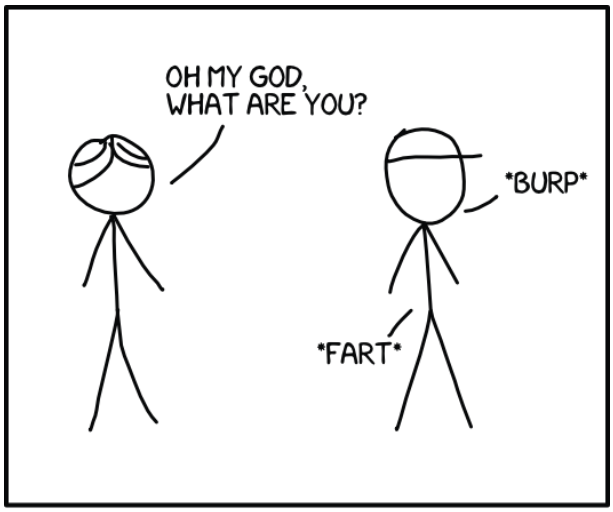
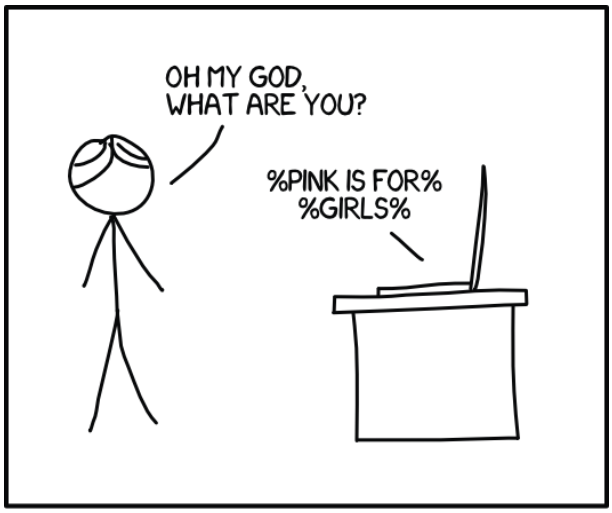
Machines Gone Wrong (Draft)

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.



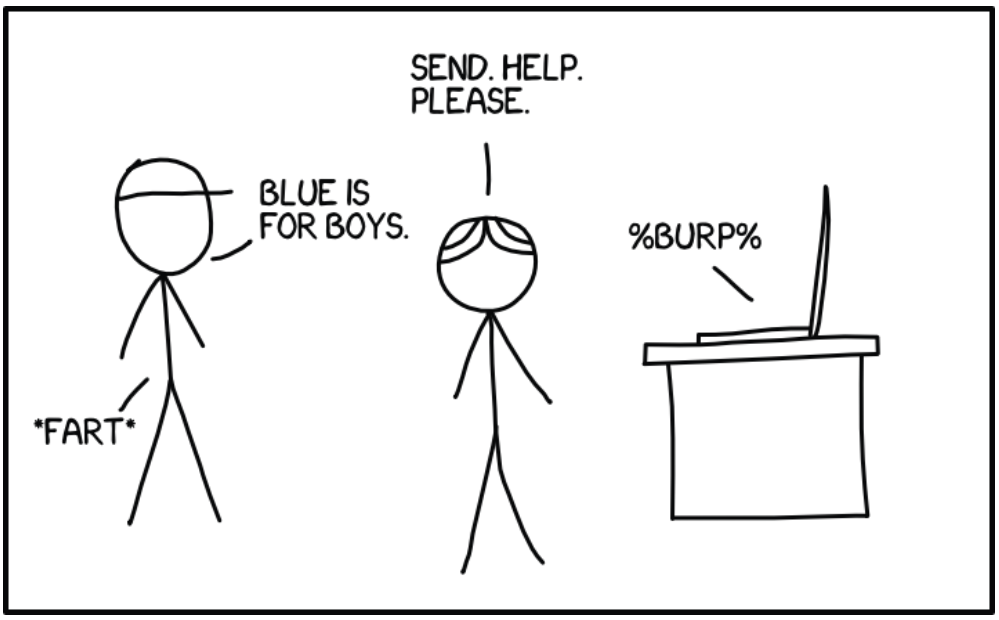
WHEN YOU REALIZE
THE LOVE OF YOUR LIFE IS HUMAN.



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, system failure 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?
- How do we decide which definition to follow?
- When is AI not the answer?

Collecting Data

- How do we know if our dataset is biased?
- How do we minimize bias in our dataset?
- What are some open-source datasets that are diverse?

Training and Evaluation

- How do we detect and reduce bias in the training process?
- How do we control the trade-off between bias and performance?

Deployment and Maintenance

- What should our clients and users know?
- How do we consistently evaluate our models for bias?

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:

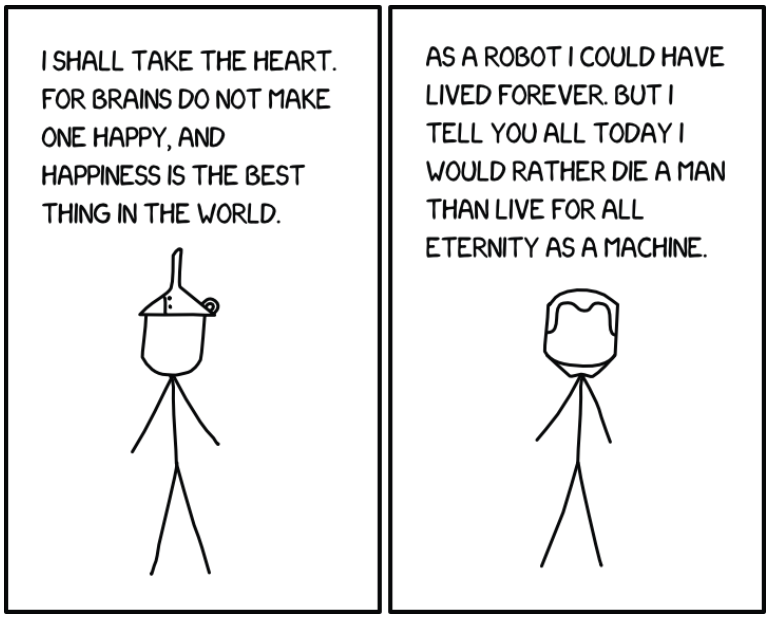
- What do we mean by AI ethics?
- What do we mean by AI systems?
- How is AI different from other technologies?
- What is the single most important question when implementing AI solutions?

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.

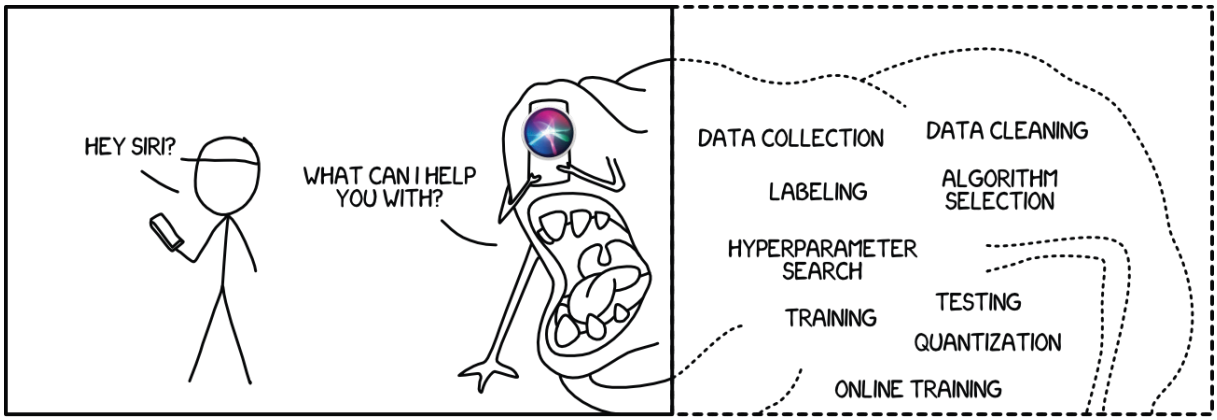
- What do we mean by AI ethics?

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?

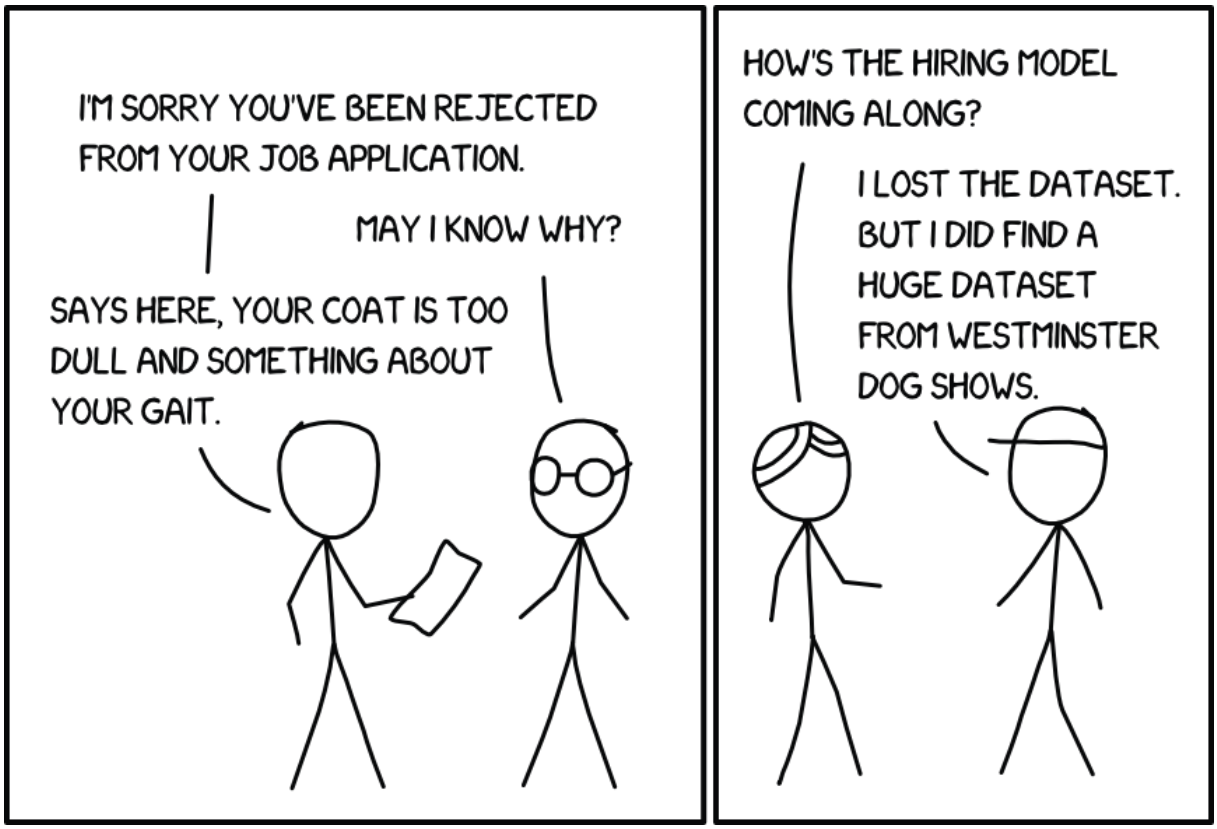
There's been many articles talking about how AI is the shit and how it's better than every other technology we've had. Here we look at three aspects that make AI stand out in terms of its social impact - an illusion of fairness, tremendous speed and scale, and open accessibility.

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

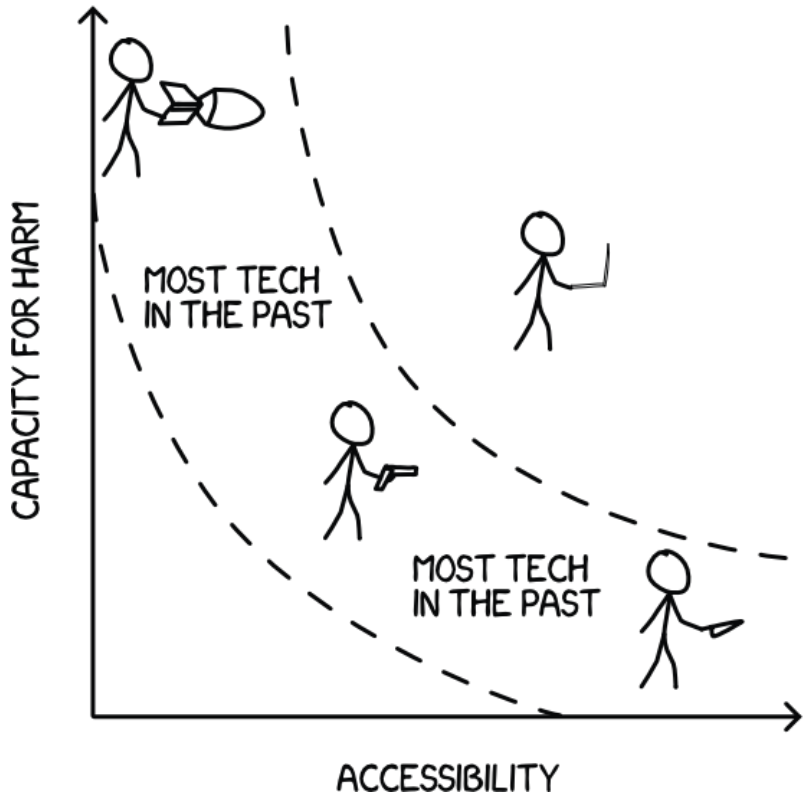
AI research has largely been open. As a self-taught coder and AI researcher, I remain eternally grateful for the kindness and generosity of the AI community. The vast majority of researchers share their work freely on arxiv.org and GitHub. Open-source software libraries and datasets are available to 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?

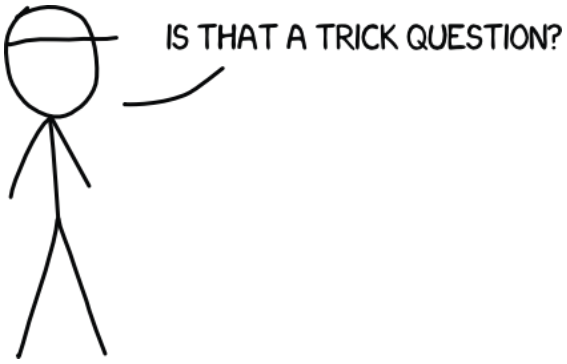
AI differs from most technologies in three aspects:

- **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

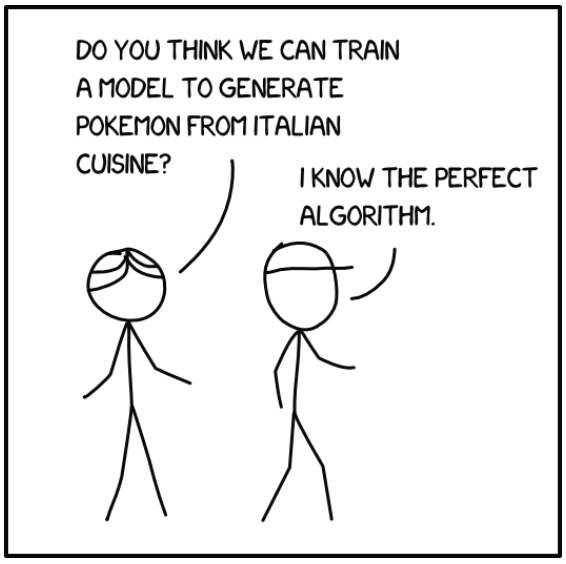
"WHEN IS AI NOT THE ANSWER?"



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?”

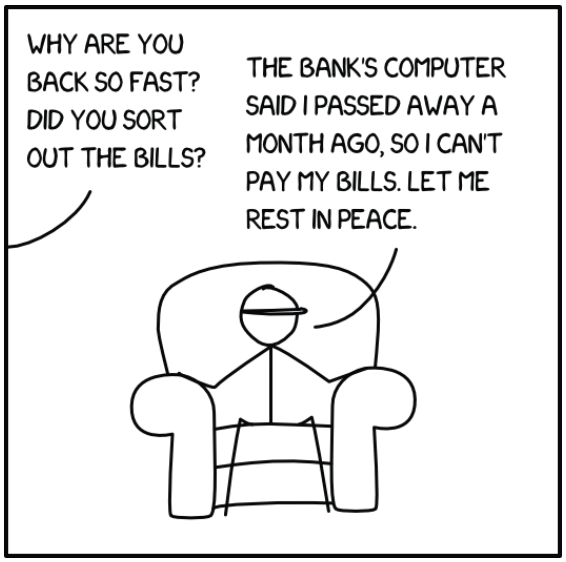
AI+Human systems are frequently perceived to be the best of both worlds. We have the empathy and explainability of humans augmented by the rigour and repeatability of AI systems. What could go wrong? Well, turns out documented experiences have shown that in such systems, humans might have a tendency 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.

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.**

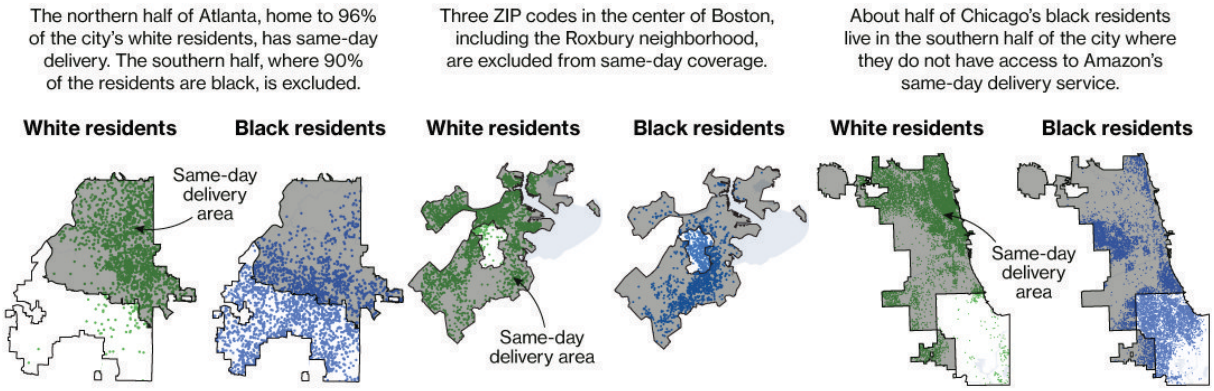
Disparate Impact

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. 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.

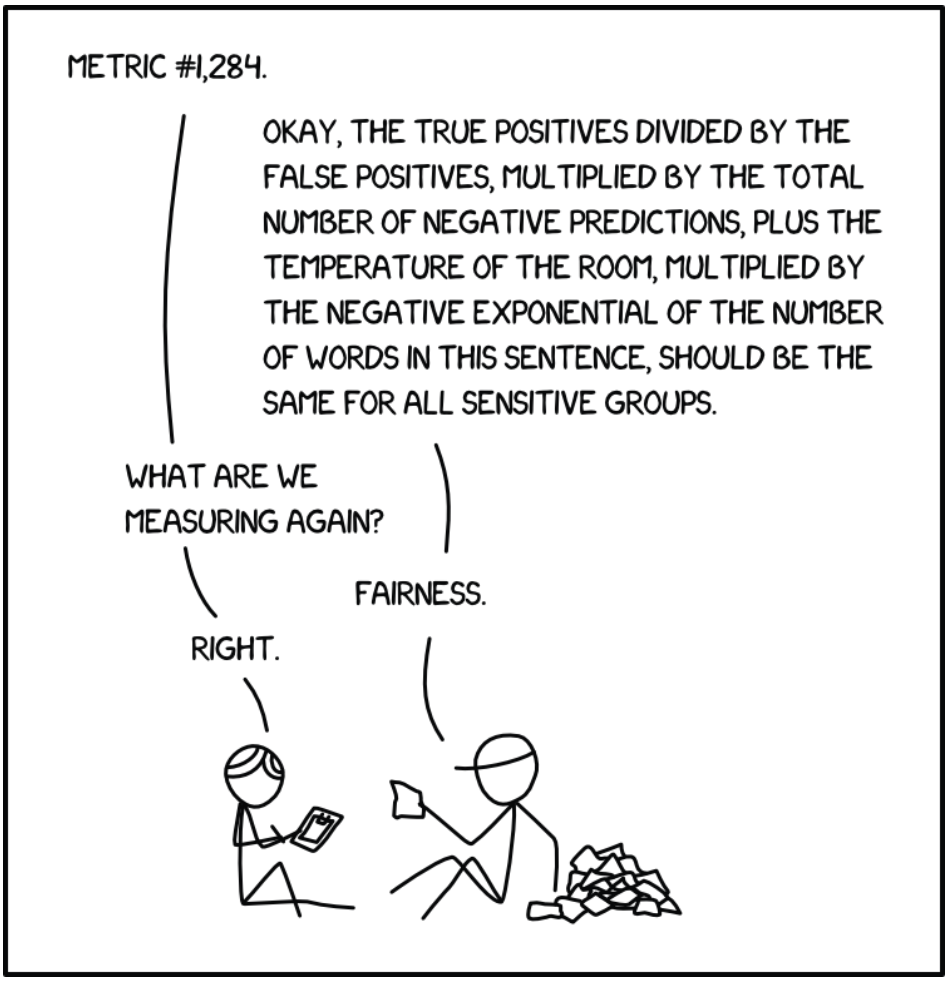
Well-calibration is a stricter form of calibration, with the added condition where the chance of being actually positive is equal to the score.

> 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.

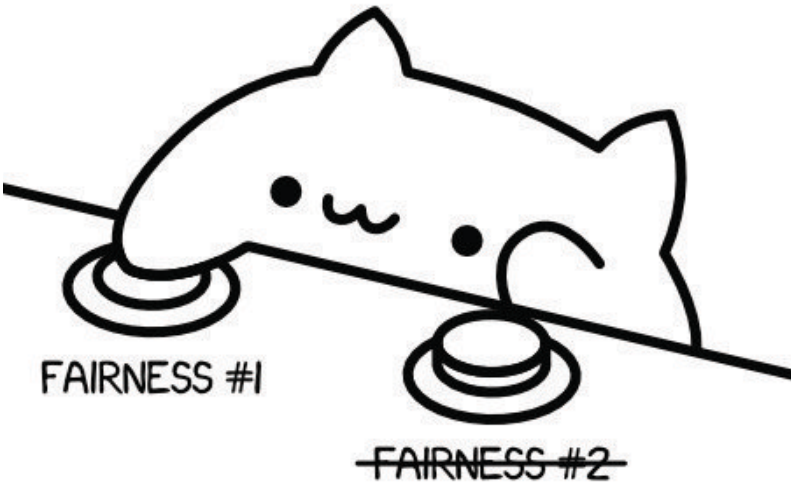
The fairness metrics can be a systematic way to check for bias, but they are only a piece of the puzzle. A complete assessment for fairness needs us to get down and dirty with the problem at hand.

- 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.

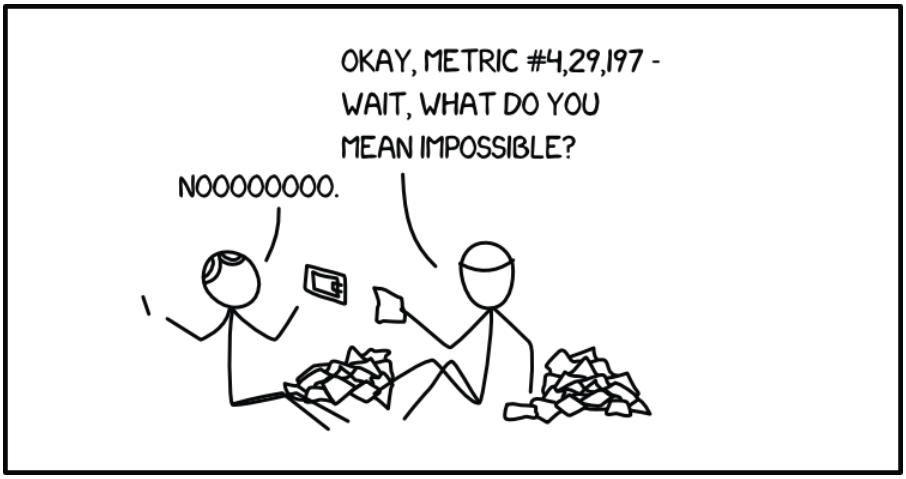
In ProPublica's well-known article Machine Bias, the subtitle reads:

> 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?

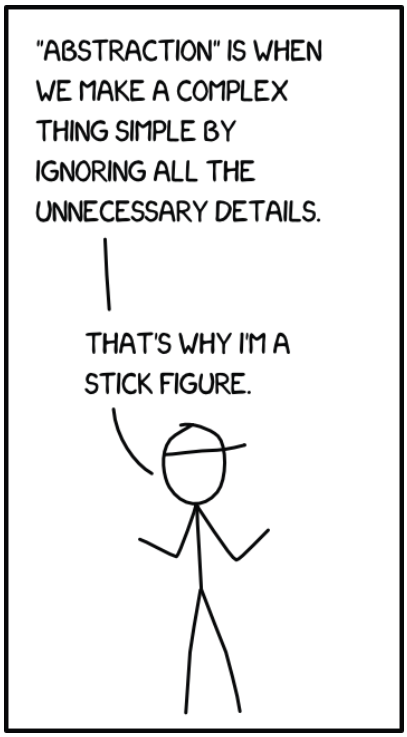


The point of all these is not to show that fairness does not make sense or that it is impossible. After all, notions of fairness are heavily based on context and culture. Different definitions that appear incompatible simply reflect this context-dependent nature.

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?

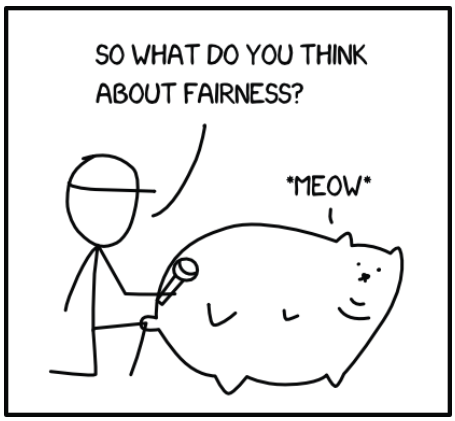
Nope we can't. Gotcha that was a trick question. The same decision can be both fair and unfair depending on the larger context, so context absolutely matters. As such, it is difficult to give advice on how to pick a fairness metric without knowing what is the context. Check out the next section for some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

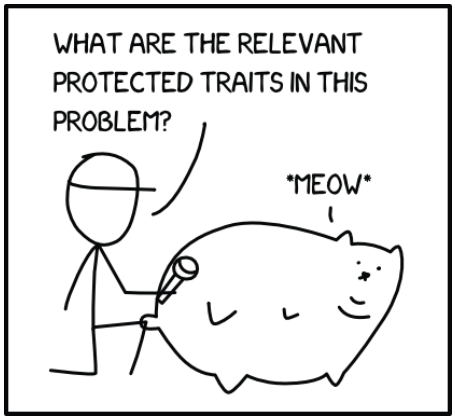
- What do false positives and false negatives mean for different users? Under what circumstances might one be worse than the other?
- Try listing out some examples of fair and unfair predictions. Why are they

fair/unfair?

- What are the relevant protected traits in this problem?
- If 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!