Ethics, Bias, and Fairness Part I: Overview, Frameworks, and Issues

Rayid Ghani and Kit Rodolfa

Carnegie Mellon University

ML
MACHINE LEARNING
DEPARTMENT

HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Reminders

- No Class Wednesday and Thursday
- Project update due October 19th (Monday)

We know data can help improve society but we have to do it responsibly and ethically

We need to be ok with not having answers but raising more questions that can be empirically informed

Most of these questions are not new

Ethical issues have always existed in every decision-making process

Why are we talking about them in this class now?

Most of these questions are not new

Ethical issues have always existed in every decision-making process but we are dealing with them now at a different scale and with a more data-driven view

Dim Sum is so much better than Sushi

Good research can only happen in academia

My phone should never collect, store, or use my location information

My phone should never collect, store, or use my location information

Self Driving cars save lives

Predicting whether someone will commit a crime in the future is unethical

Predicting whether someone will commit suicide in the future is unethical

Pittsburgh vs New York

Predictive policing systems reduce crime

Facial Recognition Systems should be banned

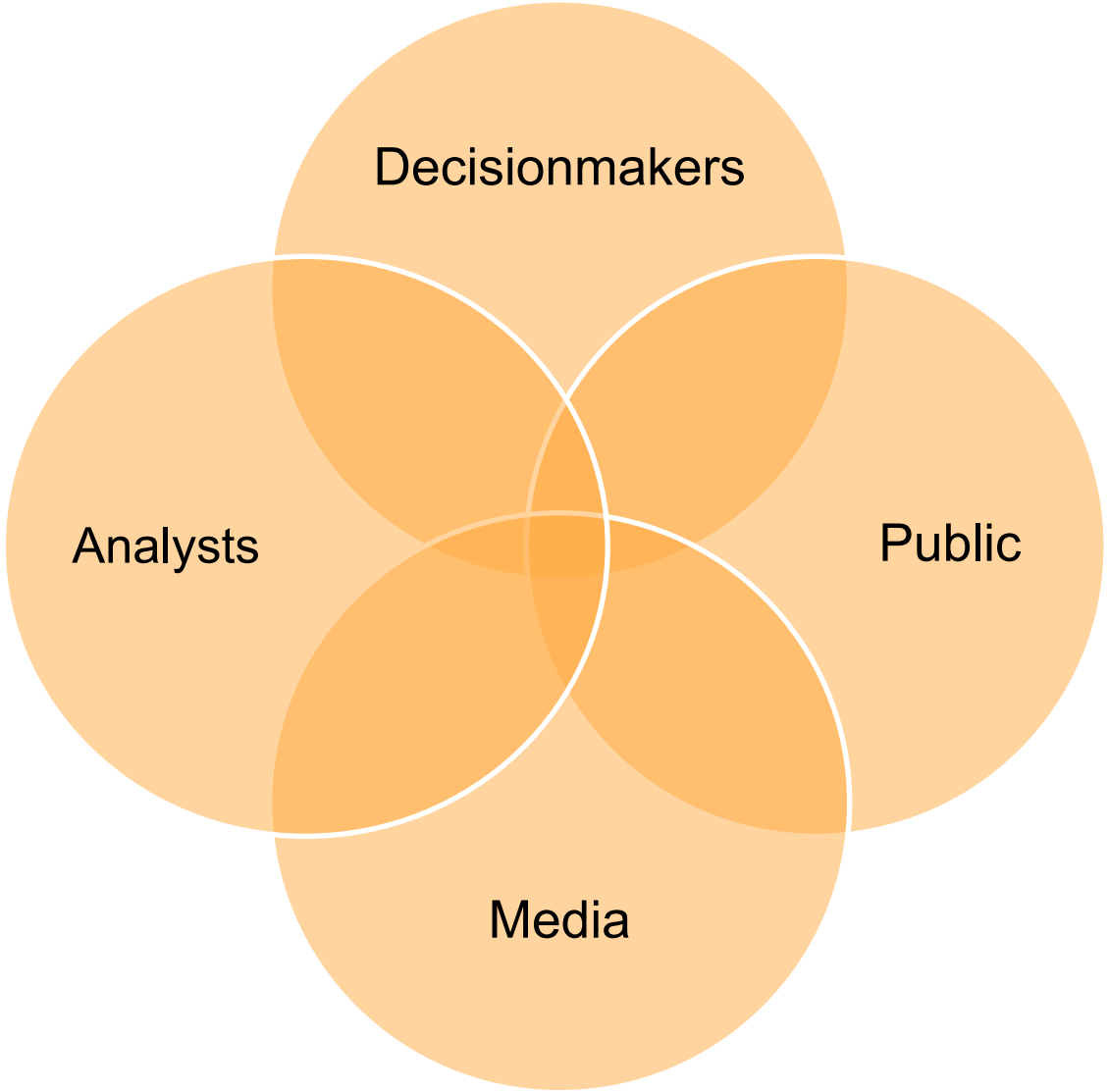
The blame for the negative consequences of the WCG system lies with the mayor

The blame for the negative consequences of the WCG system lies with the ML Developers

Our policies need to reflect our values

- What do we mean by equity?
- Is it ok to (reverse) discriminate against one group to reach equity?
- Should our right to privacy matter more than our right to life? Or to healthcare?
- Should we be allowed to use someone’s data just because it’s accessible to us?
- Is predictive policing always bad?

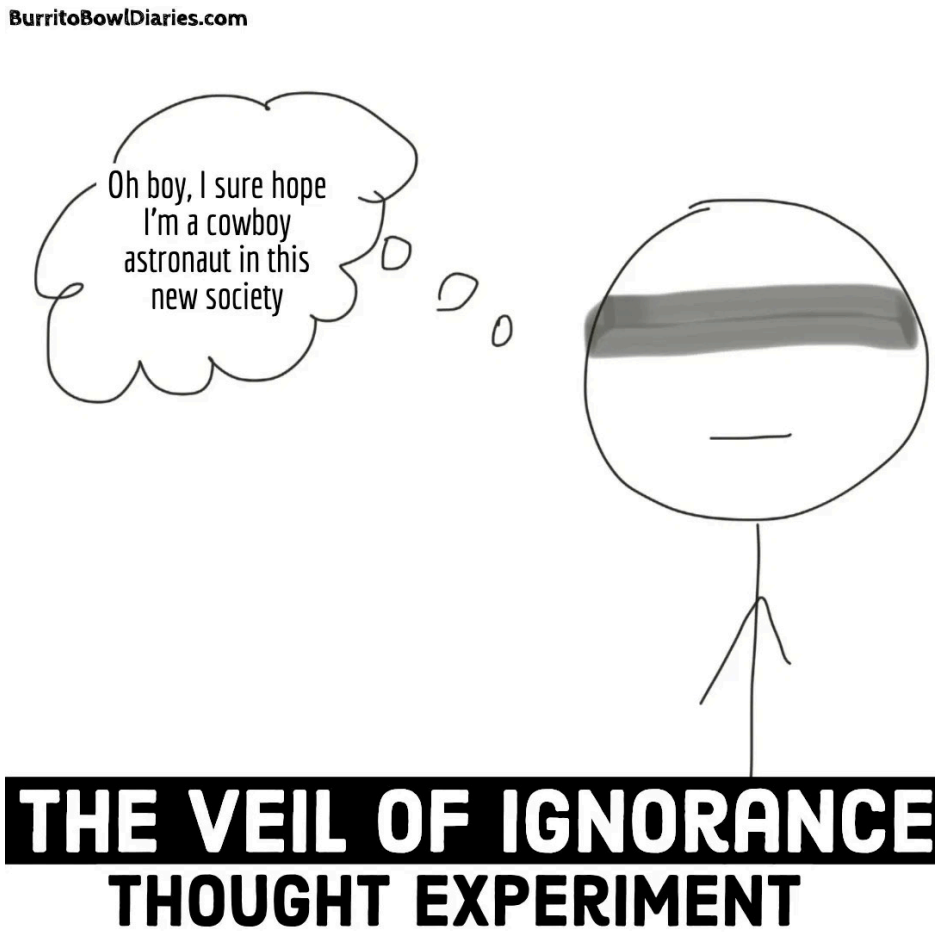
We need to think of & include the perspectives of different groups



Rawls's Theory of Justice

- Moral philosophy is much more about frameworks to help approach tough problems and balance competing obligations than clear-cut answers
- Rawls is far from the only voice here and many other perspectives are well worth exploring: utilitarianism (Singer), deontology (Kant), Scanlon, Sen, etc.

Rawls's Theory of Justice



- Make rules from the "original position" under a "veil of ignorance"
- Rawls's conclusions: equality in rights/duties & the difference principle
  - Maximize the position of the least-well-off in society ("maximin")

Some questions we should be asking

Some questions we should be asking

- Are you using data for purposes it’s intended for?
- How are you protecting the data?
- Do the people who “own” the data know you’re using it?
- Do you have their permission? How was it obtained?
- What actions are you taking on individuals based on this data?
- Do the people you’re targeting know why and if they’re being targeted?
- What recourse do they have?
- Would it make the front page of NYT if they found out what you’re doing?

Data and AI Ethics Issues

Privacy

Data Ownership

Bias, Equity, & Fairness

Transparency

Trustworthiness and Accountability

Data and AI Ethics Issues

Privacy

Data Ownership

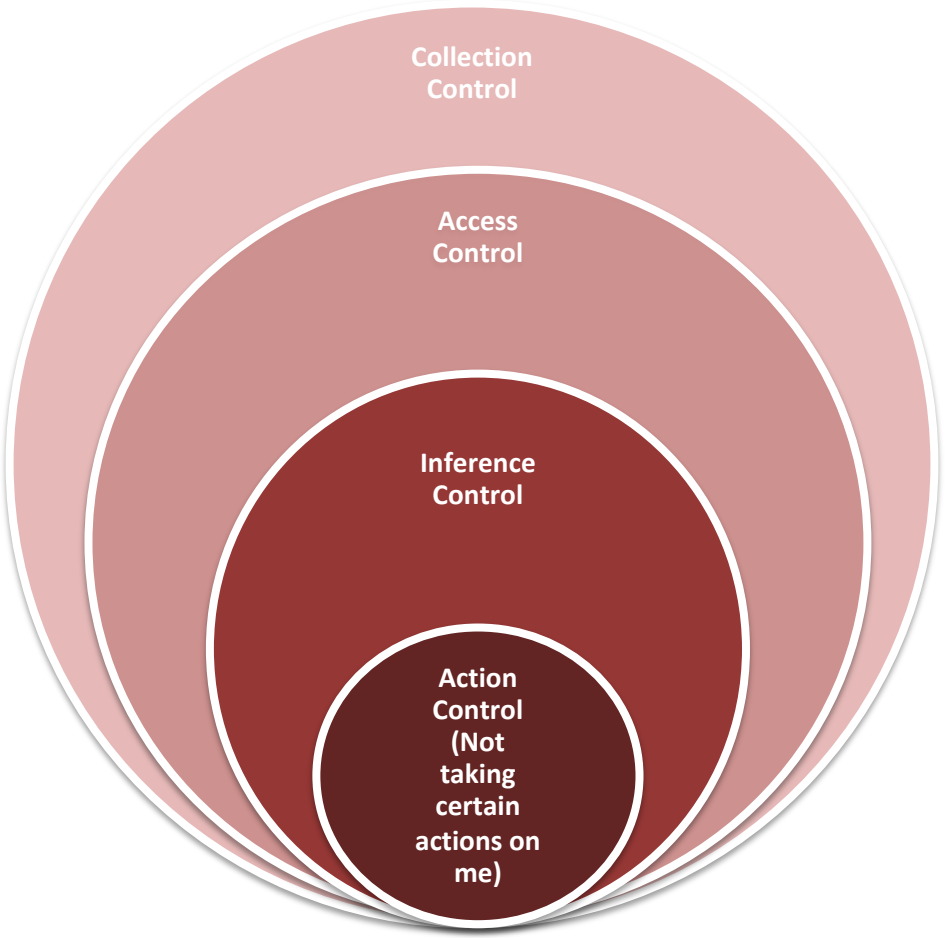
Bias, Equity, & Fairness

Transparency

Trustworthiness and Accountability

Privacy, Inference, and Levels of control

While we may want controls at all levels, we want stronger controls as we go down these levels



Data and AI Ethics Issues

Privacy

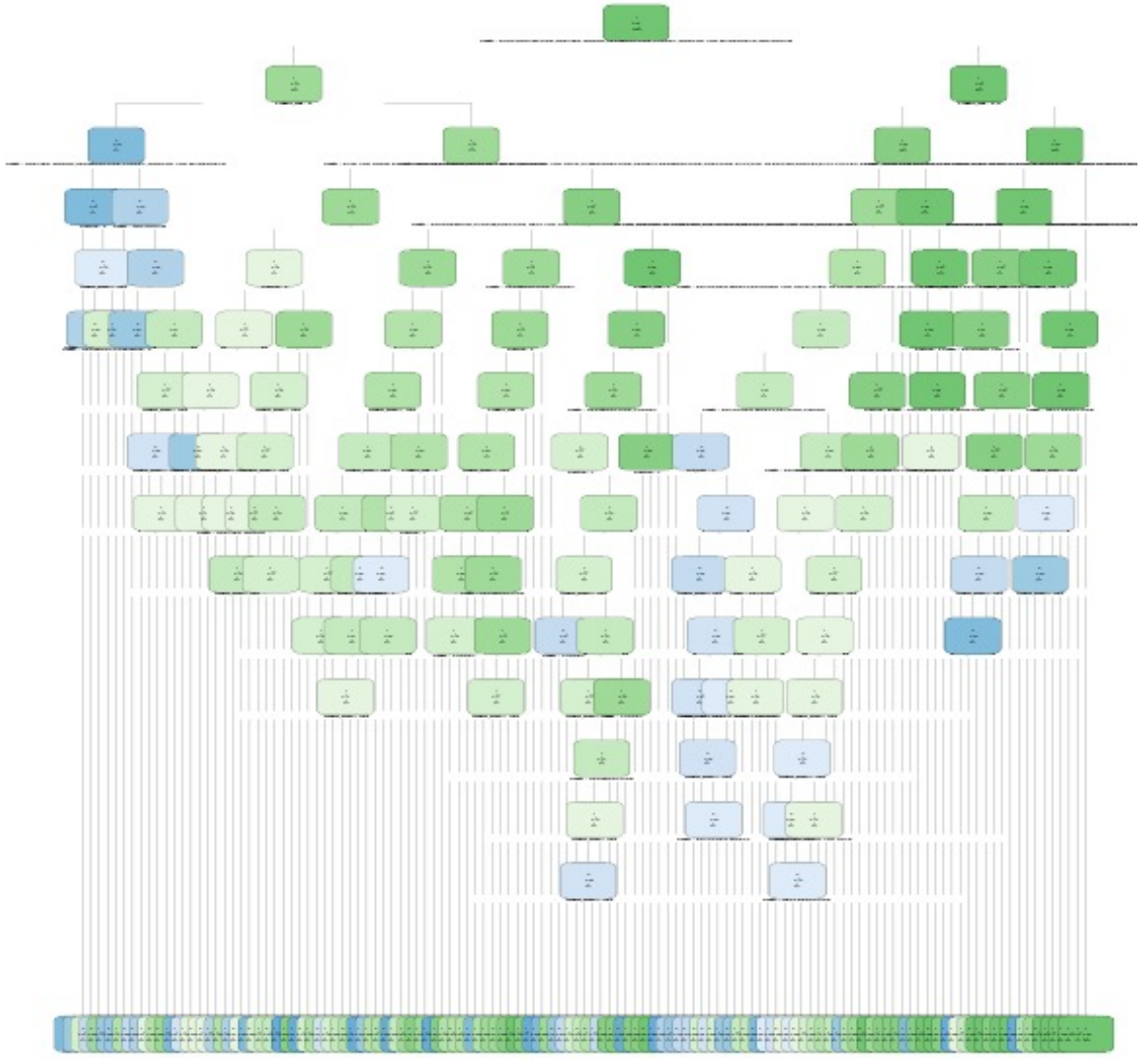
Data Ownership

Bias, Equity, & Fairness

Transparency

Trustworthiness and Accountability

Transparency is not about displaying the model that was built



What does it take for an analysis to be transparent?

- Code for the analysis
- Model that was built
- Data that was used
- …

We have to be able to explain and disclose

- Processes—what the analysis pipeline and algorithm does
- How it does it
- Who controls the algorithm
- What data does it use
- What is it being asked to optimize for
- How well did it do what it was asked to do
- What was its impact on different types of people