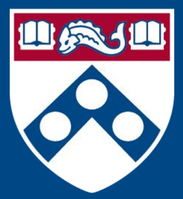
Data Protection at the Technical Level

There are a variety of techniques to make systems more privacy protective

- Federated learning
- Differential privacy
  - Mathematical technique for strategically adding noise to datasets
  - Statistical queries produce equivalent results, but it's impossible to determine whether a particular individual is part of the dataset
  - Allows you to tune accuracy/privacy tradeoff precisely

Data Protection at the Operational Level

- Privacy by design — first introduced by Ann Cavoukian, the privacy commissioner of Ontario, Canada, and now formally incorporated into GDPR
  - Incorporate privacy considerations into every decision involving personally-identifiable, or potentially personally-identifiable data, at every stage of the process
  - If everyone on your team is aware of the risks and thinking about where something could be problematic, you are much more likely to avoid a controversy
- Formal mechanisms such as data impact assessments
- The most important principle for privacy and data protection is to see them as pervasive considerations



Wharton

UNIVERSITY of PENNSYLVANIA

ONLINE