CHAPTER 26
Data Ethics

Grub first, then ethics.
—Bertolt Brecht

What Is Data Ethics?

With the use of data comes the misuse of data. This has pretty much always been the case, but recently this idea has been reified as "data ethics" and has featured somewhat prominently in the news.

For instance, in the 2016 election, a company called Cambridge Analytica improperly accessed Facebook data and used that for political ad targeting.

In 2018, an autonomous car being tested by Uber struck and killed a pedestrian (there was a "safety driver" in the car, but apparently she was not paying attention at the time).

Algorithms are used to predict the risk that criminals will reoffend and to sentence them accordingly. Is this more or less fair than allowing judges to determine the same?

Some airlines assign families separate seats, forcing them to pay extra to sit together. Should a data scientist have stepped in to prevent this? (Many data scientists in the linked thread seem to believe so.)

"Data ethics" purports to provide answers to these questions, or at least a framework for wrestling with them. I'm not so arrogant as to tell you *how* to think about these things (and "these things" are changing quickly), so in this chapter we'll just take a quick tour of some of the most relevant issues and (hopefully) inspire you to think about them further. (Alas, I am not a good enough philosopher to do ethics *from scratch*.)

No, Really, What Is Data Ethics?

Well, let's start with "what is ethics?" If you take the average of every definition you can find, you end up with something like *ethics* is a framework for thinking about "right" and "wrong" behavior. *Data* ethics, then, is a framework for thinking about right and wrong behavior involving data.

Some people talk as if "data ethics" is (perhaps implicitly) a set of commandments about what you may and may not do. Some of them are hard at work creating manifestos, others crafting mandatory pledges to which they hope to make you swear. Still others are campaigning for data ethics to be made a mandatory part of the data science curriculum—hence this chapter, as a means of hedging my bets in case they succeed.

Curiously, there is not much data suggesting that ethics courses lead to ethical behavior, in which case perhaps this campaign is itself data-unethical!

Other people (for example, yours truly) think that reasonable people will frequently disagree over subtle matters of right and wrong, and that the important part of data ethics is committing to *consider* the ethical consequences of your behaviors. This requires *understanding* the sorts of things that many "data ethics" advocates don't approve of, but it doesn't necessarily require agreeing with their disapproval.

Should I Care About Data Ethics?

You should care about ethics whatever your job. If your job involves data, you are free to characterize your caring as "data ethics," but you should care just as much about ethics in the nondata parts of your job.

Perhaps what's different about technology jobs is that technology *scales*, and that decisions made by individuals working on technology problems (whether data-related or not) have potentially wide-reaching effects.

A tiny change to a news discovery algorithm could be the difference between millions of people reading an article and no one reading it.

A single flawed algorithm for granting parole that's used all over the country systematically affects millions of people, whereas a flawed-in-its-own-way parole board affects only the people who come before it.

So yes, in general, you should care about what effects your work has on the world. And the broader the effects of your work, the more you need to worry about these things.

Unfortunately, some of the discourse around data ethics involves people trying to force their ethical conclusions on you. Whether you should care about the same things *they* care about is really up to you.

Building Bad Data Products

Some "data ethics" issues are the result of building *bad products*.

For example, Microsoft released a chat bot named Tay that parroted back things tweeted to it, which the internet quickly discovered enabled them to get Tay to tweet all sorts of offensive things. It seems unlikely that anyone at Microsoft debated the ethicality of releasing a "racist" bot; most likely they simply built a bot and failed to think through how it could be abused. This is perhaps a low bar, but let's agree that you should think about how the things you build could be abused.

Another example is that Google Photos at one point used an image recognition algorithm that would sometimes classify pictures of black people as "gorillas". Again, it is extremely unlikely that anyone at Google *explicitly decided* to ship this feature (let alone grappled with the "ethics" of it). Here it seems likely the problem is some combination of bad training data, model inaccuracy, and the gross offensiveness of the mistake (if the model had occasionally categorized mailboxes as fire trucks, probably no one would have cared).

In this case the solution is less obvious: how can you ensure that your trained model won't make predictions that are in some way offensive? Of course you should train (and test) your model on a diverse range of inputs, but can you ever be sure that there isn't *some* input somewhere out there that will make your model behave in a way that embarrasses you? This is a hard problem. (Google seems to have "solved" it by simply refusing to ever predict "gorilla.")

Trading Off Accuracy and Fairness

Imagine you are building a model that predicts how likely people are to take some action. You do a pretty good job (Table 26-1).

Table 26-1. A pretty good job

Prediction	People	Actions	%
Unlikely	125	25	20%
Likely	125	75	60%

Of the people you predict are unlikely to take the action, only 20% of them do. Of the people you predict are likely to take the action, 60% of them do. Seems not terrible.

Now imagine that the people can be split into two groups: A and B. Some of your colleagues are concerned that your model is *unfair* to one of the groups. Although the model does not take group membership into account, it does consider various other factors that correlate in complicated ways with group membership.

Indeed, when you break down the predictions by group, you discover surprising statistics (Table 26-2).

Table 26-2. Surprising statistics

Group	Prediction	People	Actions	%
A	Unlikely	100	20	20%
A	Likely	25	15	60%
B	Unlikely	25	5	20%
B	Likely	100	60	60%

Is your model unfair? The data scientists on your team make a variety of arguments:

Argument 1
: Your model classifies 80% of group A as "unlikely" but 80% of group B as "likely." This data scientist complains that the model is treating the two groups unfairly in the sense that it is generating vastly different predictions across the two groups.

Argument 2
: Regardless of group membership, if we predict "unlikely" you have a 20% chance of action, and if we predict "likely" you have a 60% chance of action. This data scientist insists that the model is "accurate" in the sense that its predictions seem to *mean* the same things no matter which group you belong to.

Argument 3
: 40/125 = 32% of group B were falsely labeled "likely," whereas only 10/125 = 8% of group A were falsely labeled "likely." This data scientist (who considers a "likely" prediction to be a bad thing) insists that the model unfairly stigmatizes group B.

Argument 4
: 20/125 = 16% of group A were falsely labeled "unlikely," whereas only 5/125 = 4% of group B were falsely labeled "unlikely." This data scientist (who considers an "unlikely" prediction to be a bad thing) insists that the model unfairly stigmatizes group A.

Which of these data scientists is correct? Are any of them correct? Perhaps it depends on the context.

Possibly you feel one way if the two groups are "men" and "women" and another way if the two groups are "R users" and "Python users." Or possibly not if it turns out that Python users skew male and R users skew female?

Possibly you feel one way if the model is for predicting whether a DataSciencester user will *apply* for a job through the DataSciencester job board and another way if the model is predicting whether a user will *pass* such an interview.

Possibly your opinion depends on the model itself, what features it takes into account, and what data it was trained on.

In any event, my point is to impress upon you that there can be a tradeoff between "accuracy" and "fairness" (depending, of course, on how you define them) and that these tradeoffs don't always have obvious "right" solutions.

Collaboration

A repressive (by your standards) country's government officials have finally decided to allow citizens to join DataSciencester. However, they insist that the users from their country not be allowed to discuss deep learning. Furthermore, they want you to report to them the names of any users who even *try* to seek out information on deep learning.

Are this country's data scientists better off with access to the topic-limited (and surveilled) DataSciencester that you'd be allowed to offer? Or are the proposed restrictions so awful that they'd be better off with no access at all?

Interpretability

The DataSciencester HR department asks you to develop a model predicting which employees are most at risk of leaving the company, so that it can intervene and try to make them happier. (Attrition rate is an important component of the "10 Happiest Workplaces" magazine feature that your CEO aspires to appear in.)

You've collected an assortment of historical data and are considering three models:

- A decision tree
- A neural network
- A high-priced "retention expert"

One of your data scientists insists that you should just use whichever model performs best.

A second insists that you not use the neural network model, as only the other two can explain their predictions, and that only explanation of the predictions can help HR institute widespread changes (as opposed to one-off interventions).

A third says that while the "expert" can offer *an* explanation for her predictions, there's no reason to take her at her word that it describes the *real* reasons she predicted the way she did.

As with our other examples, there is no absolute best choice here. In some circumstances (possibly for legal reasons or if your predictions are somehow life-changing) you might prefer a model that performs worse but whose predictions can be explained. In others, you might just want the model that predicts best. In still others, perhaps there is no interpretable model that performs well.

Recommendations

As we discussed in Chapter 23, a common data science application involves recommending things to people. When someone watches a YouTube video, YouTube recommends videos they should watch next.

YouTube makes money through advertising and (presumably) wants to recommend videos that you are more likely to watch, so that they can show you more advertisements. However, it turns out that people like to watch videos about conspiracy theories, which tend to feature in the recommendations.

At the time I wrote this chapter, if you searched YouTube for "saturn" the third result was "Something Is Happening On Saturn… Are THEY Hiding It?" which maybe gives you a sense of the kinds of videos I'm talking about.

Does YouTube have an obligation not to recommend conspiracy videos? Even if that's what lots of people seem to want to watch?

A different example is that if you go to google.com (or bing.com) and start typing a search, the search engine will offer suggestions to autocomplete your search. These suggestions are based (at least in part) on other people's searches; in particular, if other people are searching for unsavory things this may be reflected in your suggestions.

Should a search engine try to affirmatively filter out suggestions it doesn't like? Google (for whatever reason) seems intent on not suggesting things related to people's religion. For example, if you type "mitt romney m" into Bing, the first suggestion is "mitt romney mormon" (which is what I would have expected), whereas Google refuses to provide that suggestion.

Indeed, Google explicitly filters out autosuggestions that it considers "offensive or disparaging". (How it decides what's offensive or disparaging is left vague.) And yet sometimes the truth is offensive. Is protecting people from those suggestions the ethical thing to do? Or is it an unethical thing to do? Or is it not a question of ethics at all?

Biased Data

In "Word Vectors" on page 287 we used a corpus of documents to learn vector embeddings for words. These vectors were designed to exhibit *distributional similarity*. That is, words that appear in similar contexts should have similar vectors. In particular, any biases that exist in the training data will be reflected in the word vectors themselves.

For example, if our documents are all about how R users are moral reprobates and how Python users are paragons of virtue, most likely the model will learn such associations for "Python" and "R."

More commonly, word vectors are based on some combination of Google News articles, Wikipedia, books, and crawled web pages. This means that they'll learn whatever distributional patterns are present in those sources.

For example, if the majority of news articles about software engineers are about *male* software engineers, then the learned vector for "software" might lie closer to vectors for other "male" words than to the vectors for "female" words.

At that point any downstream applications you build using these vectors might also exhibit this closeness. Depending on the application, this may or may not be a problem for you. In that case there are various techniques that you can try to "remove" specific biases, although you'll probably never get all of them. But it's something you should be aware of.

Similarly, as in the "photos" example in "Building Bad Data Products" on page 357, if you train a model on nonrepresentative data, there's a strong possibility it will perform poorly in the real world, possibly in ways that are offensive or embarrassing.

Along different lines, it's also possible that your algorithms might codify actual biases that exist out in the world. For example, your parole model may do a perfect job of predicting which released criminals get rearrested, but if those rearrests are themselves the result of biased real-world processes, then your model might be perpetuating that bias.

Data Protection

You know a lot about the DataSciencester users. You know what technologies they like, who their data scientist friends are, where they work, how much they earn, how much time they spend on the site, which job postings they click on, and so forth.

The VP of Monetization wants to sell this data to advertisers, who are eager to market their various "big data" solutions to your users. The Chief Scientist wants to share this data with academic researchers, who are keen to publish papers about who becomes a data scientist. The VP of Electioneering has plans to provide this data to political campaigns, most of whom are eager to recruit their own data science organizations. And the VP of Government Affairs would like to use this data to answer questions from law enforcement.

Thanks to a forward-thinking VP of Contracts, your users agreed to terms of service that guarantee you the right to do pretty much whatever you want with their data.

However (as you have now come to expect), various of the data scientists on your team raise various objections to these various uses. One thinks it's wrong to hand the data over to advertisers; another worries that academics can't be trusted to safeguard the data responsibly. A third thinks that the company should stay out of politics, while the last insists that police can't be trusted and that collaborating with law enforcement will harm innocent people.

Do any of these data scientists have a point?

In Summary

These are a lot of things to worry about! And there are countless more we haven't mentioned, and still more that will come up in the future but that would never occur to us today.

For Further Exploration

- There is no shortage of people professing important thoughts about data ethics. Searching on Twitter (or your favorite news site) is probably the best way to find out about the most current data ethics controversy.
- If you want something slightly more practical, Mike Loukides, Hilary Mason, and DJ Patil have written a short ebook, *Ethics and Data Science*, on putting data ethics into practice, which I am honor-bound to recommend on account of Mike being the person who agreed to publish *Data Science from Scratch* way back in 2014. (Exercise: is this ethical of me?)