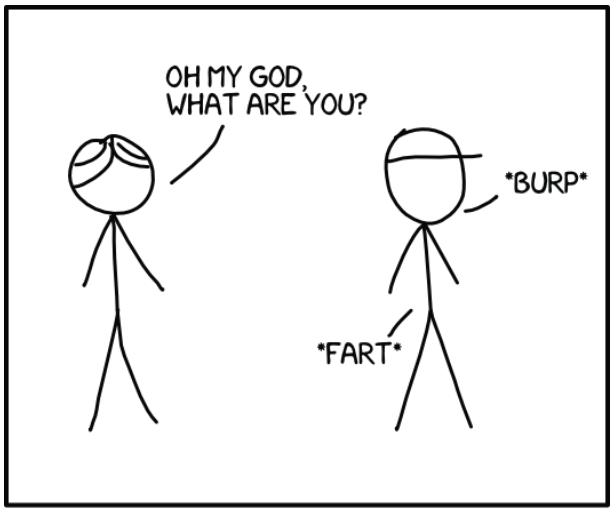
Machines Gone Wrong

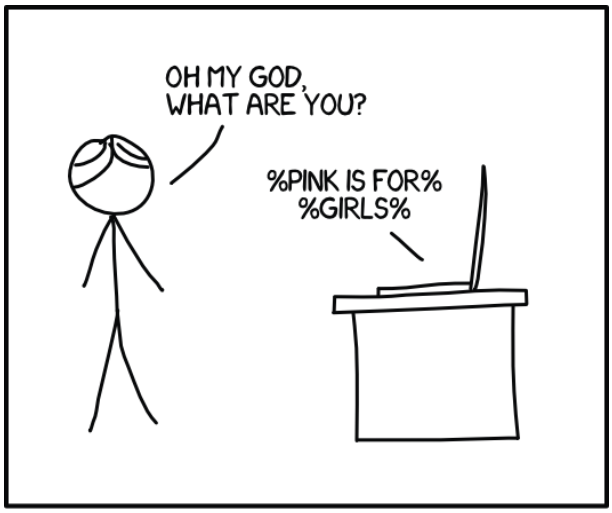
An online version of this is available at
https://machinesgonewrong.com.

Introduction

Often, when we first fall in love, the person of our affection seems to be perfect. But the happy honeymoon is cut short when we realize they are not *that* perfect. Turns out, they've got annoying habits. They wake up with bad breath. They burp. And oh my god their farts smell just as bad as ours.



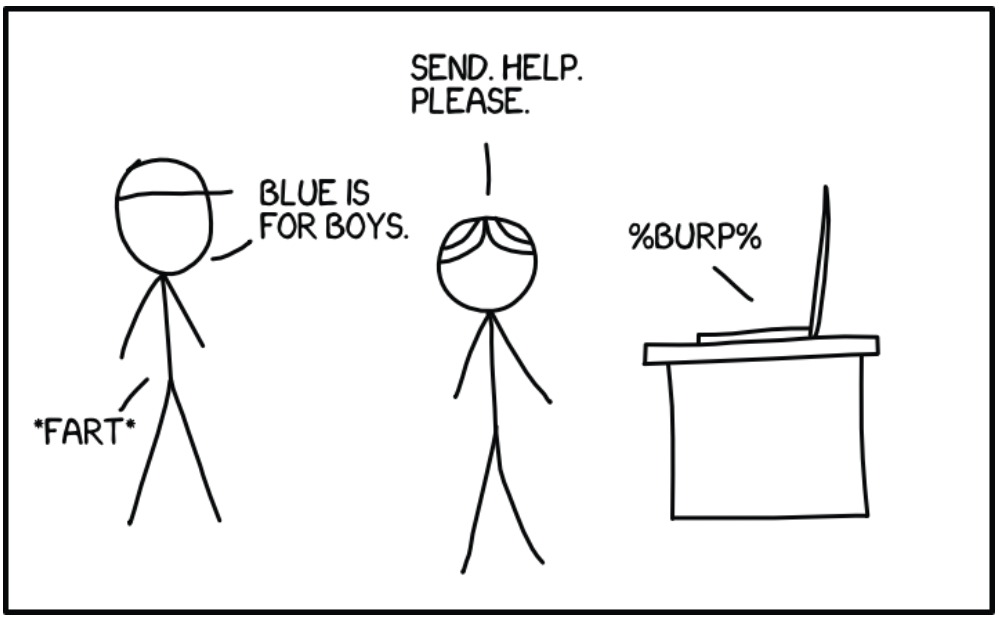
WHEN YOU REALIZE
THE LOVE OF YOUR LIFE IS HUMAN.



WHEN YOU REALIZE
THE AI IN YOUR LIFE IS BIASED.

In the same way, our honeymoon with artificial intelligence (AI) is quickly giving way to a realization that AI is not perfect. Turns out, AI is not neutral. It is not necessarily right or fair. The recommendations of AI systems can be just as sexist or racist as any human.

There are so many ways that AI can go wrong. There are so many guidelines from governments, companies, non-governmental organizations (NGOs). There are so many new algorithms, datasets and papers on ethical AI. It can all be a bit hard to take in, so this guide is here to help.



At the moment, the guide is targeted at AI practitioners and assumes some understanding of AI technologies. This mainly includes researchers and engineers. But it may also be useful for anyone helping to implement or recommend AI solutions.

The current version of the guide focuses on algorithmic bias. Future work will include other AI-related problems such as black boxes, privacy violations, ghost work and misinformation.

Here are some questions this guide tries to answer at different stages of the AI system lifecycle.

Taking Up the Project

- What are the different ways to define fairness?
- When is AI *not* the answer?

Collecting Data

- What are some possible sources of bias in datasets?
- What are some open-source datasets that are diverse?

Training and Evaluation

- What are possible sources of bias in the training process?
- What's wrong with using pre-trained models and external datasets?

Deployment and Maintenance

- What should our clients and users know?
- What are bias-related concerns when deploying an AI system?

Getting Started

AI ethics can be confusing. To practitioners, AI is kind of just clever mathematics. So how can a bunch of code and equations be ethical or unethical? Why are we so worried about AI ethics?

This section tries to give a warm-up to AI ethics before we dive into the deep end. It will cover the following:

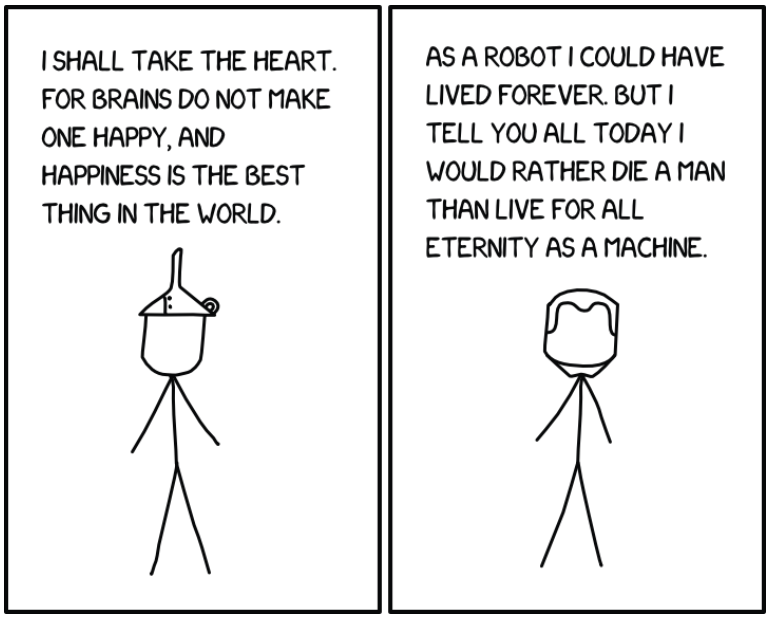
- What do we mean by AI ethics?
- What do we mean by AI systems?
- How is AI different from other technologies?
- What is the single most important question when implementing AI solutions?

Ethics of Artificial Intelligence (AI Ethics)

> On my view, computer ethics is the analysis of the nature and social impact of computer technology and the corresponding formulation and justification of policies for the ethical use of such technology.
>
> *What is Computer Ethics? - James H. Moor, 1985*

Discussions of AI ethics typically fall into two categories: how people treat AI (think Chappie and Bicentennial Man) and how AI treat people (think Terminator and HAL9000).

TREATMENT OF AI BY HUMANS



Anyone who has been touched by Robin Williams's portrayal of Andrew in Bicentennial Man might have thought about the idea of granting rights to robots and AI systems. In Life 3.0, Max Tegmark recounted a heated discussion between Larry Page and Elon Musk on robot rights.

> At times, Larry accused Elon of being “specieist”: treating certain life forms as inferior just because they were silicon-based rather than carbon-based.
>
> *Life 3.0 - Max Tegmark, 2017*

Realistically though, AI systems that require us to rethink notions of humanity and consciousness still remain on the far-flung horizon. Instead, let’s focus on the more urgent issue of how AI treat people.

AND TREATMENT OF HUMANS BY AI…

More urgently, we need to consider the effects of present AI systems on human moral ideals.

AI systems can promote human values. Low-cost automated medical diagnoses enable more accessible medical services. Fraud detection algorithms in banks help to prevent illegitimate transactions. Image recognition algorithms help to automatically detect images of child abuse and identify victims.

But AI can also violate human values. The use of generative models to create fake articles, videos and photos threatens our notion of truth. The use of facial recognition on public cameras disrupt our conventional understanding of privacy. The use of biased algorithms to hire workers and sentence criminals violate our values of fairness and justice.

The pervasive nature of AI systems means that these systems potentially affect millions and billions of lives. Many important institutions (political, judicial, financial) are increasingly augmented by AI systems. In short, it is critical to get things right before human civilization blows up in our faces. AI ethics goes beyond philosophical musings and thought experiments. It tries to fix the real problems cropping up from our new AI solutions.

… WHICH ARE ALSO DESIGNED BY HUMANS

For now at least, the implementation of AI systems is a manual non-automated process. So we really shouldn’t be thinking about how an AI system is violating human values. Keep in mind that the system was designed by humans and its designers are probably the ones who should be responsible for any ethical violations. In fact, all the instances of

“AI” above should be replaced with “human-designed AI”.

As such, AI ethics also consists of educating AI parents (aka human researchers and engineers) about how to bring up their AI babies. Because their AI babies grow up to become really influential AI adults. AI researchers and engineers have to understand the tremendous power and responsibility that they now possess.

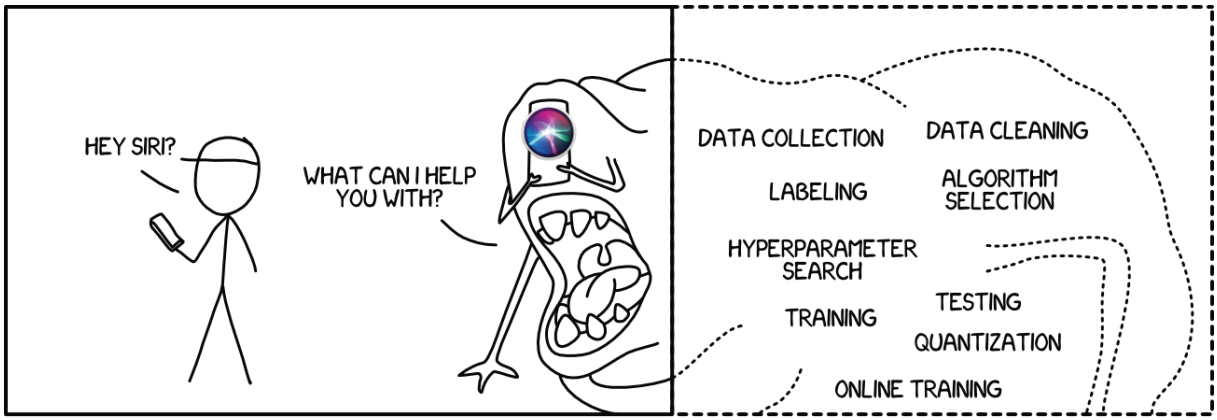
- What do we mean by AI ethics?

For the rest of this guide, AI ethics refers to the study of how AI systems promote and violate human values, including justice, autonomy and privacy. In particular, we note that current AI systems are still created, deployed and maintained by humans. And these humans need to start paying attention to how their systems are changing the world.

Artificial Intelligence Systems (AIS)

> An [Artificial Intelligence System (AIS)] is any computing system using artificial intelligence algorithms, whether it’s software, a connected object or a robot.
>
> *The Montréal Declaration - Université de Montréal, 2018*

The Montréal Declaration is a set of AI ethics guidelines initiated by Université de Montréal. In the Declaration, its 10 principles refers extensively to “AIS” instead of “AI”. This guide will do the same because the term “system” serves as a nice reminder that we are looking at a complex network of parts that work together to make a prediction.



Siri is not a tiny sprite that lives in iPhones. Siri is an entire digital supply chain from initial conception to data collection to model training to deployment to maintenance and finally retirement.

The same is true for any other AIS, including Google Translate, Amazon Rekognition and Northpointe's COMPAS. This big-picture perspective is important. It reminds us that we have to look at the entire system and infrastructure when we talk about AI ethics.

In addition to a digital supply chain, AIS also have physical supply chains that comprise energy usage, resource extraction and hardware recycling or disposal. These physical supply chains can be due to cloud servers, physical devices or simply the electricity and hardware used to train and house the models. The AI Now Institute also has a fantastic illustration titled Anatomy of an AI System that considers AIS in terms of "material resources, human labor, and data".

Finally, the "system" also includes the sociotechnical context where the AIS is applied. This refers to the culture, norms and values of the application, the domain and the geography and society that the application lives in. These values can be formalized (e.g. laws) or informal (e.g. unwritten customs and traditions). This sociotechnical context becomes critical when we talk about concepts like fairness and justice.

- What do we mean by AI systems?

The term Artificial Intelligence System (AIS) refer to the entirety of artificial intelligence applications or solutions, in terms of:

- **Digital lifecycle (conceptualization to retirement),**
- **Physical lifecycle (resource extraction to hardware disposal), and**
- **Sociotechnical context (culture, norms and values).**

What is different about AI?

There's been many articles talking about how AI is the shit and how it's better than every other technology we've had. Here we look at three aspects that make AI stand out in terms of its social impact - an illusion of fairness, tremendous speed and scale, and open accessibility.

ILLUSION OF FAIRNESS

Since machines have no emotions, we often assume that they would be impartial and make decisions without fear or favor.

This assumption is flawed. For one, guns too, have no capacity for prejudice or bias. But we don't attribute impartiality to guns. "Guns don't kill people, people kill people." A gun wielded by different people can have vastly different moral embeddings. The same can be said for AIS.

Moreover, the data used to train machine learning models can be a tremendous source of bias. A hiring model trained with sexist employment records would obviously suggest similarly sexist decisions. A recidivism model trained on racist arrest histories would obviously give racist suggestions. Like produces like. Garbage in, garbage out.

Unfortunately, AIS marketed as impartial and unbiased seem really appealing for all sorts of important decisions. This illusion of fairness provides unwarranted justification for widespread deployment of AIS without adequate control. But fairness is not inherent in AIS. It is a quality that has to be carefully designed for and maintained.

SPEED AND SCALE

The shipping industry revolutionized trade, enabling it to be conducted on an international scale across maritime trade routes. Previously lengthy land detours had much quicker maritime alternatives. But this increase in speed and scale also facilitated the rapid spread of the Black Death.

Many of today's AIS function on an unprecedented speed and scale. Google Translate serves over 500 million queries a day. Amazon's Rekognition claims to be able to perform "real-time face recognition across tens of millions of faces". Previously expensive, slow, one-to-one functions can now be automated to become cheaper, faster and serve much larger audiences. This means more people can benefit from AIS.

But just like the Black Death supercharged by rats on merchant ships, this crazy speed and scale also applies to any inherent problems. A biased translation system could serve over 500 million biased queries a day. An insecure facial recognition system can leak tens of millions of faces and related personal details. Speed and scale is a double-edged sword and it's surprising how people often forget that a double-edged sword is double-edged.

ACCESSIBILITY

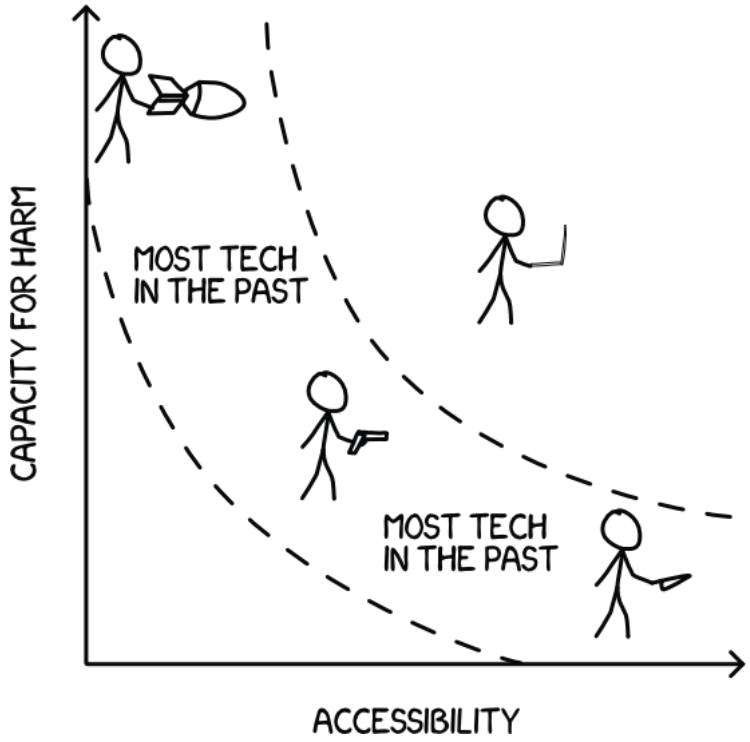
AI research has largely been open. As a self-taught coder and AI researcher, I remain eternally grateful for the kindness and generosity of the AI community. The vast majority of researchers share their work freely on arxiv.org and GitHub. Open-source software libraries and datasets are available to anyone with Internet access. There are abundant tutorials for anyone keen to train their own image recognition or language model.

Furthermore, advances in hardware mean that consumer-grade computers are sufficient to run many state-of-the-art algorithms. More resource-intensive algorithms can always be trained on the cloud via services such as Amazon Web Services, Google Cloud and Microsoft Azure.

The combination of accessible research, hardware, software and data means that many

people have the ability to train and deploy their own AIS for personal use. A powerful technology is now openly accessible to unregulated individuals who may use it for any purpose they deem fit. There has been cool examples of students using Tensorflow to predict wildfires and tons of other nice stuff.

But like speed and scale, this accessibility is also a double-edged sword. Consider the examples of DeepFakes and DeepNude. These open-source programs use Generative Adversarial Networks and variants of the pix2pix algorithm to generate realistic pornographic media of unwitting individuals. Accessible and powerful technology can also be used by irresponsible or malicious actors.



- How is AI different from other technologies?

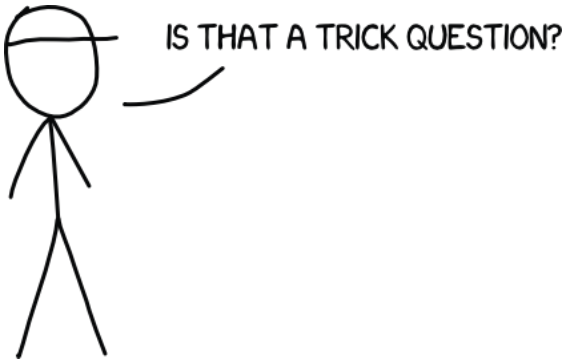
AI differs from most technologies in three aspects:

- **We tend to think AI is like totally fair and better than people.**
- **AI can be crazy fast and deployed on a massive scale.**
- **Given how powerful it is, AI is also really accessible to everyone.**

The Most Important Question

Cue drumroll

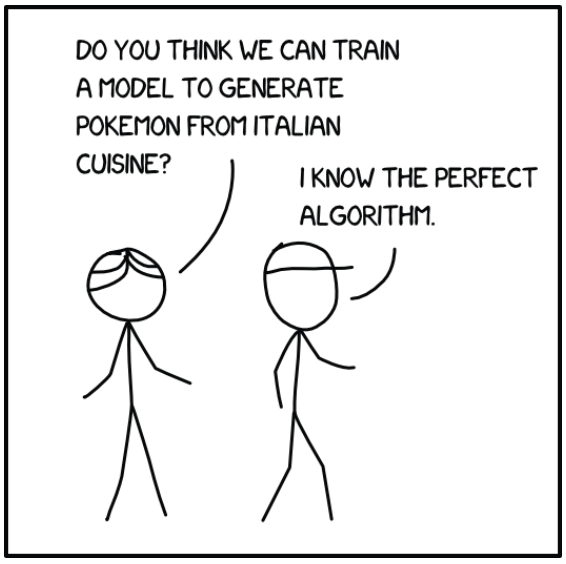
"WHEN IS AI NOT THE ANSWER?"



This is the most important question in this entire guide, and these days it can feel like the answer is, "Never."

This section here is to remind the reader that not using AIS *is* an option.

AI technologies have been used for facial recognition, hiring, criminal sentencing, credit scoring. More unconventional applications include writing inspirational quotes, coming up with Halloween costumes, inventing new pizza recipes and creating rap lyrics.



But the superiority of AIS should not be taken for granted despite all the hype. For example, human professionals are often far better at explaining their decisions, as compared to AIS. Most humans also tend to make better jokes.

It is immensely important to consider the trade-offs when deploying AIS and look critically at both pros and cons. In some cases, AIS may not actually offer significant benefits despite all the hype. Common considerations include explainability and emotional and social qualities, where humans far outperform machines.

“WHEN IS AI NOT THE ANSWER?”

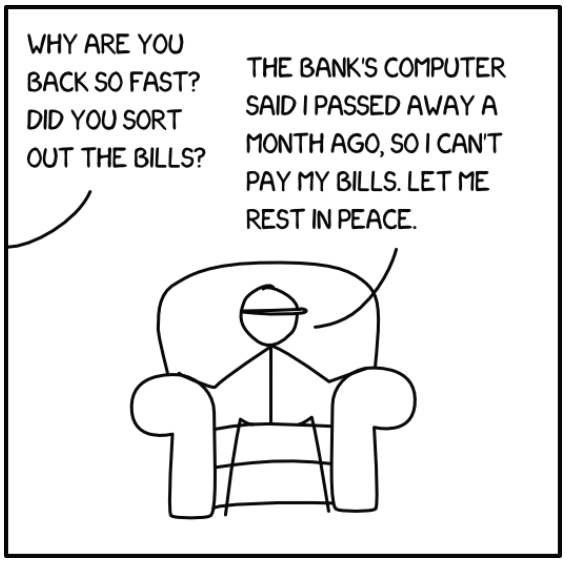
AI+Human systems are frequently perceived to be the best of both worlds. We have the empathy and explainability of humans augmented by the rigour and repeatability of AI systems. What could go wrong? Well, turns out documented experiences have shown that in such systems, humans might have a tendency to defer to suggestions made by the AIS. So rather than “ AI+Human”, these systems are more like “ AI+AgreeableHuman”.

In her book Automating Inequality, Virginia Eubanks notes that child welfare officers working with a child abuse prediction model would choose to amend their own assessments in light of the model's predictions.

> Though the screen that displays the [Allegheny Family Screening Tool (AFST)] score states clearly that the system "is not intended to make investigative or other child welfare decisions," an ethical review released in May 2016 by Tim Dare from the University of Auckland and Eileen Gambrill from University of California, Berkeley, cautions that the AFST risk score might be compelling enough to make intake workers question their own judgement.

> According to Vaithianathan and Putnam-Hornstein, **intake screeners have asked for the ability to go back and change their risk assessments after they see the AFST score**, suggesting that they believe that the model is less fallible than human screeners.
>
> *Automating Inequality - Virginia Eubanks, 2018*

Such observations are hardly surprising, given the daily exhortations of the reliability of machines. In fact, the human tendency to defer to automated decisions has been termed "automation bias". Unfortunately, this over-deference to machines potentially undermines the mutually complementary aspect of AI+Human models.



NEGLECTED RIPPLES

More generally, when discussing the pros and cons of adopting AIS solutions, we often forget to consider how the AIS might affect the humans interacting with the system i.e. cause "ripples" within the system. This is referred to the Ripple Effect Trap by Selbst et al. [7]. Examples of ripples include:

- **Automation bias**, as mentioned earlier. This refers to an unwarranted bias towards automated decisions. This might occur when people lack confidence in their own decisions, such as new or untrained personnel. It might also occur when the decision has severe consequences. People afraid of taking the blame for a wrong decision might prefer to transfer responsibility to the human-designed AIS.

- **Automation aversion**. The opposite of automation bias, this refers to a preference to disagree with automated decisions. This can arise from a fear of being displaced - "They took our jobs!" It can also be due to a bad history with poorly designed human-designed AIS or general mistrust due to negative media portrayals.

- **Overconfidence in AIS-derived decisions**. While the well-known fallibility of humans remind us to double and triple check decisions, employing human-designed AIS might create a false sense of security. This can arise over long-term experience with a generally reliable human-designed AIS. People might gradually take for

granted the reliability of the human-designed AIS. Consider the excruciating experiences of test drivers for self-driving cars, who have to be continuously alert despite a mostly safe ride.

- What is the single most important question when implementing AI solutions?

"Is using AI for this really a good idea?"

In other words, think hard about what using AI really means in the context of your problem. Like really hard. Not using AI is definitely an option.

And don't assume that AI+Human systems are definitely better than AI or humans by themselves. Instead, consider how AI and people might interact within your problem in unexpected ways. Ask prospective users what they think about AIS and factor their responses into your mental models.

References

1. The Montréal Declaration [link]
   Montréal, U.d., 2018. The Montréal Declaration for a Responsible Development of Artificial Intelligence, pp. 1-308.

2. What is computer ethics? [link]
   Moor, J.H., 1985. Metaphilosophy, Vol 16(4), pp. 266-275. Wiley Online Library.

3. Life 3.0: Being human in the age of artificial intelligence
   Tegmark, M., 2017. Knopf.

4. Automating inequality: How high-tech tools profile, police, and punish the poor
   Eubanks, V., 2018. St. Martin's Press.

5. Technological due process [link]
   Citron, D.K., 2007. Wash. UL Rev., Vol 85, pp. 1249. HeinOnline.

6. Automation bias and errors: are crews better than individuals? [link]
   Skitka, L.J., Mosier, K.L., Burdick, M. and Rosenblatt, B., 2000. The International journal of aviation psychology, Vol 10(1), pp. 85-97. Taylor & Francis.

7. Fairness and abstraction in sociotechnical systems [PDF]
   Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S. and Vertesi, J., 2019. Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 59-68.

Understanding Fairness

> All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.
>
> *Article I in the Universal Declaration of Human Rights*

To lay the ground for algorithmic bias, we first ask, "What does fairness mean?" And boy is this a big one. There are tons of definitions, so how do we know which one to pick? Why can't we all just agree on one?

This section acts as a primer to fairness, covering a few key concepts. It tries to answer the following questions:

- What is a widely used framework for fairness?
- How can we quantify fairness?
- Can't we just combine *all* of the fairness definitions?
- How do we design for fairness without context?
- How do we learn more about the context?

Disparate Treatment, Disparate Impact

Let's begin with a not-so-mathematical idea. A common paradigm for thinking about fairness in US labor law is disparate treatment and disparate impact.

Both terms refer to practices that cause a group of people sharing **protected characteristics** to be **disproportionately disadvantaged**. The phrase "protected characteristics" refers to traits such as race, gender, age, physical or mental disabilities, *where differences due to such traits cannot be reasonably justified*. Ideally, we should have a set of sensitive traits that we can check against. **But in reality, what constitutes "protected characteristics" varies by context, culture and country.** Next, the phrase "disproportionately disadvantaged" dismisses differences in treatment due to statistical randomness. To be frank, this is really vague but we will try to go into details in the next section.

The difference between disparate treatment and disparate impact can be summarized as explicit intent. Disparate treatment is explicitly intentional, while disparate impact is implicit or unintentional.

WHAT DOES THIS MEAN FOR AIS?

Let's use Amazon's Prime Free Same-Day service as an example. The Free Same Day service is a fantastic mind-blowing innovation that provides free same-day delivery. Since it's in its early stages, Amazon wants to trial the service before rolling it out to everyone. Suppose Amazon implements a model that decides which lucky neighborhoods should get first dibs on the Prime Free Same-Day service.

Disparate Treatment

Using race to decide who should get this service is certainly unjustified. So if Amazon had explicitly used racial composition of neighborhoods as an input feature for the model, that would be **disparate treatment**. In other words, disparate treatment occurs when protected characteristics are used as input features.

Obviously, disparate treatment is relatively easy to spot and resolve once we determine the set of protected characteristics. **We just have to make sure none of protected characteristics is explicitly used as an input feature.**

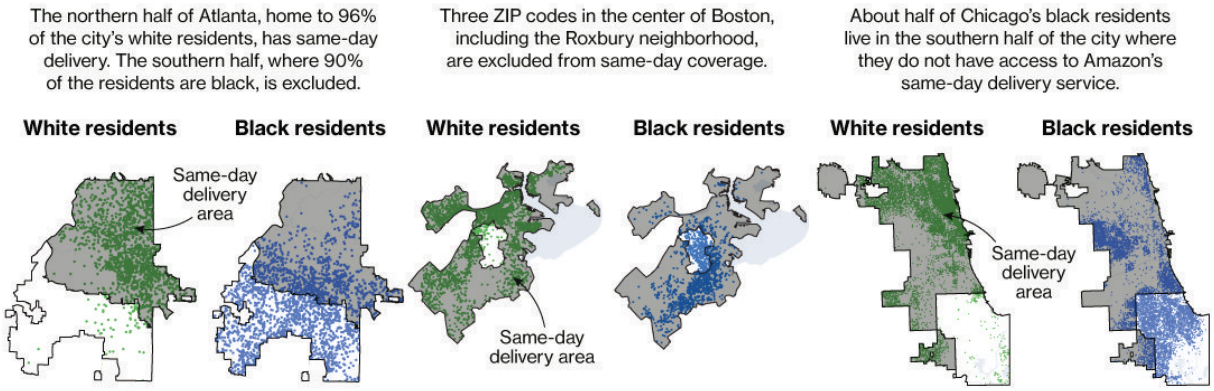
Disparate Impact

On the other hand, Amazon might have been cautious about racial bias and deliberately excluded racial features for their model. In fact, we can quote Craig Berman, Amazon's vice president for global communications, on this:

> Amazon, he says, has a "radical sensitivity" to any suggestion that neighborhoods are being singled out by race. "Demographics play no role in it. Zero."

> Amazon says its plan is to focus its same-day service on ZIP codes where there's a high concentration of Prime members, and then expand the offering to fill in the gaps over time.
>
> *Amazon Doesn't Consider the Race of Its Customers. Should It? - David Ingold and Spencer Soper, 2016*

Focusing on ZIP codes with high density of Prime members makes perfect business sense. But what if the density of Prime members correlates with racial features? The images below from the 2016 Bloomberg article by David Ingold and Spencer Soper shows a glaring racial bias in the selected neighborhoods.



Amazon Doesn't Consider the Race of Its Customers. Should It? - David Ingold and Spencer Soper, 2016

Despite not using any racial features, the resulting model appears to make recommendations that disproportionately exclude predominantly black ZIP codes. This unintentional bias can be seen as **disparate impact**.

In general, disparate impact occurs when protected characteristics are not used as input features but the resulting outcome still exhibits disproportional disadvantages.

Disparate impact is more difficult to fix since it can come from multiple sources, such as:

- A non-representative dataset e.g. using a training set that contains only white male faces but applying the trained model to everyone regardless of race or gender.
- A dataset that already encodes unfair decisions e.g. a credit scoring dataset with labels that underreports the credit score for black individuals.
- Input features that are proxies for protected characteristics e.g. postal code might be a proxy feature for race since racial and ethnicity demographics often have spatial correlations.

OKAY, BUT HOW DO WE KNOW HOW MUCH DISPARITY IS UNFAIR?

To answer that question, we have to review what we meant earlier by "disproportionately disadvantaged". In general, this has been rather hand-wavy, with good reason! What is unfair in one case might be justified in another, depending on the specific circumstances. And there are just so many factors to consider:

Let's say an insurance company uses an AIS that predicts whether an insuree will get into an accident within the next year. Insurees predicted as accident-prone could be charged higher premiums.

- If the model excessively predicts males as accident-prone, are males disproportionately disadvantaged?
- If the accuracies are different between age groups, are the age groups with worse accuracies disproportionately disadvantaged?
- What if the model overestimates accident-likelihood for certain races and underestimates it for other races? This means the first group pays higher premiums than they should, while the second group underpays. Then do we say the former group is disproportionately worse off and the latter is disproportionately better off?

On the other hand, there have been many attempts at trying to formalize and quantify fairness. Especially now that we have more computer scientists getting in on the game. The next section looks at some of these fairness metric.

- What is a widely used framework for fairness?

The terms "disparate treatment" and "disparate impact" are commonly used in US labor law, dividing discrimination into intentional and unintentional. Avoiding disparate treatment entails removing protected characteristics from the input features to the AIS. Avoiding disparate impact is slightly more complicated and we will discuss this in a later section.

A Fair Fat Pet Predictor

Suppose for a moment that our company organizes diet boot camps for overweight cats and dogs. We want to develop an AI system to help owners diagnose if a pet is overweight. Pets diagnosed as fat are then sent to our boot camps, which means less food and no treats boohoo. Furthermore, we know that dogs are more likely to be fat, as compared to cats. In fact, cats only have a 40% chance of being overweight, while dogs have a 60% chance of being overweight.

SOME BASICS BEFORE WE START

You can skip this section if you understand what are TP, FP, TN and FN. If these explanations are too long for comfort, check out the explorable on the website!

- **Positive** - What the model is predicting for. In our case, the model is predicting if a pet is fat. So a positive prediction is one that predicts a pet is fat. Despite this being super important for later definitions of fairness, this is unfortunately arbitrary because we can also say that the same model is predicting if a pet is not fat. In that case, a positive prediction is one that predicts a pet is not fat. But in general, this is clearly defined at the beginning when analyzing any model. TL;DR - for this example, positive refers to fat.
- **Negative** - Opposite of positive. In this case, negative refers to not fat.
- **Real Positives/Negatives** - The samples grouped by their actual labels. In this case, real positives refer to pets that are actually fat. Real negatives refer to pets that are actually not fat.
- **Predicted Positives/Negatives** - The samples grouped by their predictions. So predicted positives refer to pets that are predicted fat and predicted negatives refer to pets that are predicted not fat.
- **True Positives (TP)** - Predicted positives that are also real positives i.e. predicted positives that are correct. In our case, TP refers to fat pets correctly predicted fat.

- **True Negatives (TN)** - Predicted negatives that are also real negatives i.e. predicted negatives that are correct. Here, TN refers to pets that are not fat and correctly predicted as not fat.
- **False Positives (FP)** - Predicted positives that are actually real negatives i.e. predicted positives that are wrong. In our case, FP refers to pets that are not fat but misclassified as fat.
- **False Negatives (FN)** - Predicted negatives that are actually real positives i.e. predicted negatives that are wrong. Here, FN refers to fat pets wrongly predicted as not fat.

TUNING OUR MODEL FOR FAIRNESS

Here we will go through a few quantitative metrics for fairness. Again, for an interactive explanation, check out the explorable on the website!

Group Fairness

Both cats and dogs should have equal chances of being predicted fat.

The chance of a positive prediction (TP + FP) should be equal.

Equalized Odds

Both thin cats and thin dogs should have equal rates of false alarms (thin pets misdiagnosed as fat). Both fat cats and fat dogs should also have equal rates of escaping (fat pets misdiagnosed as thin).

Equal false positive rate (FPR) i.e. FP / Real Negatives and equal false negative rate (FNR) i.e. FN / Real Positives.

Conditional Use Accuracy Equality

Whether predicted fat or not, the probability of the prediction being correct should be equal for cats and dogs.

Equal positive predictive value (PPV) or precision i.e. TP / Predicted Positives and equal negative predictive value (NPV) i.e. TN / Predicted Negatives.

Overall Accuracy Equality

The probability of the prediction being correct should be equal for cats and dogs. This disregards the type of prediction.

Equal accuracy i.e. TP + TN.

Treatment Equality

The ratio of escaped fat animals to wrongly accused thin animals should be equal for cats and dogs. The idea here is that wrong predictions lead to either false alarms (FP) or escapes (FN). So the ratio of these two effects should be equal between cats and dogs.

Equal ratios of wrong predictions i.e. FP / FN.

MANY MORE METRICS

In addition to these, there are plenty more fairness metrics enumerated by Verma and Rubin and Narayanan. Some notable metrics include:

Calibration

This goes beyond true or false predictions and considers the score assigned by the model. For any predicted score, all sensitive groups should have the same chance of actually being positive.

> Suppose our fat pet predictor predicts a fatness score from 0 to 1 where 1 is fat with high confidence. If a cat and a dog are both assigned the same score, they should have the same probability of being actually fat.

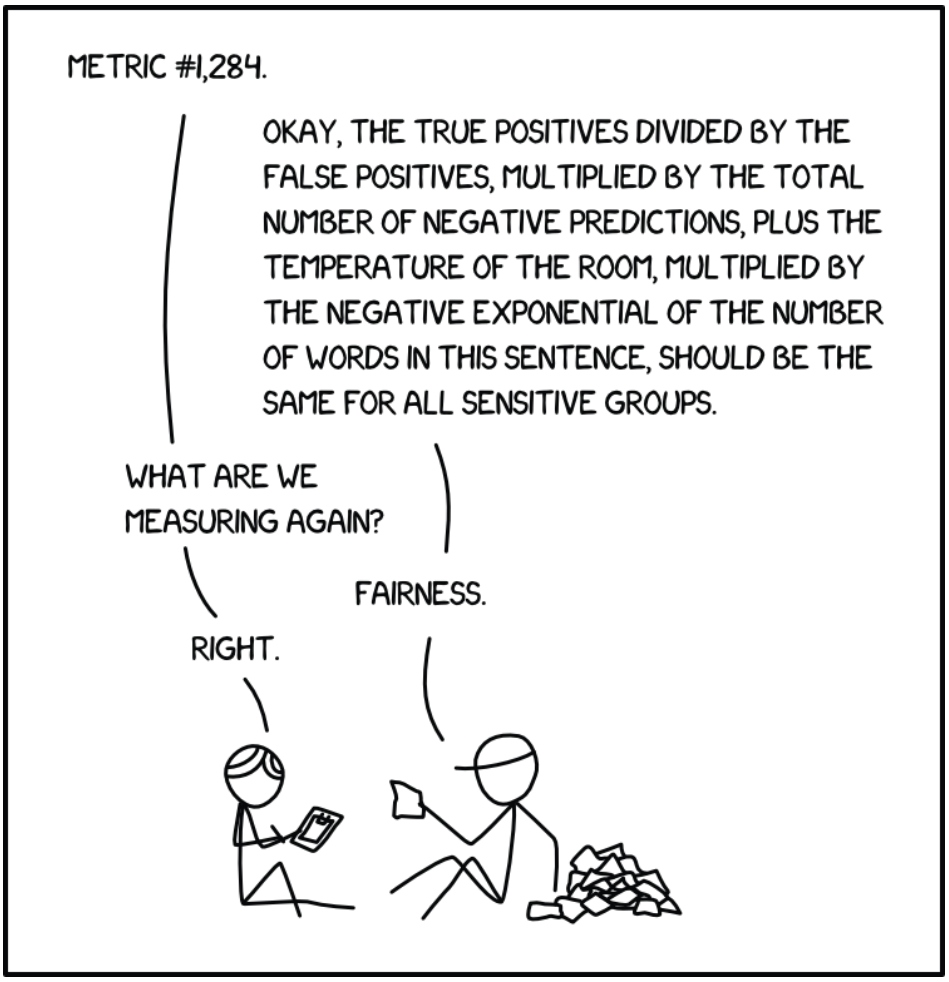
Well-calibration is a stricter form of calibration, with the added condition where the chance of being actually positive is equal to the score.

> For our fat pet predictor to be well-calibrated, the predicted fatness score has to be equal to the probability of actually being fat. For example, if a cat and a dog are both assigned the a score of 0.8, they should both have an 80% chance of being actually fat.

Fairness Through Awareness

This fairness metric is based on an intuitive rule - “treating similar individuals similarly”. Here, we first define distance metrics to measure the difference between individuals and difference between their predictions. An example of a distance metric could be the sum of absolute differences between normalized features. Then, this metric states that for a model to be fair, the distance between predictions should be no greater than the distance between the individuals.

Unfortunately, this leaves the difficult question of how to define appropriate distance metrics for the specific problem and application.



IS IT JUSTIFIED?

The awesome thing about these metrics is that they can be put into a loss function. Then we can train a model to optimize the function and voilà we have a fair model. Except, no it doesn't work like that.

A major issue with these metrics (besides the question of how to pick one) is that they neglect the larger context. In the previous section, we explained:

> The phrase "protected characteristics" refers to traits such as race, gender, age, physical or mental disabilities, where differences due to such traits cannot be reasonably justified.

Suppose an Olympics selection trial requires applicants to run 10km in 40 minutes. This selection criterion seems reasonably justified. Running speed tends to be an appropriate measure of athleticism. But the ability to run that fast is probably negatively correlated with age. Someone looking at the data alone might flag a bias against very elderly applicants. Without understanding the context, it is difficult to see how this bias might be reasonably justified.

The fairness metrics can be a systematic way to check for bias, but they are only a piece of the puzzle. A complete assessment for fairness needs us to get down and dirty with the problem at hand.

- How can we quantify fairness?

Most of the fairness metrics focus on equality in the rates of true positives, true negatives, false positives, false negatives, or some combination of these. But remember that these metrics are insufficient when they exclude the larger context of the AIS and neglect contextual justifications.

For more comprehensive reviews of existing metrics, check out Narayanan (2018) and Verma et al. (2018).

The Impossibility Theorem

For our fictional fat pet predictor, we had complete control over the system's accuracy. Even so, you may have noticed that it was impossible to fulfill all five fairness metrics at the same time. This is sometimes known as the Impossibility Theorem of Fairness.

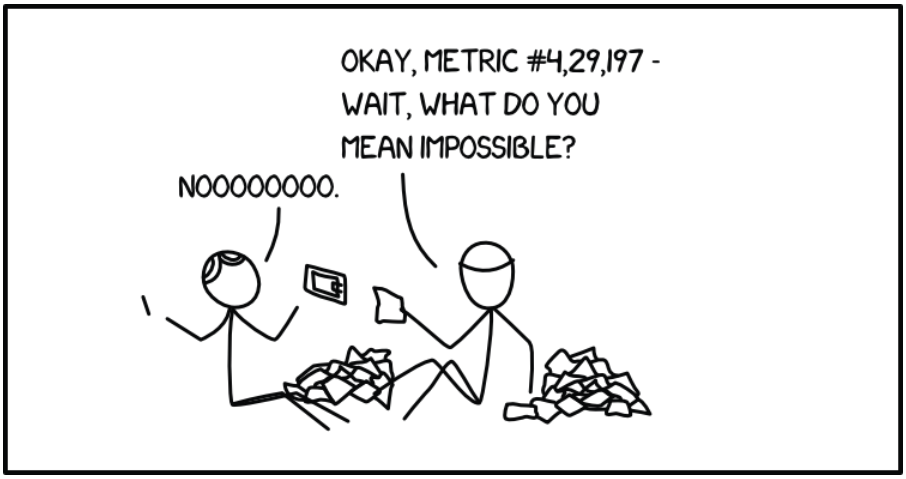
In ProPublica's well-known article Machine Bias, the subtitle reads:

> There's software used across the coutnry to predict future criminals. And it's biased against blacks.

ProPublica's article documented the "significant racial disparities" found in COMPAS, a recidivism prediction model sold by NorthPointe. But in their response, Northpointe disputed ProPublica's claims. Later on, we would discover that NorthPointe and ProPublica had different ideas about what constituted *fairness*. Northpointe used Conditional Use Accuracy Equality, while ProPublica used Treatment Equality (see previous demo for details). Northpointe's response can be found here.

Turns out, it is impossible to satisfy both definitions of fairness, given populations with different base rates of recidivism. This is similar to our previous example of fat pets. Now, different base rates of recidivism do not mean that certain individuals are more prone to re-offending by virtue of race. Instead of racial predisposition, such trends are more likely due to unequal treatment and circumstances from past and present biases. In our fat pets example, dogs might have a higher base rate for obesity not because dogs have fat genes but because dog owners tend to be overly enthusiastic about feeding their pets.

SO FAIRNESS IS IMPOSSIBLE?

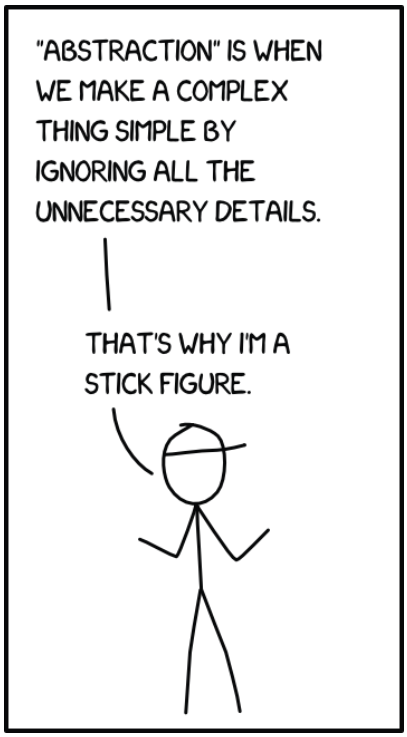


The point of all these is not to show that fairness does not make sense or that it is impossible. After all, notions of fairness are heavily based on context and culture. Different definitions that appear incompatible simply reflect this context-dependent nature.

But this also means that it is super critical to have a deliberate discussion about what constitutes fairness. This deliberate discussion must be nested in the context of how and where the AIS will be used. For each AIS, the AI practitioners, their clients and users of the AIS need to base their conversations on the same definition of fairness. **We cannot assume that everyone has the same idea of fairness.** While it could be ideal for everyone to have a say in what definition of fairness to use, sometimes this can be difficult. At the very least, AI practitioners should be upfront with their users about fairness considerations in the design of the AIS. This includes what fairness definition was used and why, as well as potential shortcomings.

Context-Free Fairness

Computer scientists might often prefer general algorithms that is agnostic to context and application. The agnostic nature of unstructured deep learning is often cited as a huge advantage compared to labor-intensive feature engineering. So the importance of context in understanding fairness can be a bane to computer scientists, who might like to "[abstract] away the social context in which these systems will be deployed" (Selbst et al., 2019).



But as Selbst et al. write in their work on fairness in sociotechnical systems:

> Fairness and justice are properties of social and legal systems like employment and criminal justice, not properties of the technical tools within. **To treat fairness and justice as terms that have meaningful application to technology separate from a social context is therefore to make a category error, or as we posit here, an abstraction error.** [emphasis mine]

On a similar note, in Peter Westen's The Empty Idea of Equality, he writes:

> For [equality] to have meaning, it must incorporate some external values that determine which persons and treatments are alike [...]

In other words, the treatment of fairness, justice and equality cannot be separated from the specific context of the problem at hand.

FIVE FAILURE MODES

In their work, Selbst et al. identify what they term "five failure modes" or "traps" that might ensnare the AI practitioner trying to build a fair AIS. What follows is a summary of the failure modes. We strongly encourage all readers to conduct a close reading of Selbst et al.'s original work. A copy can be found on co-author Sorelle Friedler's website here.

Framing Trap

> Failure to model the entire system over which a social criterion, such as fairness, will be enforced

A fair AIS must take into account the larger sociotechnical context in which the AIS might be used, otherwise it is meaningless. For example, an AIS to filter job applicants should also consider how its suggestions would be used by the hiring manager. The AIS might be "fair" in isolation but subsequent "post-processing" by the hiring manager might distort and undo the "fairness".

Portability Trap

> Failure to understand how repurposing algorithmic solutions designed for one social context may be misleading, inaccurate, or otherwise do harm when applied to a different context

This refers to our earlier observation that computer scientists often prefer general algorithms agnostic to context and application, which Selbst et al. refer to as "portability". The authors contend that the quality of portability must sacrifice aspects of fairness because fairness is unique to time and space, unique to cultures and communities, and not readily transferable.

Formalism Trap

> Failure to account for the full meaning of social concepts such as fairness, which can be procedural, contextual, and contestable, and cannot be resolved through mathematical formalisms

This trap stems from the computer science field's preference for mathematical definitions, such as the many definitions of fairness that we have seen earlier. The authors suggest that such mathematical formulations fail to capture the intrinsically complex and abstract nature of fairness, which is, again, nested deeply in the context of the application.

Ripple Effect Trap

> Failure to understand how the insertion of technology into an existing social system changes the behaviors and embedded values of the pre-existing system

This is related to the Framing Trap in that the AI practitioner fails to properly account for "the entire system", which in this case includes how existing actors might be affected by the AIS. For instance, decision-makers might be biased towards agreeing with the AIS's suggestions (a phenomenon known as automation bias) or the opposite might be true and decision-makers might be prone to disagreeing with the AIS's suggestions. Again, this stems from designing an AIS in isolation without caring enough about the context.

Solutionism Trap

> Failure to recognize the possibility that the best solution to a problem may not involve technology

Hence we crowned the most important question in this entire guide as, "When is AI not the answer?". AI practitioners are naturally biased towards AI-driven solutions, which could be an impedement when the ideal solution might be far from AI-driven.

- How do we design for fairness without context?

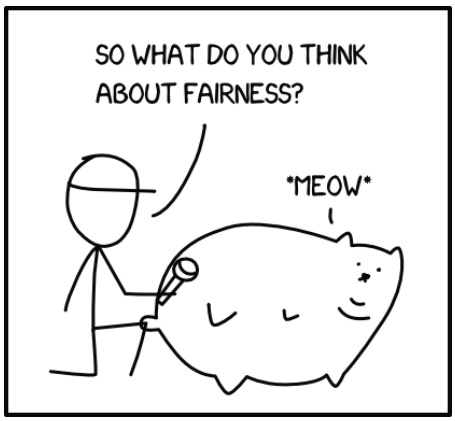
Nope we can't. Gotcha that was a trick question. The same decision can be both fair and unfair depending on the larger context, so context absolutely matters. As such, it is difficult to give advice on how to pick a fairness metric without knowing what is the context. Check out the next section for some questions to help with understanding the context.

Learning about the Context

By the time you read this, "context" should have been burned into your retina. But just in case you cheated and came straight here without reading any of the previous sections:

CONTEXT IS IMPORTANT WHEN DISCUSSING FAIRNESS!

So here is a list of questions and prompts to help you learn more about the sociotechnical context of your application. Don't be limited to these though, go beyond them to understand at much about the problem as you can. Also, these prompts should be discussed as a group rather than answered in isolation. Involve as many people as you can!



General Context

- What is the ultimate aim of the application?
- What are the pros and cons of an AIS versus other solutions?
- How is the AIS supposed to be used?
- What is the current system that the AIS will be replacing?
- Create a few user personas - the technophobe, the newbie etc. - and think about how they might react to the AIS across the short-term and long-term.
- Think of ways that the AIS can be misused by unknowning or malicious actors.

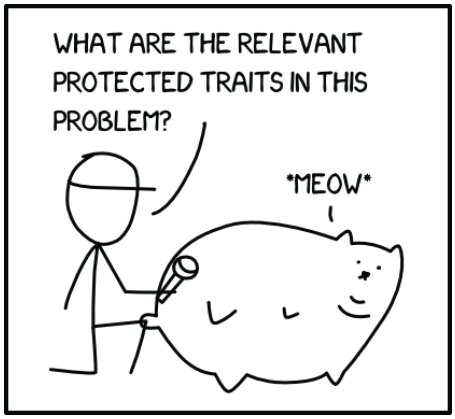
About Fairness

- What do false positives and false negatives mean for different users? Under what circumstances might one be worse than the other?

- Try listing out some examples of fair and unfair predictions. Why are they fair/unfair?
- What are the relevant protected traits in this problem?
- Which fairness metrics should we prioritize?
- When we detect some unfairness with our metrics - is the disparity justified?

Bonus Points!

- Find a bunch of real potential users and ask them all the prompts above.
- Post all of your answers online and iterate it with public feedback
- Ship your answers with the AIS when it is deployed



- How do we learn more about the context?

See above. Most of all, take a genuine interest in your application and its users!

References

1. Fairness definitions explained [link]
   Verma, S. and Rubin, J., 2018. 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), pp. 1-7.

2. Fairness through awareness [PDF]
   Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R., 2012. Proceedings of the 3rd innovations in theoretical computer science conference, pp. 214-226.

3. On the relation between accuracy and fairness in binary classification [PDF]
   Zliobaite, I., 2015. arXiv preprint arXiv:1505.05723.

4. The problem of infra-marginality in outcome tests for discrimination [link]
   Simoiu, C., Corbett-Davies, S., Goel, S. and others, ., 2017. The Annals of Applied Statistics, Vol 11(3), pp. 1193-1216. Institute of Mathematical Statistics.

5. Algorithmic decision making and the cost of fairness [PDF]
   Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. and Huq, A., 2017. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 797-806.

6. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments [link]
   Chouldechova, A., 2017. Big data, Vol 5(2), pp. 153-163. Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA.

7. Equality of opportunity in supervised learning [PDF]
   Hardt, M., Price, E., Srebro, N. and others, ., 2016. Advances in neural information processing systems, pp. 3315-3323.

8. Counterfactual fairness [PDF]
   Kusner, M.J., Loftus, J., Russell, C. and Silva, R., 2017. Advances in Neural Information Processing Systems, pp. 4066-4076.

9. Fairness in criminal justice risk assessments: The state of the art [link]
   Berk, R., Heidari, H., Jabbari, S., Kearns, M. and Roth, A., 2018. Sociological Methods & Research, pp. 0049124118782533. Sage Publications Sage CA: Los Angeles, CA.

10. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment [PDF]
Zafar, M.B., Valera, I., Gomez Rodriguez, M. and Gummadi, K.P., 2017. Proceedings of the 26th International Conference on World Wide Web, pp. 1171-1180.

11. Inherent trade-offs in the fair determination of risk scores [PDF]
Kleinberg, J., Mullainathan, S. and Raghavan, M., 2016. arXiv preprint arXiv:1609.05807.

12. Fairness testing: testing software for discrimination [PDF]
Galhotra, S., Brun, Y. and Meliou, A., 2017. Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, pp. 498-510.

13. Avoiding discrimination through causal reasoning [PDF]
Kilbertus, N., Carulla, M.R., Parascandolo, G., Hardt, M., Janzing, D. and Scholkopf, B., 2017. Advances in Neural Information Processing Systems, pp. 656-666.

14. Fair inference on outcomes [PDF]
Nabi, R. and Shpitser, I., 2018. Thirty-Second AAAI Conference on Artificial Intelligence.

15. Translation tutorial: 21 fairness definitions and their politics [link]
Narayanan, A., 2018. Proc. Conf. Fairness Accountability Transp., New York, USA.

16. Evaluating the predictive validity of the COMPAS risk and needs assessment system [link]
Brennan, T., Dieterich, W. and Ehret, B., 2009. Criminal Justice and Behavior, Vol 36(1), pp. 21-40. Sage Publications Sage CA: Los Angeles, CA.

17. Machine bias [link]
Angwin, J., Larson, J., Mattu, S. and Kirchner, L., 2016. ProPublica, May, Vol 23.

18. Amazon doesn't consider the race of its customers. Should It? [link]
Ingold, D. and Soper, S., 2016. Bloomberg News.

19. Fairness and abstraction in sociotechnical systems [PDF]
Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S. and Vertesi, J., 2019. Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 59-68.

20. The empty idea of equality [link]
Westen, P., 1982. Harvard Law Review, pp. 537-596. JSTOR.

Understanding Bias I

> Accordingly, we use the term bias to refer to computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others.
>
> *Bias in Computer Systems - Friedman & Nissenbaum, 1996*

What is so bad about algorithmic bias anyway? How has it affected the world? To figure out what algorithmic bias is, it can be useful to consider some real-world examples.

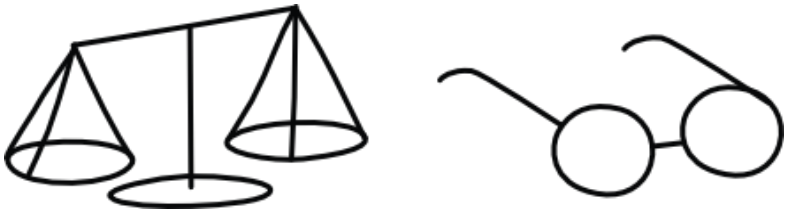
In this chapter, we take a look at what some consequences of algorithmic bias look like.

- How might we analyze the harm caused by algorithmic bias?
- What is an example of allocative harm?
- What is an example of representative harm?

Two Types of Harm

AIS are increasingly used to help **allocate** resources. Credit scoring models that help banks filter loan applications "allocate" loans. Hiring models help companies to "allocate" jobs. Medical diagnosis models help to "allocate" appropriate treatment. The AIS in these examples help identify who to give what. We are affected by these systems because we are denied or given something as a result of an AIS decision.

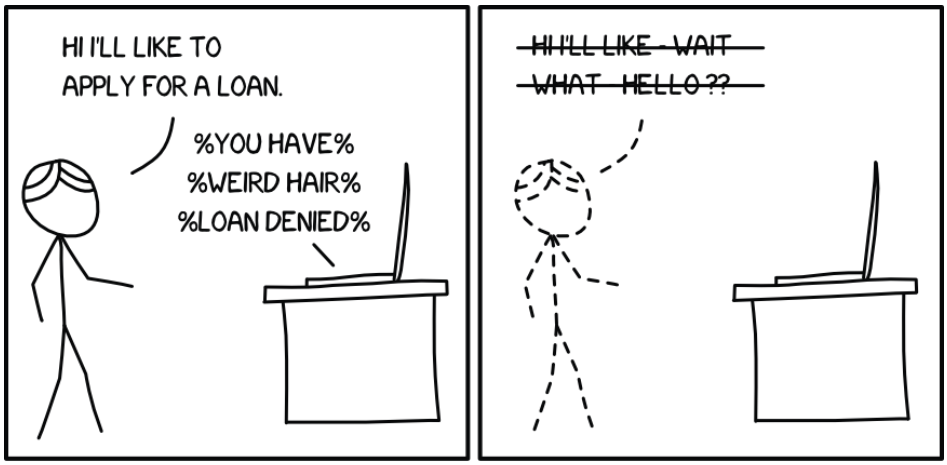
On a more abstract level, AIS are also increasingly affecting the way we perceive or **represent** the world. Think Google Search, Facebook's News Feed and YouTube's Recommended feed. This is also known as "filtering" [2]. The modern person connected to the Internet has access to a vast amount of information but limited time and attention. These AIS prevent us from being overwhelmed and help us focus on the most relevant articles and news. We are affected by these systems because these filters shape our perceptions and thoughts about the world.



We can classify the consequences of algorithmic bias in the same way. This was proposed by Kate Crawford in her NIPS 2017 keynote The Trouble with Bias [1]. Crawford first defined algorithmic bias as "a skew that produces a type of harm". She then further classifies algorithmic biases into harms of **allocation** and harms of **representation**. Over the next two sections, we will use the same framework to look at real-world examples of algorithmic bias. Since context has often been emphasized in the previous sections, we will try to see how context can be explored in these examples.

Harms of Allocation	Harms of Representation
Immediate	Long term
Easily quantifiable	Difficult to formalize
Discrete	Diffuse
Transactional	Cultural

Comparison between the two types of harm, from Crawford's NIPS 2017 keynote [1].



- How might we analyze the harm caused by algorithmic bias?

A framework proposed by Kate Crawford classifies algorithmic bias by the type of harm caused. Harms of allocation refers to unfairly assigned opportunities or resources due to algorithmic intervention. Harms of representation refers to algorithmically filtered depictions that are discriminatory.

Harms of Allocation

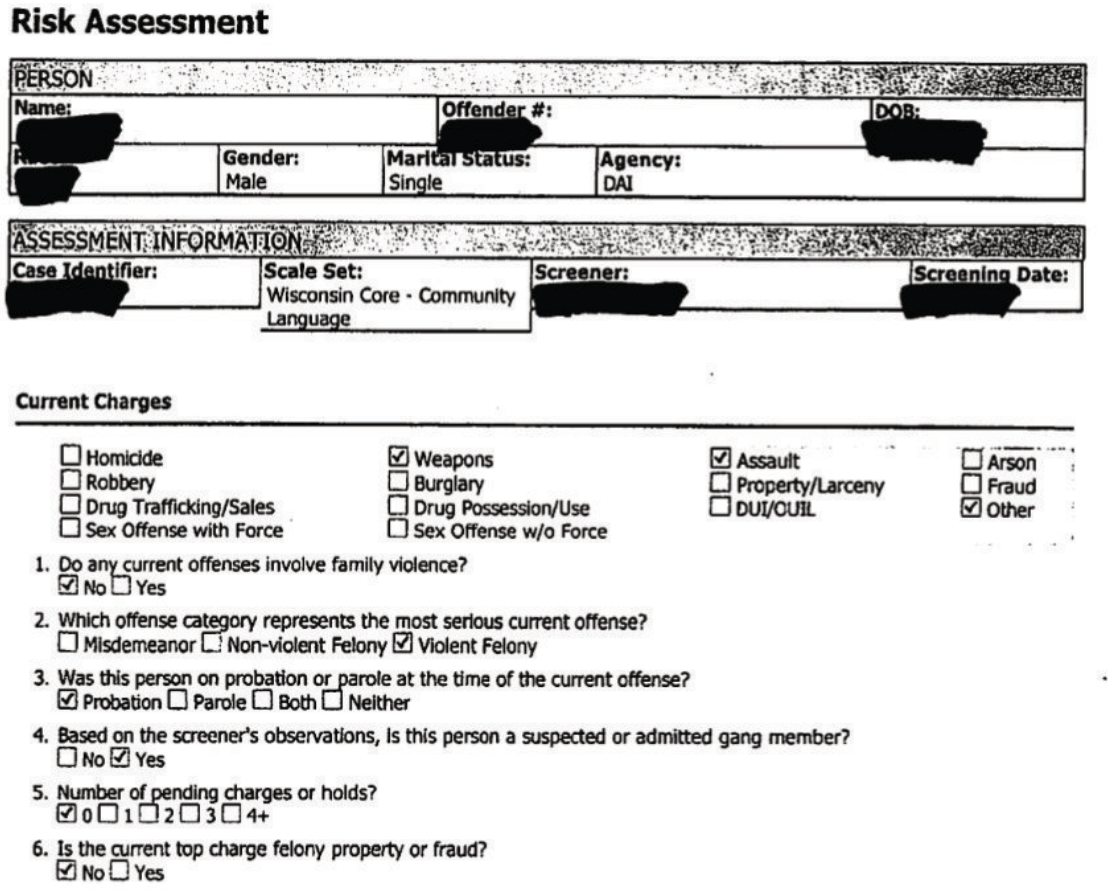
> An allocative harm is when a system allocates or withholds certain groups an opportunity or a resource.
>
> *The Trouble with Bias, Kate Crawford at NIPS2017 [1]*

> Automated eligibility systems, ranking algorithms, and predictive risk models control which neighborhoods get policed, which families attain needed resources, who is short-listed for employment, and who is investigated for fraud.
>
> *Automating Inequality - Virginia Eubanks, 2018 [15]*

Harms of allocation arise from the unjust distribution of opportunities and resources, such as jobs, loans, insurance and education. An allocative harm can range from a small but significant and systematic difference in treatment, all the way to complete denial of a particular service.

COMPAS

The Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS, algorithm is probably the most infamous case study in algorithmic bias. In areas where COMPAS was used, defendants typically answer a COMPAS questionnaire when they are first booked in jail.



Risk Assessment

PERSON

Name:	Offender #:	DOB:
[redacted]	[redacted]	[redacted]

[redacted]	Gender: Male	Marital Status: Single	Agency: DAI

ASSESSMENT INFORMATION

Case Identifier:	Scale Set:	Screener:	Screening Date:
[redacted]	Wisconsin Core - Community Language	[redacted]	[redacted]

Current Charges

☐ Homicide	☑ Weapons	☑ Assault	☐ Arson
☐ Robbery	☐ Burglary	☐ Property/Larceny	☐ Fraud
☐ Drug Trafficking/Sales	☐ Drug Possession/Use	☐ DUI/OUIL	☑ Other
☐ Sex Offense with Force	☐ Sex Offense w/o Force		

1. Do any current offenses involve family violence?
   ☑ No ☐ Yes
2. Which offense category represents the most serious current offense?
   ☐ Misdemeanor ☐ Non-violent Felony ☑ Violent Felony
3. Was this person on probation or parole at the time of the current offense?
   ☑ Probation ☐ Parole ☐ Both ☐ Neither
4. Based on the screener's observations, is this person a suspected or admitted gang member?
   ☐ No ☑ Yes
5. Number of pending charges or holds?
   ☑ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4+
6. Is the current top charge felony property or fraud?
   ☑ No ☐ Yes

Part of a COMPAS questionnaire.

Using the responses, the COMPAS model outputs several scores related to recidivism. These include scores for Risk of Recidivism and Risk of Violent Recidivism, which go from 1 to 10, with 10 being highest risk. The scores were given to judges and they often had a huge influence on the sentence:

> After Brennan's testimony, Judge Babler reduced Zilly's sentence, from two years in prison to 18 months. "Had I not had the COMPAS, I believe it would likely be that I would have given one year, six months," the judge said at an appeals hearing on Nov. 14, 2013.
>
> *Machine Bias - Julia Angwin et al., 2016 [16]*

If we think of COMPAS as a model for potentially "allocating" freedom, harms of allocation can become very severe. In ProPublica's exposé on COMPAS, the journalists argued that the algorithm was "biased against blacks".

> In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.
>
> - The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.
> - White defendants were mislabeled as low risk more often than black defendants.

In short, black defendants were more likely to be wrongly accused of reoffending, while white defendants were more likely to "escape detection". We cited this example in an earlier section (The Impossibility Theorem), where we also mentioned that Propublica and Northpointe employed different definitions of fairness. Putting aside the debate of which definition of fairness to apply, there are also other considerations.

Proxy Labels

The term "recidivism" refers to the likelihood of a criminal committing another crime, after they have been convicted. To train a recidivism prediction model, the training data should ideally have labels denoting whether a convicted criminal has reoffended. But in reality, we don't know when someone has committed a crime, only when someone has been arrested. So, we use a proxy. Instead of labels denoting whether a convicted criminal has reoffended, the labels denote whether a convicted criminal has been convicted again. That might be the closest we can get, but is it close enough?

Let's think about some of the differences between "reoffending" and "being convicted again".

1. A criminal who has reoffended might not necessarily be caught. This means that we are missing out on the smart and lucky criminals who escape conviction.

2. The system is imperfect. Unfortunately, innocent people sometimes get wrongly accused and wrongly convicted. This means that we could have people labeled "convicted again", who have not actually "reoffended".

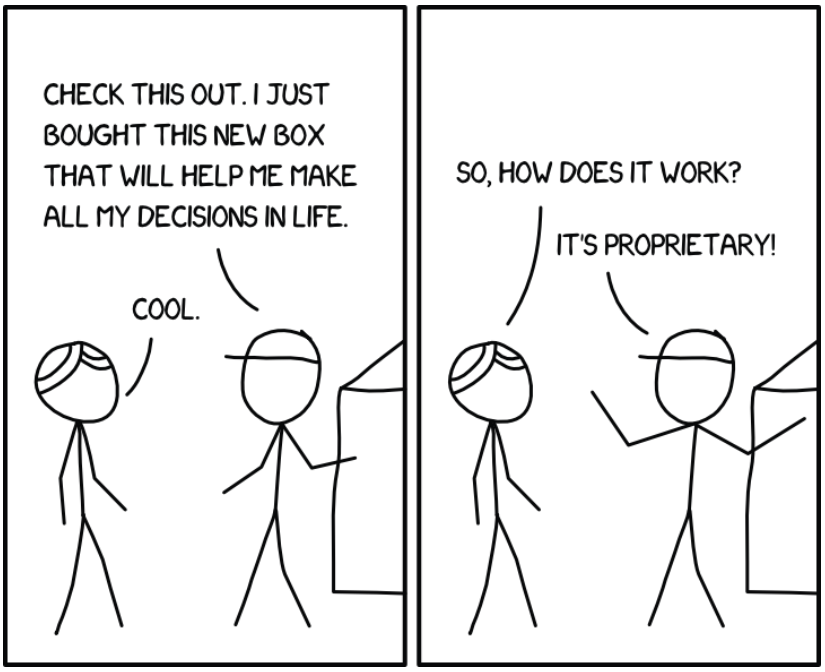
Okay, now let's go one step further and think about how a trait like race might affect these two differences. Racism in the police has been well-documented in literature [17,18,19]. In recent years, institutional racism and the related problem of police brutality have also inspired social movements such as "Black Lives Matter". In light of these issues, how might the above differences play out?

1. If racism has a major influence on police practices like stop-and-frisk, we might find that white re-offenders have a higher chance of not getting caught, as compared to black re-offenders. This might cause our dataset to underestimate the number of white re-offenders.

2. And likewise, we might find that black individuals are subject to wrongful arrests more frequently than white individuals. In that case, our dataset might be overestimating the number of black repeat offenders.

In other words, by using the proxy label of "being convicted again" rather than "reoffending", we could be exaggerating the recidivism rate of black individuals and systematically biasing the dataset along racial lines. Obviously all of this is hypothetical and requires more substantial evidence. Nevertheless, when faced with problems like these, it might be prudent to ask if an algorithmic solution is really the answer.

Public Disclosure



Despite the important role that risk scores like COMPAS play in the criminal justice system, there is little public information about these systems.

> [Researchers Sarah Desmarais and Jay Singh's] analysis of [19 risk methodologies] through 2012 found that the tools "were moderate at best in terms of predictive validity," Desmarais said in an interview. And she could not find any substantial set of studies conducted in the United States that examined whether risk scores were racially biased. "The data do not exist," she said.
>
> *Machine Bias - Julia Angwin et al., 2016 [16]*

Important information that probably should be available include:

- What goes into the risk score?
- How is it calculated?
- What is the accuracy?
- How is this accuracy measured?
- What definition of fairness was used to develop the scores?
- Why this definition instead of other definitions?
- What are potential fairness violations?

Not having to disclose such information allows bias to remain undetected. Because this information is missing, alternative actors such as ProPublica take up the mantle to evaluate these systems. But this often happens only after the AIS have been in use for some time and harm has been done.

Then again, a potential problem is that public disclosure might undermine the validity of the scores. Understanding how the risk scores are calculated might enable malicious individuals to game the scores. Nevertheless, considering what is at stake, we have to put some thought into how appropriate disclosure can be made about these scores.

The Greater Good

> "'Greater good?' I am your wife! I'm the greatest good you're ever gonna get!"
>
> *Honey Best, Frozone's wife in The Incredibles*

For the criminal justice system, we can think of its overarching aim as the greater good of promoting societal safety. The sentencing process can be seen as one of its major tools:

> Four major goals are usually attributed to the sentencing process: retribution, rehabilitation, deterrence, and incapacitation.
>
> *Sentencing and Corrections in the 21st Century: Setting the Stage for the Future - Doris Layton Mackenzie, 2001*

When we use a tool like COMPAS to decide the length of a prison sentence, we seem to focus on *retribution* and *incapacitation*, and neglecting *rehabilitation*. Is that really serving the greater good of societal safety? By reducing the issue of societal safety to recidivism prediction, we get a quantifiable problem that might be simpler to solve. But this neglects the greater objective and other alternative problems and solutions. We can see this as an instance of Selbst et al.'s Framing Trap, which we covered previously.

When we consider the greater objective of societal safety, alternative solutions might come to mind. Rather than using COMPAS for determining jailtimes, it can help design specific intervention and rehabilitation measures customized for each defendent. In fact, this might have been what COMPAS was *designed* for, which brings us to our next section.

Human-Algorithm Interaction

In the earlier quote from the article, we see how Judge James Babler from Barron County, Wisonsin, had been influenced by COMPAS to give a more severe sentence than he would have otherwise given. The more severe sentence was only retracted after Tim Brennan, Northpointe's founder, had "testified that he didn't design his software to be used in sentencing". This is reflected in Chapter 4 of Northpointe's Practitioner's Guide to COMPAS Core, which lists different interventions for specific aspects. Throughout the chapter, there are repeated references to non-incarceration interventions. For example, under the Financial Problems section, we see the following recommendation:

> Education on money management and fulfilling court ordered financial commitments is part of the necessary approach when considering interventions. Assuming someone knows how to manage their finances is an erroneous starting place, vocational training may also play a role in creating a successful change plan.
>
> *Practitioner's Guide to COMPAS Core - Northpointe, 2015*

So what went wrong?

Maybe Brennan had been too idealistic when thinking about how judges might be using COMPAS scores. Maybe Brennan didn't think that the scores could be interpreted as a measure for how long someone should be jailed. Whatever it is, the ones who deployed COMPAS had not appropriately considered how it might be used and how it might influence others. Recall Selbst et al.'s Ripple Effect Trap mentioned earlier. Here we neglected the "ripple effects" that COMPAS had on judges and underestimated COMPAS's potential for allocative harm. When we take these into consideration, we might have changed aspects of the system. For example, instead of risk scores, COMPAS could explicitly output the recommended intervention. That could reduce the chance of misunderstanding or misusing the risk scores.

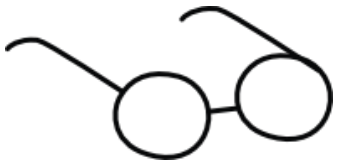
- What is an example of allocative harm?

COMPAS is a classic case of allocative harm in algorithmic bias literature, concerning the "allocation" of freedom. By examining the larger sociotechnical context of the criminal justice system that COMPAS is employed in, we identified many potential problems relating to algorithmic bias, such as:

- **Differences between proxy labels and actual labels**
- **Public disclosure of fairness considerations**
- **Neglecting the larger objective**
- **Failing to comprehensively consider how the AIS affects the system**

For more examples of allocative harms, check out Cathy O'Neil's Weapons of Math Destruction [9] and Viriginia Eubanks's Automating Inequality [15].

Harms of Representation



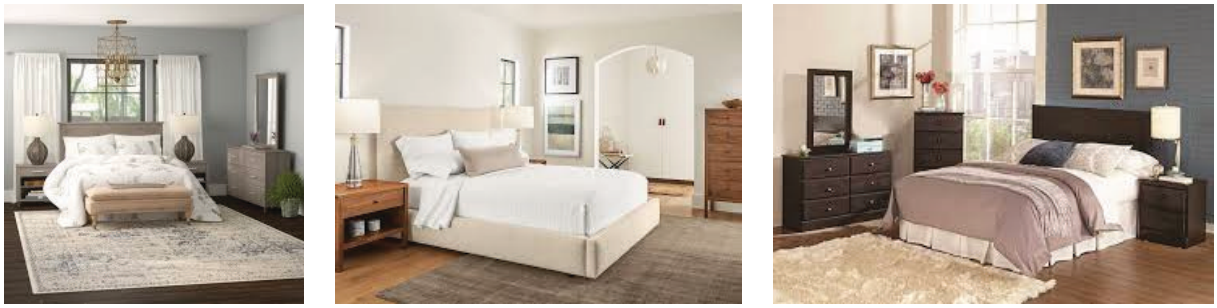
> [Representative harms] occur when systems reinforce the subordination of some groups along the lines of identity.
>
> *The Trouble with Bias, Kate Crawford at NIPS2017 [1]*

> If you control the flow of information in a society, you can influence its shared sense of right and wrong, fair and unfair, clean and unclean, seemly and unseemly, real and fake, true and false, known and unknown.
>
> *Future Politics - Jamie Susskind, 2018 [2]*

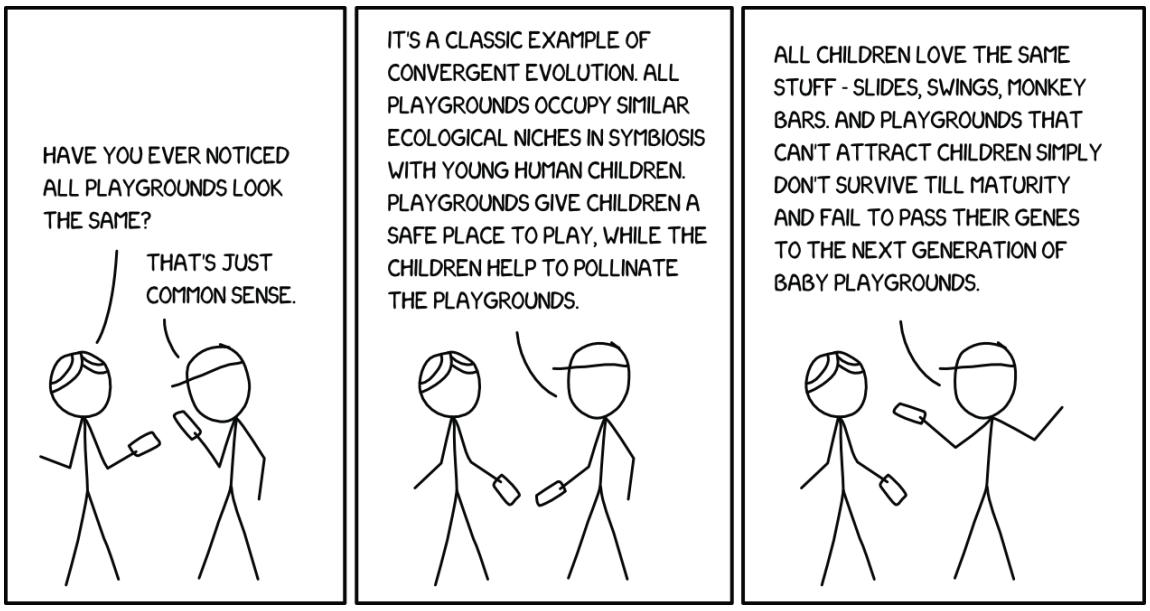
GOOGLE IMAGE SEARCH

Most of us have had experience with Google Image search. Maybe it was to find some stock photos or wallpapers. Or maybe it was to look up what some exotic animal looked like. One thing we might have noticed is that the search results often return stereotypical images of our query. Searching “playground” would give us photos of the classic outdoor playground with small slides and steps. Searching “bedroom” would return photos of nicely made beds and tidy rooms that would seem perfectly natural in a furniture catalogue.

Google Image Search for “playground”.



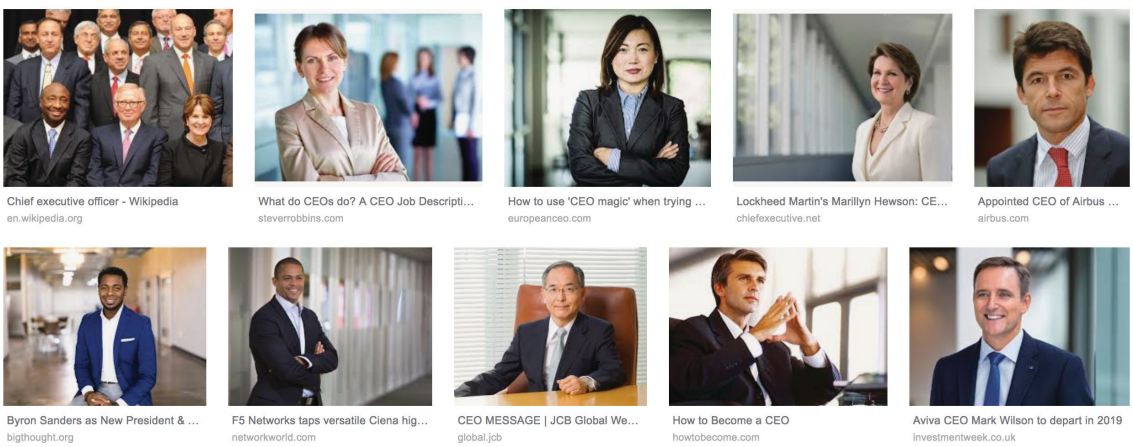
Google Image Search for “bedroom”.



Such stereotypes go beyond objects and places, extending to queries of people as well. Studies have found that Google's Image Search perpetuated and exaggerated gender and racial stereotypes for certain keywords, such as "CEO", "doctor" and "nurse" [3,4]. We know that these words are gender-neutral. But most of us might also know that these words tend to embody certain stereotypes, such as the male doctor and the female nurse. Let's consider the simple and vivid example of Google's Image Search for the term "CEO".

In April 2015, a Google Image search for the term "CEO" surfaced results that were manifestations of both racial and gender biases. An overwhelming majority of the images were photos of white males in suits. Since these biases have been flagged by several researchers, they appear to have been mitigated somewhat and a recent search shows a far more diverse result (see below).

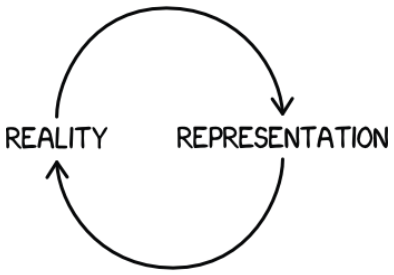
Results from Google Image Search for "CEO" in April 2015 (retrieved from here [6]) were dominated by photos of white males.



Results from Google Image Search for "CEO" in July 2019 show a more diverse distribution, in terms of race and gender.

Harms of representation are dangerous because they shape how we see the world. And in turn, how we see the world shapes the world. A generation raised solely on fairy tales of damsels in distress might not recognize the existence of heroines and men in need of saving. A generation raised solely on image search results of white male CEOs may find it difficult to entertain the possibility of a non-male non-white CEO. By limiting our cognitive vocabulary, these harmful representations become additional psychological obstacles that must be overcome.

Furthermore, when these harmful representations manifest themselves as biased actions and decisions they become self-fulfilling prophecies. Fed on a diet of white male CEO images, non-male non-white individuals might never fight for the position and we may never encourage them to go for it. We might even discourage them from pursuing what seems like an unrealistic ambition. Over time, there are fewer and fewer non-white non-male CEOs and the biases embodied by the search results turn out to be an accurate prophecy.



In that case, what does an unharmful representation look like? Two possible alternatives to consider are accurate representations and ideal representations.

Accurate Representations

Yes, the Google Image results in April 2015 were dominated by white males. But technically, in 2014, only 4% of the 500 companies on the US S&P 1500 had female CEOs [8]. This means that if the search results replicated this 4% proportion of females, we might consider this as an **accurate** representation.

On the other hand, search results that have zero female images would be obviously inaccurate. Such results would be perpetuating false and exaggerated gender stereotypes.

Ideal Representations

In March 2015, the New York Times ran an article titled “Fewer Women Run Big Companies Than Men Named John” [7]. This contributed to a growing literature on gender inequality. Such literature describes an ideal world where the gender distribution of CEOs is equal, or at least similar to the gender distribution of the general population. Search results that reproduce this equality would be an **ideal** representation.

Representations both embed and influence unwritten norms and values. Following the cycle between representation and reality, we can make the world a better place by first seeing it as a better place. In our example, the presence of more gender- and race-diverse search results for “CEO” can encourage non-white non-male candidates to go from minority to mainstream.

Accuracy versus Idealism

There is merit behind both an accurate representation and an ideal representation. But in an imperfect world, representations cannot be both accurate and ideal. Decisions and compromises have to be made about which is more important for the given application.

Imagine if a company’s internal personnel directory tries to give an ideal and fair representation of a query for the company’s regional managers. That would probably defeat the purpose of the directory. On the other hand, people often use Google to learn more about the world. Maybe presenting a more equal representation could eventually make the real world a more equal place. As always, making the right choice requires knowledge about the context.

- What is an example of representative harm?

Biased results in Google Image Search can be seen as an instance of representative harm. The harm caused is more subtle and indirect but no less dangerous than harms of allocation. A biased representation can influence people's behaviors and in turn, change the world for the worse.

Fixing harms of representation requires a conversation about the tradeoffs between an accurate representation and an ideal one. Once again, detecting such harms and fixing them requires thinking beyond the scope of mathematical algorithms and venturing into social implications.

References

1. The trouble with bias [link]
   Crawford, K., 2017. Conference on Neural Information Processing Systems, invited speaker.

2. Future politics: Living together in a world transformed by tech
   Susskind, J., 2018. Oxford University Press.

3. Competent men and warm women: Gender stereotypes and backlash in image search results [link]
   Otterbacher, J., Bates, J. and Clough, P., 2017. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 6620-6631.

4. Unequal representation and gender stereotypes in image search results for occupations [link]
   Kay, M., Matuszek, C. and Munson, S.A., 2015. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 3819-3828.

5. Who's a CEO? Google image results can shift gender biases [link]
   Langston, J., 2015. University of Washington.

6. Google Image Search Has A Gender Bias Problem [link]
   Cohn, E., 2015. Huffpost.

7. Fewer Women Run Big Companies Than Men Named John [link]
   Wolfers, J., 2015. The New York Times.

8. Women on US boards: what are we seeing? [PDF]
   Young, E.&., 2015. Ernst & Young.

9. Weapons of math destruction: How big data increases inequality and threatens democracy
   O'Neil, C., 2016. Broadway Books.

10. Semantics derived automatically from language corpora contain human-like biases [PDF]
    Caliskan, A., Bryson, J.J. and Narayanan, A., 2017. Science, Vol 356(6334), pp. 183-186. American Association for the Advancement of Science.

11. Men also like shopping: Reducing gender bias amplification using corpus-level constraints [PDF]
Zhao, J., Wang, T., Yatskar, M., Ordonez, V. and Chang, K., 2017. arXiv preprint arXiv:1707.09457.

12. Gender bias in coreference resolution: Evaluation and debiasing methods [PDF]
Zhao, J., Wang, T., Yatskar, M., Ordonez, V. and Chang, K., 2018. arXiv preprint arXiv:1804.06876.

13. Word embeddings quantify 100 years of gender and ethnic stereotypes [link]
Garg, N., Schiebinger, L., Jurafsky, D. and Zou, J., 2018. Proceedings of the National Academy of Sciences, Vol 115(16), pp. E3635-E3644. National Acad Sciences.

14. Women also snowboard: Overcoming bias in captioning models [PDF]
Hendricks, L.A., Burns, K., Saenko, K., Darrell, T. and Rohrbach, A., 2018. European Conference on Computer Vision, pp. 793-811.

15. Automating inequality: How high-tech tools profile, police, and punish the poor
Eubanks, V., 2018. St. Martin's Press.

16. Machine bias [link]
Angwin, J., Larson, J., Mattu, S. and Kirchner, L., 2016. ProPublica, May.

17. Black and blue: An analysis of the influence of race on being stopped by the police [link]
Norris, C., Fielding, N., Kemp, C. and Fielding, J., 1992. British Journal of Sociology, pp. 207-224. JSTOR.

18. In Proportion: Race, and Police Stop and Search [link]
Waddington, P.A., Stenson, K. and Don, D., 2004. British journal of criminology, Vol 44(6), pp. 889-914. Oxford University Press.

19. Driving while black: Bias processes and racial disparity in police stops [link]
Warren, P., Tomaskovic-Devey, D., Smith, W., Zingraff, M. and Mason, M., 2006. Criminology, Vol 44(3), pp. 709-738. Wiley Online Library.

Understanding Bias II

Are there some specific things to look out for when developing a fair AIS?

This chapter looks at some possible sources of algorithmic bias across different stages of developing an AIS.

- What are important considerations for data?
- What are important considerations for algorithm design?
- What are important considerations for deployment?

Bias from Data

In this section, we look at sources of bias across the data preparation process. We assume that we already have a well-defined problem and a rough idea of how the AIS will be deployed. We should also have an initial list of protected traits to evaluate sources of bias.

DEFINING THE POPULATION

The "population" refers our AIS's target audience or all the possible inputs to our AIS. By "defining the population", we are referring to understanding how the population is distributed amongst different features. For example, our earlier fat pet predictor is targeted at all cats and dogs. So the population would be all present and future cats and dogs. A simple baseline is to document the distributions of protected traits in the population.

> Facial recognition is currently being used for passenger boarding for certain airlines and airports [1]. Privacy concerns aside, suppose we define the population as airline passengers aged 18 to 50. We might be happy when the system works for this defined population. But if our passengers actually include very juvenile or very elderly passengers, the AIS might fail for these groups. Specifically, it would help to document the passenger population along the protected traits of race, gender, age and face-related anomalies, and evaluate our AIS against the actual population.

Defining the population is critical because this has downstream effects on how we collect data and design the model. An AIS based on an ill-defined population is likely to fail for the actual target audience.

Historical bias [2] can make it difficult to accurately define the population. Effort must be put into understanding the sociotechnical context of the problem. In the airline boarding example above, it is intuitive to use past passenger records to characterize our population. But imagine if the records only document the purchaser's information. We might then miss out on very young passengers who are unlikely to be buying their own tickets.

Bias can crop up when the *defined* population is not the *actual* population. Be wary of unintended historical bias when defining the population.

THE TARGET VARIABLE

How do we label our dataset? Sometimes this is simple and our objective translates to a clear label. A dataset for a fat pet predictor just has to label overweight pets. A dataset for a spam filter just has to label… well… spam. But sometimes, the objective is more abstract or difficult to formalize [3].

Substitutes

When the target variable cannot be easily or accurately measured, we might employ substitutes.

> A common example is collecting data for a recidivism prediction algorithm. Ideally, the label should be whether an individual has committed a crime again. But lacking omniscience, we have to make do with whether an individual has been arrested again. This is obviously an imperfect substitute. A crafty criminal might be able to escape a second arrest. An unlucky individual might be wrongly arrested and convicted. See the previous section for more details.

In such cases, we have to be acutely aware that we are using an **imperfect substitute**. This should also be communicated to users of the AIS.

Subjective Objectives

In some cases, the target variable is actually a subjective judgement. This increases the chances of bias creeping into the dataset via subjective labels.

For example, an AIS for filtering job applicants will require a dataset with labels of "good" and "bad" applicants. This can be very subjective, differing from employer to employer. Past employment history could have embedded biases along gender, race, age and other attributes.

Where possible, subjective labels should be replaced with clearly defined and well-justified criteria. Otherwise, datasets with subjective labels should be closely inspected for biases along the protected traits identified earlier.

- What are important considerations for data?

Defining the Population. **This refers to how we define the target scope of inputs to the AIS.**

Training Dataset versus Population. **This looks at what are the differences between the training data and the defined population.**

The Target Variable. **This relates to the purpose of the AIS and looks at the differences between the labels used and the actual labels that we are targeting.**

Bias from Algorithm Design

Here, we examine how bias might creep into the algorithm design.

INPUT FEATURES

Should we use all the features that are available in our dataset? How do we know which ones are okay to use?

Protected Traits

In order to prevent disparate treatment, we might want to remove protected traits from being used in our model. But in some cases, the use of protected traits is justified. For instance, certain medical conditions, such as lactose intolerance, are more common in some ethnicities and nationalities compared to others. The presence of these traits would be extremely useful for the diagnosis of these conditions.

The point of identifying protected traits is not to blindly remove them from the model. Rather, knowing about these traits helps us to understand more about the social context and think through the justifications for using them.

Proxies

Even when we explicitly exclude protected traits from the model, proxies might be a hidden cause of bias. These proxies are correlated to the protected traits, which allows the model to use them as substitutes even when the protected traits are removed. For example, income level is often correlated with race, gender and age. Hence income level might act as proxies for these traits. We can check for correlations between all the features and our protected traits to identify proxies.

Just like for protected traits, even if some features act as proxies, we do not necessarily want to remove them. But being aware of these correlations can help us diagnose biases that we may discover later.

AGGREGATION

One of the sources of bias raised by Suresh and Guttag [2] was Aggregation Bias:

> Aggregation bias arises when a one-size-fits-all model is used for groups with different conditional distributions, p(Y|X). Underlying aggregation bias is an assumption that the mapping from inputs to labels is consistent across groups. In reality, this is often not the case.

One Model

As raised by Suresh and Guttag, adopting a one-size-fits-all model assumes that "the mapping from inputs to labels is consistent across groups". When that assumption is false, adopting this model can lead to poor performance for everyone, since the model is struggling to compromise across diverse groups. Alternatively, the model might only be

optimized for the dominant group in the dataset and sacrifice performance for the minority groups.

Multiple Models

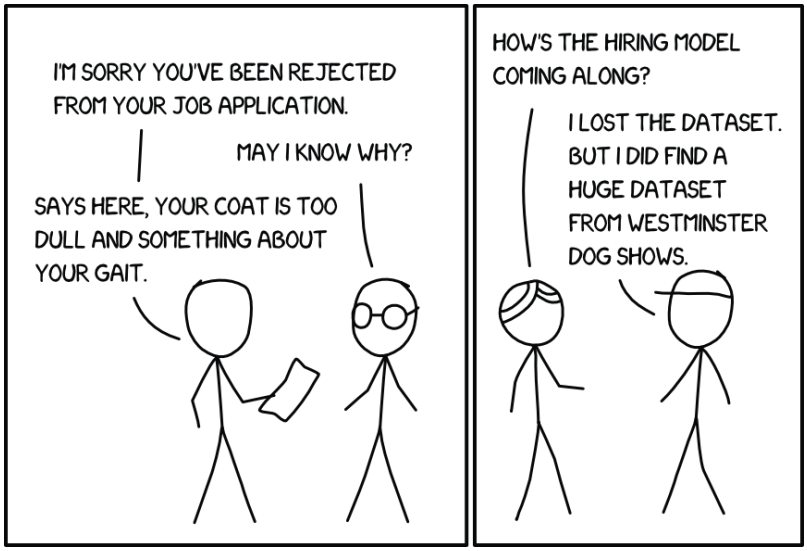
Adopting multiple models to cater to different groups also come with certain conditions. This typically works well only if there is sufficient data, which is often true for dominant groups but less so for minorities. The disparity in amount of data can then lead to a disparity in model accuracies. A possible tweak might be to pretrain a model using a general dataset, before tuning the model for each group.

Unfortunately, in some contexts, using different models for different groups can be contentious and seen as a form of discrimination. In the context of the US labor law, this practice is known as subgroup norming and is illegal under the Civil Rights Act of 1991 [7].

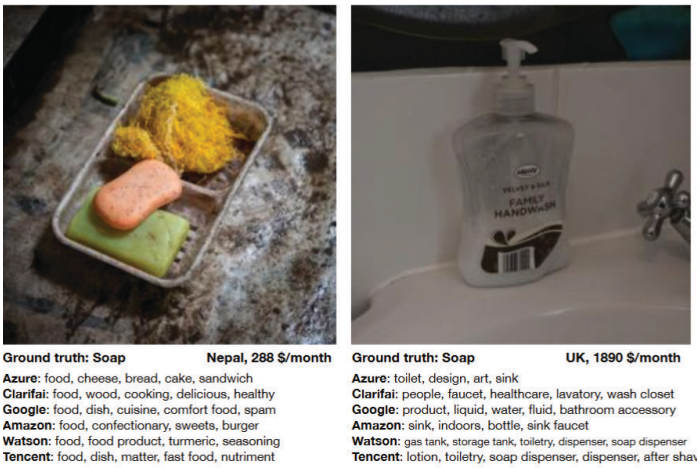
TRANSFERING MODELS AND DATASETS

Since larger datasets often mean better performance, a common trick is to import datasets and pre-trained models from other contexts. For example, the Keras library contains pre-trained image models and spaCy has pre-trained "neural models for tagging, parsing and entity recognition".

However, inappropriately transfering these datasets and pre-trained models might cause issues when the previous contexts are different from the new contexts. This is part of the Portability Trap from Selbst et al.'s five failure modes that we covered earlier.



For example, many pre-trained image models, including the ones from Keras, are trained on ImageNet. While ImageNet is definitely diverse and massive, we should be aware that the images could be an American-centric or Western-centric way of looking at the world. For one, all the labels are in English. Some of the categories such as 'recreational vehicle, RV, R.V.' and 'maypole' could be unfamiliar to other non-Western cultures. Some categories might also mean different things in different cultures and contexts. These problems were highlighted by DeVries et al. when they tested image recognition models against Gapminder's Dollar Street images, which comprises images from 60 different countries [5]. These problems also motivated Google's Inclusive Images Challenge [4]. When we transfer datasets and pre-trained models, we are often using substitutes and proxies, which can be insufficient or completely inappropriate.



Comparing images of soap from different cultures in the Dollar Street dataset, from DeVries et al. [5].

This is not to say that we should never import any datasets and pre-trained models. We just have to be more conscious about what are the differences between the contexts of these resources, versus the context that we are actually designing for.

- What are important considerations for algorithm design?

Input Features. **This concerns the use of protected traits and their proxies, as input features to the AIS.**

Aggregation. **This examines the way we aggregate the dataset and whether to use a single model or multiple models for different input groups.**

Transfering Models and Datasets. **This concerns the disparities due to using datasets and pre-trained models from different contexts.**

Bias from Deployment

Okay, now that we have trained our model, how do we evaluate it? What are important considerations when deploying the chosen model?

EVALUATION

Let's go over some fundamentals first. One of the obvious things to do is to evaluate the model on the test set. And this test set needs to be separate from the training set and the validation set. Just as how we analyzed the training set earlier, we need to think about how the test set might differ from our population.

Beyond analyzing the accuracy and other performance-related metrics, we should employ some of the fairness metrics that we have reviewed in the previous section. These metrics can be used to check for disparities along the protected traits we have identified. Since some of these fairness metrics might be mutually exclusive, there needs to be a careful conversation about which metrics to prioritize. This process should ideally be documented for public disclosure. Results from prioritized and non-prioritized fairness metrics should also be disclosed to inform users about possible problems. If it helps, releasing this information is not just an altruistic gesture. Greater transparency can make for more loyal and supportive users and reduce the chance of backlash from unofficial exposés.

Understanding more about the performance of the AIS helps users make an informed decision about how (and whether) to use the AIS, which brings us to our next section.

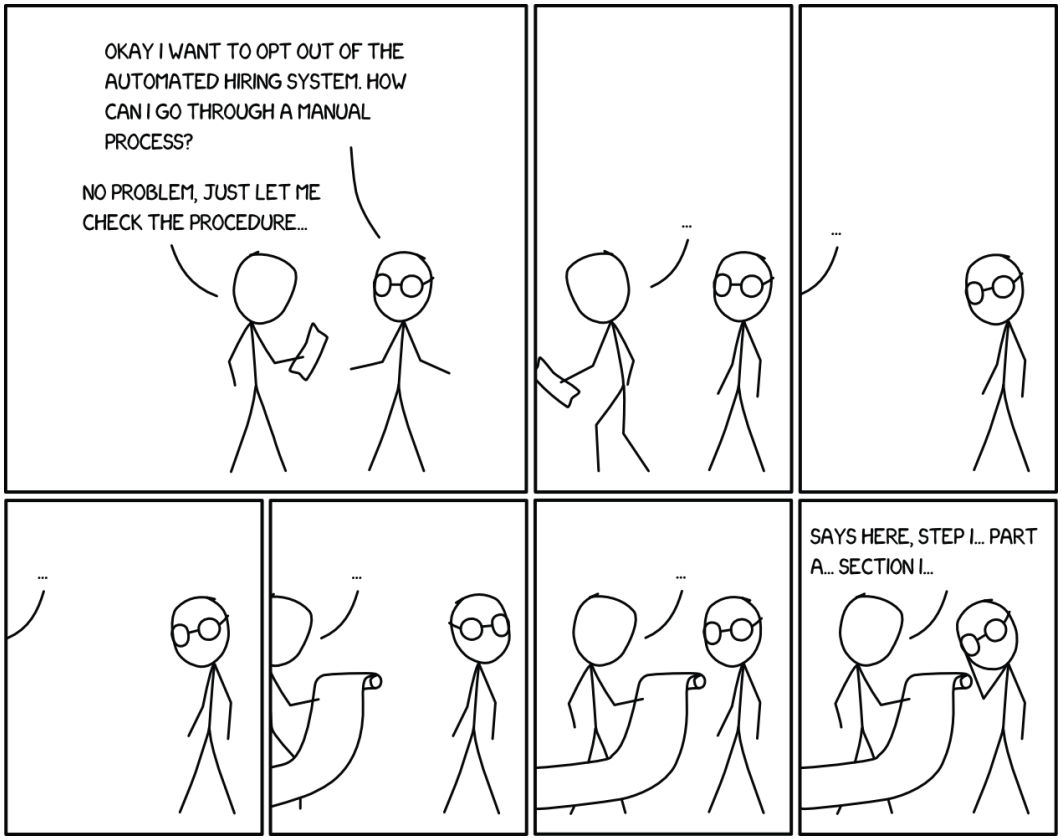
GRACEFUL DEGRADATION

We can think of graceful degradation in two ways.

The first involves system failures - how can the AIS fail gracefully? This involves building in back-ups and alternatives to the AIS, such as the human operatives who step in when Google Duplex is unable to handle a certain conversation. Most engineers are probably familiar with this concept.

The second type of graceful degradation looks at how users of the AIS can opt out of the AIS and still receive adequate service or treatment. This looks at the important issue of an AIS itself being a form of bias against individuals reluctant or unable to use such

systems. In a Wired article, Allie Funk documented the tremendous difficulties she faced when trying to board a Delta Air Lines flight without using facial recognition [1].



Current smartphones that employ facial recognition also allow users to use passcode access. This reduces the harm caused to individuals who cannot or do not want to use facial recognition. In the same way, instead of saying that users can "choose" to opt out and then leave them with no reasonable alternative, AIS should allow users to opt out gracefully. Consider how users can choose to use only part of the AIS or make it easy to adopt other viable alternatives.

FEEDBACK

Allowing users to opt out empowers them rather than subject them to the whims of the AIS. Another important mode of empowerment is allowing users to provide feedback about the AIS and for this feedback to manifest as tangible improvements. In terms of fairness, user feedback can help to surface instances of bias. Without real user feedback, any concept of fairness is ultimately subjected to the limited experiences of the engineers and designers.

On a more algorithmic note, lack of proper feedback can lead to scenarios where deployed models reinforce self-fulfilling prophecies.

Consider the case of hot-spot predictors for policing, which was mentioned by Cathy O'Neil in Chapter 5 of Weapons of Math Destruction. These models, such as PredPol, CompStat and HunchLab, predict crime hot-spots, which are then allocated more attention by the police via patrols. This sounds great since the police can utilize its limited resources more effectively.

But let's consider what happens if a prediction model gets it wrong initially. Suppose we have two areas, Area A and Area B, with equal rates of crime. Suppose the model says that Area A is a hot-spot and neglects Area B. Area A gets more patrols and because there are more patrols, more crime is detected and more arrests are made. These arrests are logged into a dataset, which is fed back into the model. The model sees that Area A has more arrests than Area B and continues predicting it as a hot-spot. We never get the chance to find out that both areas actually have the same crime rate!

In the words of O'Neil:

> This creates a pernicious feedback loop. The policing itself spawns new data, which justifies more policing.

Without proper feedback, the model cannot correct itself. It could be screwing up while its evaluated performance *appears* to be good due to the self-reinforcing feedback loop. In the context of fairness, this can cause biases to appear justified when they are actually artifacts of the model's decisions.

Proper feedback is not just a way of appeasing customers. It is critical to the maintenance and improvement of a deployed AIS.

- What are important considerations for deployment?

Evaluation. **This looks at differences between the test set and the population. It also looks at the role of fairness metrics when evaluating the deployed model.**

Graceful Degradation. **This looks at how the AIS can fail gracefully, as well as how users can opt out gracefully.**

Feedback. **This concerns feedback mechanisms for the AIS, which is needed to correct and improve the model.**

References

1. I Opted Out of Facial Recognition at the Airport — It Wasn't Easy [link]
   Funk, A., 2019. Wired.

2. A Framework for Understanding Unintended Consequences of Machine Learning [PDF]
   Suresh, H. and Guttag, J.V., 2019. arXiv preprint arXiv:1901.10002.

3. Big data's disparate impact [PDF]
   Barocas, S. and Selbst, A.D., 2016. Calif. L. Rev., Vol 104, pp. 671. HeinOnline.

4. No classification without representation: Assessing geodiversity issues in open data sets for the developing world [PDF]
   Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J. and Sculley, D., 2017. arXiv preprint arXiv:1711.08536.

5. Does Object Recognition Work for Everyone? [PDF]
   de Vries, T., Misra, I., Wang, C. and van der Maaten, L., 2019. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 52-59.

6. Weapons of math destruction: How big data increases inequality and threatens democracy
   O'Neil, C., 2016. Broadway Books.

7. The science and politics of race-norming. [PDF]
   Gottfredson, L.S., 1994. American Psychologist, Vol 49(11), pp. 955. American Psychological Association.

Summary Checklist

Here is a checklist of questions and prompts to ask when implementing an AIS. While there is no strictly correct answer, a good rule of thumb is that we should be okay with publishing our answers publicly.

This checklist is best completed as a group exercise and with extensive inputs from users and people who might interact with the proposed AIS.

SECTION 1 - UNDERSTANDING THE CONTEXT

General Context

1. **What is the ultimate aim of the application?**
   For example, for recidivism prediction, the true objective might be to make the society a safer place. As part of that, we want to identify individuals who might be prone to reoffending and offer them additional help to reduce future crime. Note the many implicit assumpations here. We assume that our sub-goal contributes to our objective. We also assume that reoffending is something that can be reliably predicted.

2. **What are the pros and cons of an AIS versus other solutions?**
   The main point here is to first weigh all the possible solutions instead of just implementing an AIS immediately.

3. **How is the AIS supposed to be used?**
   By answering this question, we can begin to think of ways that we can 'nudge' users towards the desired usage, as well as ways that the AIS can be misused.

4. **What is the current system that the AIS will be replacing?**
   How is the problem being solved at the moment? How is the proposed AIS better than this solution? How is it worse?

5. **Who will interact with the AIS?**
   This probably includes more than just the direct users that benefit from the AIS. Hiring models, for instance, interact with both employers (direct users) and job applicants.

6. **Create a few user personas - the technophobe, the newbie etc. - and think about how they might react to the AIS across the short-term and long-term.**
   This question examines the 'ripples' that the AIS might cause when it is implemented, ranging from the short-term to the long-term.

7. **Think of ways that the AIS can be misused by unknowning or malicious actors.**
   How can we design the AIS to prevent these misuses? If the potential harm is too great, we might want to reconsider adopting an AIS solution.

SECTION 1 - UNDERSTANDING THE CONTEXT

About Fairness

1. **What do false positives and false negatives mean for different users? Under what circumstances might one be worse than the other?**
   In recidivism prediction models for instance, false positives mean innocent people were wrongly accused. When we step from theory to the real world, we need to see that these mathematical concepts have very real meanings.

2. **Try listing out some examples of fair and unfair predictions. Why are they fair/unfair?**
   This is the first step towards trying to understand what are the protected traits in this context and how we should define fairness.

3. **What are the relevant protected traits in this problem?**
   Common protected traits include gender, skin color, ethnicity, age and physical ability. But remember that this really depends on the context and the culture that the application is situated in.

4. **Which fairness metrics should we prioritize?**
   Prioritizing means that some metrics are invariably compromised or violated. These decisions and their resultant shortcomings should be made known to users.

SECTION SECTION 2 - PREPARING THE DATA

1. **What is our population?**
   Note that this refers to the population that comprises all the possible inputs to the proposed AIS. This is important because it affects how we collect our data and evaluate our models later on. See Understanding Bias II for details.

2. **How does our dataset distribution differ from our population distribution?**
   In most cases, the dataset collected is different from the population. This is okay, but we have to be clear about how it is different and be aware of possible problems that might arise from the mismatch. See Understanding Bias II for details.

3. **Are we measuring the features/labels the same way for different groups?**
   Bias can creep in when we collect data differently for different groups. Check out Understanding Bias I for an example.

4. **How are our annotated labels different from the ideal labels?**
   Often, the labels that we really want is impossible or prohibitively expensive to obtain and we settle for proxy labels. Here, we ask, 'Are we using proxy labels?' and 'What are possible problems from using proxy labels?' See Understanding Bias II for details.

SECTION 3 - TRAINING THE MODEL

1. **How do our input features relate to our protected traits?**
   In cases where input features are protected traits, we need to justify their use in the model or remove them. We also need to check for correlations between protected traits and our input features, to identify proxies for the protected traits. These proxies can also be a source of algorithmic bias. See Understanding Bias II for details.

2. **Do we use the same model or different models for different inputs?**
   Using the same model assumes that the mapping between input samples and output prediction is the same for all groups, which might not be the case. On the other hand, training different models requires sufficient data for each model. See Understanding Bias II for details.

3. **If we are importing a pre-trained model or external data, what are possible conflicts between these imports and our current context?**
   Using pre-trained models and external datasets is a common practice. But these imported models and data can potentially carry hidden biases. See Understanding Bias II for details.

SECTION 4 - EVALUATING THE MODEL

1. **How does our test distribution differ from our population distribution?**
   Similar to Section 2 above, we need to think about the differences between our test dataset and our real population and possible problems that might occur.

2. **What can we say about the fairness of our final model?**
   More than just accuracy and other performance metrics, results from fairness metric evaluations should also be documented and made available to users. See Understanding Bias II for details.

3. **When we detect some unfairness with our metrics - is the disparity justified?**
   This lends some consideration for context to the quantification of fairness. Ultimately, how unjust a disparity is depends on the extent of disparity relative to its justification.

SECTION 5 - DEPLOYING THE SOLUTION

1. **How do we detect errors from the AIS after deployment?**
   The job's not over when the model is deployed. After emerging from the laboratory, the model needs to be continuously evaluated based on real-world data, to identify unexpected problems or model failure. Importantly, the model should not be caught in a self-enforcing feedback loop. See Understanding Bias II for details.

2. **What are alternative solutions in case of failure?**
   Just like any other technology, the AIS can and will break down. How can we design for graceful degradation for all types of failures (e.g. wrong predictions, total failure)?

3. **How can we allow users to gracefully opt out of the AIS?**
   Presently, there are people who are uncomfortable with certain AIS due to privacy and other concerns. How can we design for 'graceful degradation' that allows these users to opt out with minimal hassle? See Understanding Bias II for details.

Resources

This guide is just a brief introduction to main concepts in algorithmic bias. Here we list several more detailed resources that we found helpful while researching this guide, including other guides, prominent institutions and conferences, relevant datasets, software tools and academic publications.

OTHER WEBSITES AND GUIDES

- Useful or interesting links related to algorithmic bias.
- Survival of the Best Fit - a game about algorithmic bias in hiring
- Google's People + AI Guidebook and Inclusive ML Guide
- The Financial Modelers' Manifesto written by Emanuel Derman and Paul Wilmott was written for quants and financial engineers amidst the fallout of the subprime mortgage crisis, but the lessons are very applicable to today's AI engineers

> **The Modelers' Hippocratic Oath**
>
> - I will remember that I didn't make the world, and it doesn't satisfy my equations.
> - Though I will use models boldly to estimate value, I will not be overly impressed by mathematics.
> - I will never sacrifice reality for elegance without explaining why I have done so.
> - Nor will I give the people who use my model false comfort about its accuracy. Instead, I will make explicit its assumptions and oversights.
> - I understand that my work may have enormous effects on society and the economy, many of them beyond my comprehension.
>
> *Financial Modelers' Manifesto - Derman and Wilmott, 2008*

ORGANIZATIONS AND CONFERENCES

- The AI Now Institute is working actively on AI ethics and has many great publications
- FAT ML and ACM FAT* are two of the main conferences in AI ethics - check out the conference websites for related publications

DATASETS FOR THE MORE BIAS-AWARE

- Gapminder's Dollar Street images, which was used by DeVries et al. [2] in *Does Object Recognition Work for Everyone?* and comprises over 16,000 images from 60 different countries across 138 categories - a downloadable set can be found via my GitHub repository

- Google's Open Images Extended - Crowdsourced, - Google has also provided some notes on possible biases in this dataset - retrieved from the Kaggle FAQ:

> While we have targeted specific geographical locations in the collection of the Challenge Stage 1 dataset, it does have some particular areas of over and under representation that we found in preliminary analysis and wish to describe briefly here. These include:
>
> - Images of people tend to under-represent people who appear to be elderly.
> - Images tagged Child tend to be seen mostly in the context of play.
> - Some Person-related categories, including Bartender, Police Officer, and several sports related tags, appear to be predominantly (but by no means entirely) male.
> - Some Person-related categories, including Teacher, appear to be predominantly (but by no means entirely) female.
> - Some Person-related categories, including Teacher, appear to be predominantly (but by no means entirely) female.
> - Images with people seem to be taken predominantly in urban rather than rural areas.
> - Images of people in traditional locale-specific dress such as Sari's in India are relatively under-represented in this Challenge Stage 1 data set.
> - In images tagged Wedding, there does not appear to be representation of same-sex marriages.

- Joy Buolamwini's Gender Shades dataset [1] can be requested here

TOOLS

Tools for diagnosing and mitigating algorithmic bias, complete with detailed tutorials.

- IBM's AI Fairness 360 Open Source Toolkit
- Microsoft's InterpretML
- Tensorboard's What If

READINGS

Academic publications related to algorithmic bias that we found useful.

- Do Artifacts have Politics? (Winner, 1980) [4]
- Bias in Computer Systems (Friedman and Nissenbaum, 1996) [3]
- Technologies of Humility (Jasanoff, 2007) [5]
- Big Data's Disparate Impact (Barocas and Selbst, 2016) [6]
- Inherent Trade-offs in the Fair Determination of Risk Scores (Kleinberg et al., 2016) [7]
- Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment (Barabas et al., 2017) [8]
- Fairness Definitions Explained (Verma et al., 2018) [9]
- Fairness and Abstraction in Sociotechnical Systems (Selbst et al., 2019) [10]
- A Framework for Understanding Unintended Consequences of Machine Learning (Suresh and Guttag, 2019) [11]

References

1. Gender shades: Intersectional accuracy disparities in commercial gender classification [link]
   Buolamwini, J. and Gebru, T., 2018. Conference on fairness, accountability and transparency, pp. 77-91.

2. Does Object Recognition Work for Everyone? [PDF]
   DeVries, T., Misra, I., Wang, C. and van der Maaten, L., 2019. arXiv preprint arXiv:1906.02659.

3. Bias in computer systems [link]
   Friedman, B. and Nissenbaum, H., 1996. ACM Transactions on Information Systems (TOIS), Vol 14(3), pp. 330-347. ACM.

4. Do artifacts have politics? [PDF]
   Winner, L., 1980. Daedalus, pp. 121-136. JSTOR.

5. Technologies of humility [PDF]
   Jasanoff, S., 2007. Nature, Vol 450(7166), pp. 33. Nature Publishing Group.

6. Big data's disparate impact [PDF]
   Barocas, S. and Selbst, A.D., 2016. Calif. L. Rev., Vol 104, pp. 671. HeinOnline.

7. Inherent trade-offs in the fair determination of risk scores [PDF]
   Kleinberg, J., Mullainathan, S. and Raghavan, M., 2016. arXiv preprint arXiv:1609.05807.

8. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment [PDF]
   Barabas, C., Dinakar, K., Ito, J., Virza, M. and Zittrain, J., 2017. arXiv preprint arXiv:1712.08238.

9. Fairness definitions explained [link]
   Verma, S. and Rubin, J., 2018. 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), pp. 1-7.

10. Fairness and abstraction in sociotechnical systems [PDF]
Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S. and Vertesi, J., 2019. Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 59-68.

11. A Framework for Understanding Unintended Consequences of Machine Learning [PDF]
Suresh, H. and Guttag, J.V., 2019. arXiv preprint arXiv:1901.10002.

About

Hi there! I'm Lim Swee Kiat, a graduate student enrolled in the MSc. Urban Science, Policy and Planning course at the Singapore University of Technology and Design.

Machines Gone Wrong is part of my final project, supervised by Prof. Lim Sun Sun.

This guide is meant to introduce AI practitioners to the concerns of AI ethics, specifically algorithmic bias. As an undergraduate engineer and subsequent AI researcher, I found the social and ethical aspects lacking in the curriculum. This is my response to the problem and served as a way to document my learning journey about the topic.

Acknowledgements and Inspirations

The making of this guide was assisted and inspired by a ton of people.

- All the wonderful professors, lecturers, staff and my fellow classmates in the 2019 batch of SUTD's MUSPP course, for teaching me all about the urban and helping me through my struggles with interdisciplinary work, with special mention to Prof. Ate Poorthuis for all of his hard work in getting the program up and running!
- Randall Munroe for his wonderful xkcd comics, and whose style I have unashamedly copied.
- Bret Victor's 2011 essay on Explorable Explanations, and Nicky Case and co. behind explorabl.es.
- Chris Olah, Shan Carter and co. for creating Distill, which was a huge inspiration for this guide.