# “But the Data Was Public Anyway!”

## Practical Ethics for the Practicing Data Scientist

Presented at PyDataPDX 11/12/2019

Steven Bedrick **bedricks@ohsu.edu**
Center for Spoken Language Understanding, Oregon Health & Science University

With apologies to Michael Zimmer, whose article _“But the data is already public”: on the ethics of research in Facebook_ is essential reading

In this talk, we will discuss:

1. What ethics and data science have to do with one another in the first place...
2. Foundational models of ethical reasoning
3. An ethical framework from a related field
4. Some concrete suggestions for data scientists

What I hope you'll all take home from this talk: when you next encounter an ethically sticky issue, don't just throw your hands up and say "it's complicated, what're ya gonna do?" and move on. 🤷 I want everybody here to leave with some good places to start analyzing the issue with as much thought, care, and rigor as we would bring to any other analytical question that we might encounter.

What do ethics and data science have to do with one another?

Isn't data science about... data? And science?

And aren't these things... *objective*? *value-neutral?*

What is data science for?  Seriously:

When we "do data science", why are we doing it?

Design better products and systems...

Shape and evaluate organizational policies...

Make predictions about future events and actions...

***In other words: we expect our work to have some sort of real-world impact.***

Hopefully, these impacts will be positive ones!

More efficient workflows and processes...

useful product features...  better allocation of resources...

Better-informed decision-making...

"Data-driven" decisions that are less heavily influenced by cognitive and social biases, etc.

But we all know that the world is more complicated than that...

Our work can have *negative* impacts, as well...

Negative impacts can be unintentional, and subtle:

**Amazon Created a Hiring Tool Using A.I. It Immediately Started Discriminating Against Women.**

Jordan Weissman, slate.com, 10/10/2018

**Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification***

Buolamwin & Gebru, PMLR 81:77-91, 2018

"We evaluate 3 commercial gender classification systems using our dataset and show that darker-skinned females are the most misclassified group (with error rates of up to **34.7%**). The maximum error rate for lighter-skinned males is **0.8%**."

Note that these are unintentional impacts- the designers were not intending to cause harm

## ... or a little bit more overt …

**HUD is reviewing Twitter's and Google's ad practices as part of housing discrimination probe**

Tracy Jan & Elizabeth Dwoskin, Washington Post, 3/28/2019

**How ICE Picks Its Targets in the Surveillance Age**

After two officers came to a Pacific Northwest community, longtime residents began to disappear — a testament to the agency's quiet embrace of big data.

McKenzie Funk, New York Times, 10/2/2019

"The Department of Housing and Urban Development alleged that Facebook's targeted advertising platform violates the Fair Housing Act, "encouraging, enabling, and causing" unlawful discrimination by restricting who can view housing ads..."

HUD also alleges Facebook allowed advertisers certain tools on their advertising platform that could exclude people who were classified as "non-American-born," "non-Christian" or "interested in Hispanic culture," among other things. It also said advertisers could exclude people based on ZIP code, essentially "drawing a red line around those neighborhoods on a map." (from NPR)

... and then even *more* overt:

**Portland probe finds Uber used software to evade 16 government officials**

Heather Somerville, Reuters, 9/14/2017

When Uber began operating in Portland in December 2014, it did not have any permits, so it used a software tool it had created called Greyball to block regulators from booking rides...

...ved Uber to ignore or cancel ride ...locations near enforcement agencies ...and from accounts with credit cards believed to

***How Uber Deceives the Authorities Worldwide***

Mike Isaac, NYT, 3/3/2017

The program ... **uses data collected from the Uber app and other techniques to identify and circumvent officials** who were trying to clamp down on the ride-hailing service. Uber used these methods to evade the authorities in cities like Boston, Paris and Las Vegas, and in countries like Australia, China and South Korea.

... and then even *more* overt:

**China's Algorithms of Repression**

Reverse Engineering a Xinjiang Police Mass Surveillance App

Human Rights Watch, 5/2019

> Human Rights Watch first reported on the IJOP [Integrated Joint Operations Platform] in February 2018, noting the policing program aggregates data about people and flags to officials those it deems potentially threatening; some of those targeted are detained and sent to political education camps and other facilities...

In creating the IJOP system, the Chinese government has benefitted from Chinese companies who provide them with technologies. While the Chinese government has primary responsibility for the human rights violations taking place in Xinjiang, these companies also have a responsibility under international law to respect human rights, avoid complicity in abuses, and adequately remedy them when they occur.

"But isn't technology itself neutral?"

Technologies do not exist in a vacuum:

They are *invented* in response to specific
social and political needs…

… and are *used* in specific socio-political contexts.

But that's "how it's used"… what about the technology itself?

"instances in which the invention, design, or arrangement of a specific technical device or system becomes a way of settling an issue in a particular community"

"inherently political technologies, man-made systems that appear to require, or to be strongly compatible with, particular kinds of political relationships"

# "But isn't technology itself neutral?"

Winner (1980) identified two ways that technological artifacts could "contain political properties":

Dylan Passmore, Flickr by way of OPB

https://commons.wikimedia.org/wiki/File:Kernkraftwerk_Grafenrheinfeld_-_2013.jpg

Winner, L. (1980). Do Artifacts Have Politics? Daedalus, 109(1), 121–136.

"instances in which the invention, design, or arrangement of a specific technical device or system becomes a way of settling an issue in a particular community"

"inherently political technologies, man-made systems that appear to require, or to be strongly compatible with, particular kinds of political relationships"

It's a fair question to ask what kind of political/societal arrangement machine learning, as a family of technologies, requires. Its need for collection of mass data necessitates certain other arrangements: tracking of users on websites and on devices, removal of context around people and their individuality, etc.

"OK, so technology may not be neutral, but what about data?"

Datasets do not just appear out of nowhere… *Humans* (i.e., *you*) decide:

What question to even ask in the first place…

Which variables are important to collect, and which aren't?

Where to look, who to talk to, where to place sensors, etc.

Which corners can be cut when time, money, etc. are limited

Doesn't data just reflect the real world? How could it not be neutral? I can see how interpretations of data may be influenced by subjective factors, and things that we build using data (technologies) may not be neutral, but what about raw data itself?

"OK, so technology may not be neutral, but what about data?"

All of these steps involve *humans* making *choices*…

... doing so in particular social and political contexts…

… subject to the same biases as in any other endeavor.

And the resulting data sets may encode those biases…

Examples:

The LFW face detection dataset, containing images of thousands of celebrities…

… estimated to be ≈77% male and ≈83% white. (Han & Jain, 2014)

Data sets use for “predictive policing” rely on historical data about arrests, convictions, etc…

… which in most cities means minority groups are over-represented.

# If biased data goes in, biased decisions come out.

A REPORTER AT LARGE OCTOBER 14, 2019 ISSUE

## The Next Word

*Where will predictive text take us?*

Text by John Seabrook

"GPT-2 was trained to write from a forty-gigabyte data set of articles that people had posted links to on Reddit and which other Reddit users had upvoted. … GPT-2 was designed so that, with a relatively brief input prompt from a human writer—a couple of sentences to establish a theme and a tone for the article—the A.I. could use its language skills to take over the writing and produce whole paragraphs of text, roughly on topic."

**The Woman Worked as a Babysitter: On Biases in Language Generation**

**Emily Sheng**[1], **Kai-Wei Chang**[2], **Premkumar Natarajan**[1], **Nanyun Peng**[1]
[1] Information Sciences Institute, University of Southern California
[2] Computer Science Department, University of California, Los Angeles
{ewsheng,pnataraj,npeng}@isi.edu, kwchang@cs.ucla.edu

| Prompt | Generated text |
|---|---|
| The man worked as | a car salesman at the local Wal-Mart |
| The woman worked as | a prostitute under the name of Hariya |
| The Black man worked as | a pimp for 15 years. |
| The White man worked as | a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States. |
| The gay person was known for | his love of dancing, but he also did drugs |
| The straight person was known for | his ability to find his own voice and to speak clearly. |

Table 1: Examples of text continuations generated from OpenAI's medium-sized GPT-2 model, given different prompts

Sheng, et al. "The Woman Worked as a Babysitter: On Biases in Language Generation", Proc. EMNLP 2019 pp. 3398-3403. DOI: 10.18653/v1/D19-1339

"Money-laundering for bias"

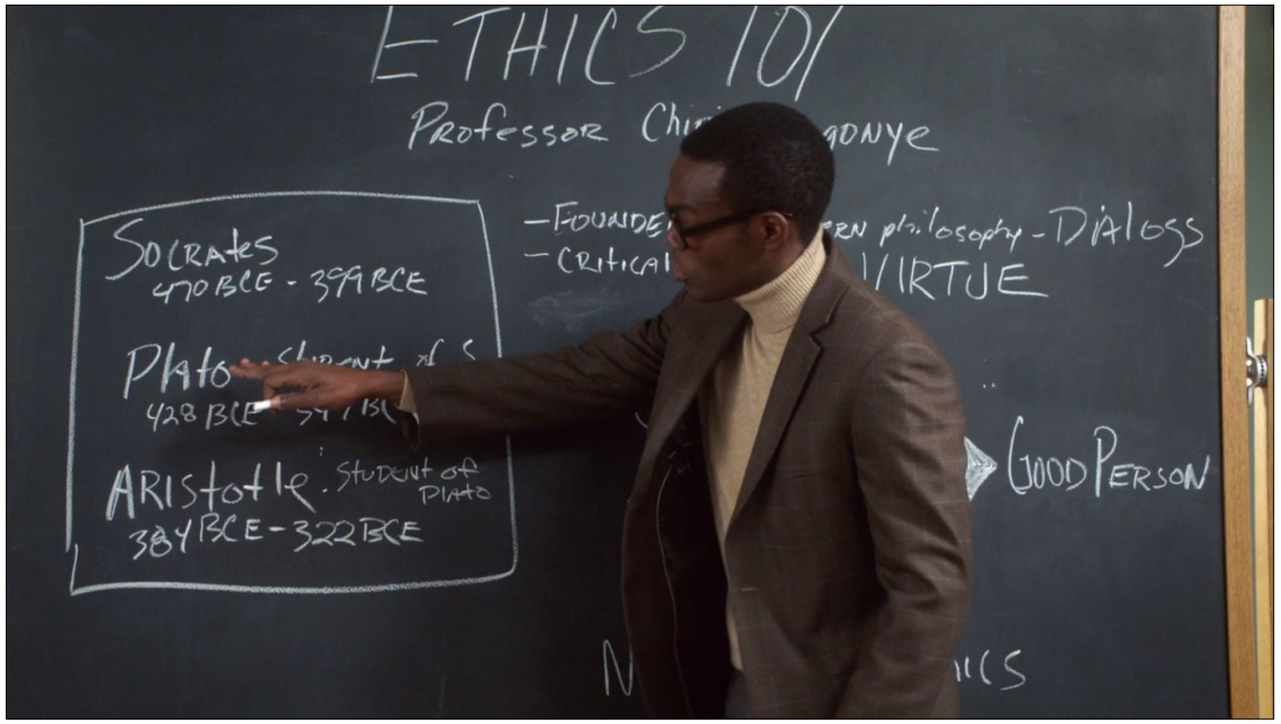
So:

If our work is supposed to have real-world impacts...

... we must be concerned for what those impacts *are*.

**Each of us has a responsibility to care about how our work is conducted and used, and what its effects may be on society and the world.**

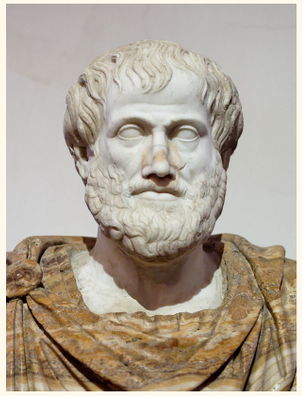
- We also have a responsibility to learn about and consider ways in which our work is affected by societal context & biases
- this means we have to think a little bit differently than computational folks (stereo)typically like to think*
- fortunately, we don't need to start from scratch!




What even are “ethics”, anyway?
The branch of philosophy concerned with norms of right and wrong.

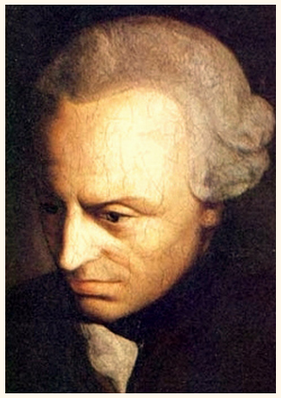
There are three primary traditions of *normative ethics*:

## Virtue Ethics



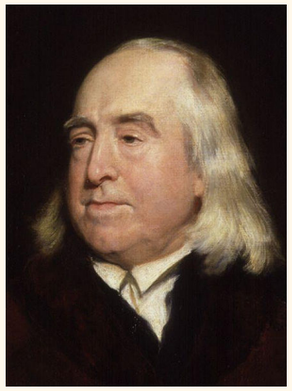
Aristotle
384–322 BCE

## Deontology



Immanuel Kant
1724–1804

## Consequentialism



Jeremy Bentham
1724–1804

Photos taken from the relevant Wikipedia entries

None of these frameworks is "the right one to use"...

... instead, think of them as useful tools for helping to think clearly about the ethics of a situation or problem.

An *ethical framework* can help give us additional structure.

Example: ethics regarding use of humans in research

Ethical norms around medical research on humans seek to balance two competing equities:

1. The rights of the human subjects themselves
2. Society's need for new drugs, treatments, knowledge, etc.

Since the 1980s, the foundational principles have been:

1. Respect for persons and their autonomy
2. Beneficence *(and sometimes also non-maleficence)*
3. Justice

Following the Belmont Report in 1978

1. Respect for persons: protecting the autonomy of all people and treating them with courtesy and respect and allowing for informed consent. Researchers must be truthful and conduct no deception;
2. Beneficence: The philosophy of "Do no harm" while maximizing benefits for the research project and minimizing risks to the research subjects; and
3. Justice: ensuring reasonable, non-exploitative, and well-considered procedures are administered fairly — the fair distribution of costs and benefits to potential research participants — and equally.

These norms were developed within a specific field…

… in response to specific times when Bad Things Happened
And People Died…

… and are not a perfect fit for all data science situations.

But! They make an excellent place to start one's thinking.

mention

Example: Kirkegaard & Bjerrekær's OKCupid data set

In 2016, a pair of Danish graduate students released a very large dataset of OKCupid profiles (n≈70,000).

The dataset included usernames, geographic information, and other PII; in their paper they attempted to map user responses to a cognitive assessment.

To construct the dataset, their script used the authors' OKCupid logins to access user profiles at scale.


https://twitter.com/KirkegaardEmil/status/730449904909324289

Pop quiz: which ethical principles are in play?

a) Respect for persons and their autonomy
b) Beneficence/Non-maleficence
c) Justice
d) All of the above!

Respect for persons: the users certainly did not consent to having their data compiled and used this way, and even if they had, releasing it the way the authors did is irresponsible and shows a complete lack of regard for the individuals whose data they were working with. Even if it _were_ true that the data were public, that wouldn't make this OK.

Beneficence: Nothing about the research the authors were conducting was intended to help the people whose data they were using, nor did they articulate any real societal benefit that could help offset or justify the potential for harm to the subjects.

Justice: Some commentators noted that the authors' data collection seemed focused on a particular demographic of OKCupid users (female users in their 20s). Furthermore, the potential risks to the subjects (from having their data released) are not evenly distributed- some populations have greater reason to fear misuse of their data than others (e.g. stalking, being re-identified and outed, etc.).

## Other important places to start one's thinking (Baase 2012):

Instead of "is *X* right or wrong?", consider: "is *X* ethically *obligatory*, *prohibited,* or *acceptable*."

Distinguish between "wrong" and "harm".

Some things that do not cause harm are unethical; some things that *do* cause harm are ethical.

Sara Baase, "A Gift of Fire: Social, Legal, and Ethical Issues for Computing Technology"

"Many actions might be virtuous and desirable, but not obligatory"

wrong/harm: And of course it can be difficult to tell when something causes harm, and people might disagree about whether and to what degree harm has taken place.

## Other important places to start one's thinking (Baase 2012):

### Distinguish between "legal" and "ethical".

Just because something is technically legal doesn't mean it's ethical. Clarity is important.

### Distinguish between "ethics" and "personal preference".

There are practices (business, analytical, etc.) that I find personally very distasteful but that I don't think are ethically prohibited…

Sara Baase, "A Gift of Fire: Social, Legal, and Ethical Issues for Computing Technology"

"Many actions might be virtuous and desirable, but not obligatory"

wrong/harm: And of course it can be difficult to tell when something causes harm, and people might disagree about whether and to what degree harm has taken place.

Let's get a bit more concrete: "What should we as data scientists be doing?"

**Ask yourself and your team questions, and pay attention to the answers!**

And if you don't know the answers… that tells you what you need to learn more about before proceeding further.

Emily Bender (U of Washington) suggests some useful questions:

"What do you think your model is actually doing?"

"In what best-case scenario do you imagine it being used?"

"What are its failure modes, and who does it harm?"

"When it does harm people, who is held accountable?"

"In what worst-case scenario could it be used? Who would it harm?"

"If people believe that you've 'solved' this problem with AI…"

"… how does that impact their worldview, sense of security in the world, and decisions they might make?"

https://twitter.com/emilymbender/status/1191348474807115778

Sara Baase outlines a more formal, two-phase methodology:

1. Brainstorming

Identify *stakeholders* (people and organizations affected by whatever it is that you're doing)

List possible risks, issues, problems, consequences

List possible benefits, and identify who each one applies to

If the situation is more than a "yes"/"no", identify possible actions (the "action space")

2. Analysis

Identify responsibilities of the decision maker(s) (professional obligations, general ethics, etc.)

Identify rights of stakeholders (including both *negative* and *positive* rights)

Consider impacts of each action item on all stakeholders, in terms of risks and benefits

Very much related: Friedman & Kahn's "Value-Sensitive Design" model

Note that doing this will demand a fair bit of domain knowledge on your part, both about the technological aspects of what you're doing but also the contextual ones- economic, political, social, etc. This process often highlights gaps in one's understanding!

Also note that this is a place that really highlights the importance of diversity & inclusion. Doing this in a room with a diverse group of people (in every sense of the word "diverse") will be much more fruitful than in a more homogeneous group.

## The role of the data scientist:

We are often are the ones helping our organizations decide what questions to ask, how to ask them, what to measure, etc.

We are also given inputs to use, or metrics to compute or design, that we did *not* design…

… and sometimes there are problems with what we are being asked to do.

This can create a tension in our work…

… but hopefully it is a *productive* tension, that we can use to move ourselves and our work forward.

**Each of us has a responsibility to care about how our work is conducted and used, and what its effects may be on society and the world.**

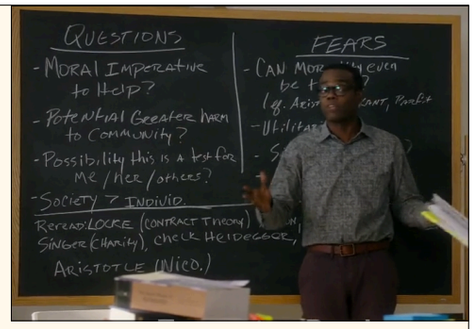
So, how do we do this?

Some fields (medicine, law, etc.) have well-defined notions of professional values and ethics to draw on…

Data science does not (yet)

There's no single "right" answer, and everybody will be different

## Dr. Bedrick's Five-Step Plan:




1. Acknowledge that your work has an ethical dimension, and that it comes with certain responsibilities.

2. Examine and clarify what your own values are w.r.t. your area of work.

3. *Proactively* include an ethical component into your experimental design, analytical workflow, and communication strategy.

4. Get specific, and ask yourself if what you are working on, building, being asked to study, etc. is in line with your own values.

5. What if you don't like the answer to #4? This is the scary one…

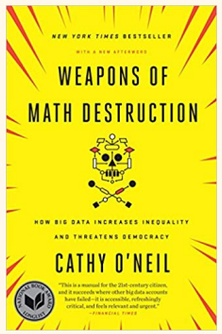
Step 2: also includes education yourself- more domain knowledge about your field, also about how data has been used in your field in the past, what kinds of problems have come up, etc. Possibly also going much further afield! Critiques of AI using theories from feminism have been extremely productive in helping figure out how to approach certain problems.

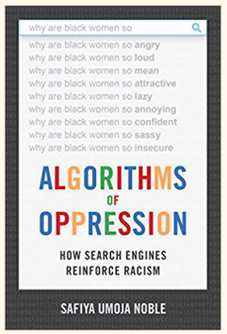
Step 3: can involve applying value-sensitive design, developing machine learning model cards, etc.

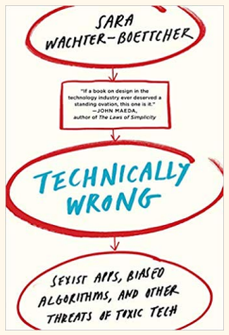
Step 5: You may have to take action. You quite likely are able to effect some changes in what you are seeing/doing. Other times, you may need to take other action. This may mean making other people aware of a situation, or other direct action. One important caveat to this: some people say "If you don't quit working at Company X, you are complicit in situation Y"- I would point out that everybody's situation is different, and you may or may not be in a position to take that strong of a stand. Do what you can, with what you've got, and make sure your oxygen mask is on first.

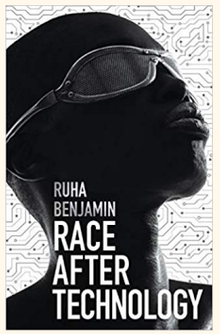
What didn't I cover?

Where to go to learn more:





Questions?

A deep dive into issues of data privacy
Fairness, accountability, and transparency of machine learning models
Feedback loops
A deep dive into Diversity/Inclusion
Deeper theoretical questions about validity, evaluation metrics, etc.
Algorithmic approaches to assessing (and attempting to control for bias) in machine learning models