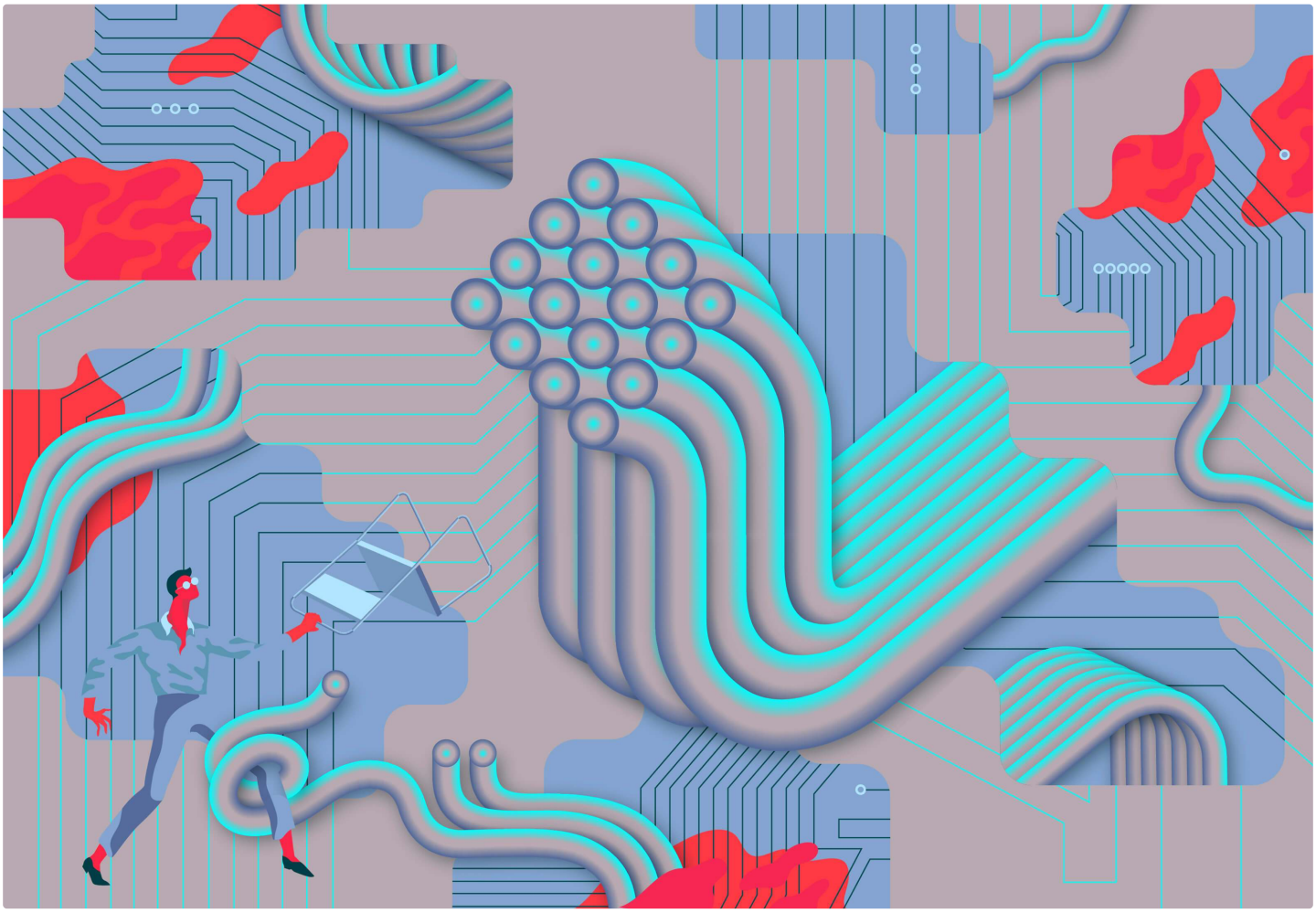


ANA KOVA FOR QUARTZ

STATE OF PLAY

The quest to make AI less prejudiced

✦ Member exclusive by Helen Edwards for AI's power problem

In 2016, researchers from Princeton University and the University of Bath made waves in the AI research

community with a landmark study. They looked at a common tool used by AI researchers to represent language, derived from a large database of text from the internet, and they found associations that strongly correlated with human biases—including mundane things like the fact that people find flowers more pleasant than bees and that weapons are less pleasant than musical instruments.

They also found associations that we would recognize today as stereotypes: female names are more likely to be associated with family than careers or with arts rather than sciences. And the biased associations they uncovered mapped onto real-world discrimination. Previous research had found that US job candidates with traditionally European-American names were 50% more likely to get job interviews. They were able to replicate that finding, using just the fact that European-American names were more closely related to pleasant words in their data.

These were disturbing revelations for AI. But the researchers' point was to start a conversation about bias not just in algorithms but in humans. Because prejudice is a human trait, which is dependent on cultural norms and an individual's actions, addressing bias in AI is not solely a technical challenge.

AI is transforming industries and has an uncanny knack for finding patterns, edge cases, and counter-intuitive

outcomes. It is immensely powerful and, used wisely, it can provide a platform for progress. But there is overwhelming evidence of its ability to harm, too. AI discriminates and stereotypes. It reinforces historical biases and it works better for some people than others.

A growing cadre of academics, activists, technologists, lawyers, and designers are confronting these biases and attempting to understand and mitigate them. It's pollyannaish to think we can just "de-bias" AI with the right software. Instead, the attempt to grapple with AI bias will force us to confront the biases in ourselves.

Table of contents

How bias happens

Bias has many causes but two categories are particularly important. If historical human bias is reflected in a dataset

used to train AI, the AI will likely exhibit the same bias. And if the dataset isn't representative—or even unbalanced—it will learn to predict things about some groups better than others.

Historical Bias

Many instances of bias are a result of long-standing bias in human societies. Human culture, language, associations, traditional roles and stereotypes all contribute to it.

A prime example of historical bias is Amazon's abandoned recruitment algorithm, which "learned" to downgrade resumes associated with women. That happened because the data reflected a historical bias: men were historically seen as a better "fit" for employment. Amazon scrapped the initiative in 2017 because executives lost confidence that the AI wouldn't just keep finding more ways to discriminate.

The problem in this case wasn't, as far as we know, that the dataset didn't have women in it. It was that women had historically been hired at lower rates, a pattern the algorithm effectively learned to mimic. Anywhere this sort of historical bias is embedded in the data, AI will likely be biased. (Amazon told Reuters that the tool "was never used

by Amazon recruiters to evaluate candidates" but otherwise declined to comment on its reporting.)

Another example of historical bias comes from health care. Recent research focused on the US population found that a commonly used algorithm assigned black people lower risk scores compared to white people of similar health. The algorithm used health costs to predict who would need more care. But black people spend $1,800 per year on average less than white people for the treatment of a chronic condition. This is for a variety of reasons, including lower incomes, less flexibility in transport and work schedules, as well as unconscious bias and overt discrimination by the health care practitioners, resulting in disparity of treatment.

In this case, the algorithm interpreted the lower cost of treating black patients as indication that they were healthier and therefore did not need as much additional care as white people. While this bias is obvious in hindsight, the designers of the algorithm did not recognize it ahead of time.

Bias in representation

Bias can occur because the data are not representative of specific groups. If a dataset contains lots of examples of

male CEOs and few examples of female CEOs, for example, a statistical model trained on that data to recognize CEOs is likely to be deeply biased. (Even datasets that are representative can be biased if they're not balanced between groups, because the algorithm has an incentive to learn patterns associated with the majority group to maximize accuracy.)

Another major source of bias are the labels that AI uses to help it learn about the world initially. Kate Crawford is an AI researcher at Microsoft and one of the founders of AI Now, an institute at New York University focused on the social ramifications of AI. In 2019, Crawford, together with artist Trevor Paglen, undertook an art project called "the archeology of datasets." They took one of the largest, most used image datasets, ImageNet, and studied the values by which it was labeled and constructed. Their tool, called ImageNetRoulette, "often returns misogynistic, racist and cruel labels." These labels are now an embedded part of many image recognition systems and the project was instrumental in demonstrating how systemic bias from AI systems is now a ubiquitous feature of our world.

Bias can be introduced because of missing context. In her book *Algorithms of Oppression*, Safiya Umoja Noble interviews Kandis, a black woman who owns the only local African American hair salon within a predominantly white neighborhood near a prestigious college town in the US.

When she was asked about her experience with the business review site Yelp, Kandis highlighted a data gap that she saw as important to how Yelp's algorithms prioritize her business:

> "Black people don't 'check in' and let people know where they're at when they sit in my chair. They already feel like they are being hunted; they aren't going to tell The Man where they are."

Representation in data is critical, and these cases highlight that there are many social and economic factors that affect people's representation in datasets. There are deep disparities across groups based on their access to the internet and their online behavior, which means dealing with bias not solely a technical process.

How bias harms

Even when statistically neutral and technically de-biased, if AI is used inappropriately or is poorly designed, it can perpetuate discrimination and result in unfair outcomes. And AI raises the stakes because bias can now harm more people, faster.

Human bias in recruitment is a huge problem which denies millions a fair and equitable chance to prove themselves. AI has a role to play in solving this problem by efficiently sorting desirable candidates, predicting those who are most likely to succeed, reducing the impact of human prejudice and unconscious bias and screening people through AI-interview processes. This is big business: the top two companies based on funds raised are HireVue ($93 million) and pymetrics ($56.6 million).

But the biases of a recruiting algorithm created by a popular hiring platform will have far greater impact than even the hardest-working recruiter. Its biases become a single point of failure.

Hiring algorithms assess candidates' suitability and personality based on videos and games, and by analyzing body language, speech patterns, mouse movements, eye tracking, tonality, emotional engagement and expressiveness. Hundreds of thousands of data points are gathered in a half hour interview or online game-playing exercise.

One of the biggest problems with this process of AI screening is that the science isn't keeping up. "Academic research has been unable to keep pace with rapidly evolving technology, allowing vendors to push the boundaries of assessments without rigorous independent research," say

researchers from Cornell and Microsoft in a recent paper.

While the intent is good, there is a lack of data on to whether AI is improving recruitment diversity and fairness or introducing new sources of bias, which are then applied at scale.

And even when AI isn't biased in a technical sense, it can be used to discriminate.

"Affinity profiling" is the practice of grouping people based on their assumed interests rather than on their personal traits. This is common in online advertising and it has the potential to be discriminatory if people do not see certain ads or receive different prices based on their affinity. This "discrimination by association" can be difficult to detect, which means that, from a legal perspective, it can be hard to remedy.

Affinity profiling is common in online advertising. For instance, as recently as 2016, Facebook allowed advertisers to use "ethnic affinity" as a proxy by which to target people by race. In 2018, the company removed 5,000 of these affinity categories to address concern from activists and lawmakers. The company also added a requirement that advertisers behind ads for housing, jobs, and credit cards comply with a non-discrimination policy.

Even so, AI's ability to detect subtle correlations between interests and personal traits like race and gender make it exceedingly difficult to identify when someone has been discriminated against. This form of discrimination matters because of its opacity, which effectively allows discrimination to hide in another guise. Scholar and teacher Chris Gilliard calls it "friction-free racism." Affinity profiling is indirect and is seen as neutral so it "allows people to feel comfortable with racism," Gilliard says.

Dealing with bias

Individuals, companies, and societies are beginning to grapple with AI bias through new programming tools, design techniques, laws, and more.

In practice, there are a variety of strategies that people use when dealing with AI bias and harm that comes from AI systems. New technical tools are being developed. Human-centered AI design practices and ethical AI design are being refined as more people work on AI-enabled products. Lawyers are finding ways to challenge unfair outcomes from automated decision-making systems. And people are reflecting on AI's role in the power structures of society, and how fighting bias means fighting power.

Debiasing and fairness testing

“Debiasing is now table-stakes,” says John C. Havens, the director of emerging technology and strategic development at IEEE, a prominent association of technical professionals. Havens plays a major role in advocating for standard practices in AI ethics. He strongly recommends that all AI models be rigorously examined for bias and tested for fairness and, with new technical tools available, implies there is no excuse for not doing so.

A lot can be achieved with purely technical tools. Documenting sources of bias, testing for fairness, de-biasing models, and archiving previous model versions are considered best practice.

A critical step in that process concerns understanding the data representations and choices the AI is making. There are a host of tools that have been developed to help there, including IBM’s AI Explainability 360, Google’s What-If tool and LIME (Local Interpretable Model-Agnostic Explanations) from the University of Washington, for example. These tools are designed to help data scientists understand the model’s most important features and how it makes predictions. They all combine visualizations with

sophisticated data tools which allow data scientists and engineers to examine and manipulate data.

Even so, most AI experts acknowledge that an unbiased dataset is not possible. Bias, in some form, will always exist, which means that it's vital to understand how bias affects different groups. This is the role of fairness testing.

There are multiple technical definitions of fairness, all based on what happens to different populations when AI makes an incorrect prediction. The most simple idea of fairness is to ensure some form of parity across a predetermined list of groups, often based on legally protected categories like race or gender or in other domains where discrimination is known to be common.

AI can make four types of predictions. As an example, imagine you are a recruiter and you use a pre-recruitment algorithm that tests candidates, gives them a score, and then recommends candidates for you to interview based on their score. It can:

- Recommend a candidate that it correctly predicts would be good at the job; a true positive (TP)
- Recommend a candidate that won't be good at the job; a false positive (FP)
- Not recommend a candidate that wouldn't be good at the

job; a true negative (TN)

- Not recommend a candidate that would be good at the job; a false negative (FN)

Statistical fairness tests use error rates (false positives and false negatives) to test various ratios of failure between different groups. There are many different types of fairness tests but they fall into three broad categories: individual fairness, where similar predictions are given to similar individuals; group fairness, where different groups are treated equally; and subgroup fairness, which tries to balance both approaches by picking the best properties of the individual and the group and testing across various subgroups.

These are some examples of commonly used metrics:

- Group fairness: Equal positive prediction rates (TP + FP)
- Equalized odds: Equal false positive rates (FP / (TN + FP)) and equal false negative rates (FN / (TP + FN))
- Conditional use accuracy equality: Equal positive predictive values, also known as Precision (TP / (TP + FP)) and Equal negative predictive values (TN / (TN + FN))
- Overall accuracy equality: Equal accuracies (TP + TN)

- Treatment equality: Equal ratios of wrong predictions (FP / FN)

The problem is that there is usually a conflict between accuracy and fairness. The core tension is making optimal decisions for the system as a whole versus maintaining parity between groups. Therefore, even with a plethora of technical definitions, fairness testing remains context and value dependent. It involves making decisions about the kinds of mistakes that are made and how these mistakes are distributed between different groups. In *The Ethical Algorithm*, Michael Kearns and Aaron Roth point out that the tension between fairness and accuracy will never go away, but that it can now be measured and managed better than in the past.

> "In the era of data and machine learning, society will have to accept, and make decisions about, trade-offs between how fair models are and how accurate they are. In fact, such trade-offs have always been implicitly present in human decision-making; the data-centric, algorithmic era has just brought them to the fore and encouraged us to reason about them more precisely."

Perhaps one of the biggest challenges in the technical domain is actually a human one. Maria Axente, Responsible AI Lead with PwC UK in London, works with both executives

and technical teams to help them understand the importance of the data supply chain. She says she starts any conversation about tech ethics by asking, “How biased is the context in which a technology is created?”

It’s vital that data scientists are encouraged to think beyond the strict boundaries of their role and to consider the consequences of their work. This helps to reduce bias being introduced as a product of someone’s background or because of a certain preference which unconsciously contributes to design choices, which then means that unintended bias is amplified at scale without any human making a conscious choice. As Kearns says, “If you don’t say anything about fairness, AI won’t either.”

Human-centered design

As technology and AI diffuse through everything—from devices in our homes, to the apps we use to track our health, to sophisticated equipment in industry—design plays an increasingly important role. We expect things “just to work,” and that expectation extends to AI. “Almost all design stems from making sure that a user can figure out what to do, and can tell what’s going on,” Cliff Kuang and Robert Fabricant write in their history of user-centered design, *User Friendly*.

Design has long sought to knit human psychology and product functionality together. A fundamental idea in design is feedback: how we adjust our predictions based on our experience. AI presents a unique design challenge because it can be hard to figure out what's going on and bias can amplify inaccurate and unreliable feedback. AI also speeds up the cycle of feedback—think of how rapidly "likes" on Facebook can result in ads related to those likes.

Human-centered AI design is an emerging practice with the goal of making better AI. People who work in the area generally value putting humans first, with the belief that AI should serve humans, rather than the other way around. Bias that is not understood or revealed may not satisfy the user's need and may be confusing, which disrupts the core tenet of design.

Many designers consider the idea of unbiased AI to be unrealistic and counter to technology goals. Josh Lovejoy, head of Design, Ethics and Society at Microsoft, says that bias reflects a latent prioritization. The goal of the designer is to reveal this in a productive way. Instead of aiming to "de-bias" an AI, designers need to be explicit about priorities and heuristics.

Instead of setting a goal for an AI to be as accurate as possible across as big and diverse a population as possible, Lovejoy suggests a different trade-off: more models with

each designed for “narrow-utility.”

He illustrates this idea with an example: designing an AI to detect fake names in online forms.

In commonly used datasets there are patterns between the length of a surname and the likelihood of that name being marked as a likely fraud case, according to Lovejoy. Training data that is US-centric is biased: very long or very short names, and names that contain hyphens, are more likely to be flagged as fake, giving rise to clusters of error. But the pattern varies in different parts of the world. One answer is to have a greater number of narrow models instead of generalized deeper models. These models are localized, which has the effect of reducing bias. This is less convenient for the technologists but better for users—training data may be more difficult to gather, models are more fragmented, but each individual user has more influence over the AI, which increases how useful the AI ultimately is.

Human-centered design has a bias towards understanding the natural ways that humans interact with a product and provide feedback that helps humans develop an accurate mental model of the system. A mental model is our intuition about how something works, such that we can make an accurate prediction about what it will do. AI introduces additional complexity because it fails in unpredictable ways.

Lovejoy challenges technologists and other AI designers to maximize ways that users can play a bigger role in the AI. The more a user can interact, the more the user plays a role in the AI, the more the AI becomes a collaborator with a human who participates by making active choices. This gives the user a better mental model of the AI which builds trust.

How humans build mental models of AI systems is an active area of research, according to Tom Griffiths, professor of information technology in Princeton University's psychology department, and co-author of *Algorithms to Live By*. "Humans are good at thinking about management of our own cognition," Griffiths says. We can reason about how we, ourselves, would solve a problem and use this reasoning to build an internal model of how an AI would solve the same problem. The key to this process is interaction. "We underestimate how good we are at interacting with black boxes," Griffiths says. "We already engage with human beings as other black boxes."

Milena Pribic, a design advisor at IBM focused on AI, sees trust as "an act that a user takes"—so something that can be measured and monitored in user actions. In this way, bias is intentional. It is set by product design teams as part of a design. The AI has a personality, a style, a tone, a goal of its own. Without conscious design choices made up front that help set a user's mental model, people can come to an AI with an anthropomorphic bias or have expectations that are

too high. The result is frustration and “trolling” of an interaction which ultimately amplifies any existing bias in the AI.

AI is different because of the “elongated engagement” that users have with an AI. The first encounter can be very different from the thousandth and the user needs to have autonomy to set and guide the relationship.

Legal solutions

AI challenges the legal landscape. Tech giants, antitrust, privacy, surveillance, ad micro-targeting, discrimination, and bias are all hot topics. While new laws are likely required, it’s not clear when, how, and to what extent new regulations will be introduced to deal with the new challenges of AI.

Algorithmic decision-making that results in discrimination or disparate treatment, that is done without notice, or in ways that humans cannot understand and explain, is increasingly being investigated by journalists, legal scholars, human rights-focused non-profits, as well as challenged in the courts.

In the US, some of the most prominent and successful cases against AI-enabled discrimination have been taken against state governments. These cases have had one common factor: automated systems, which stand in for human decision makers, and which have denied people their constitutional rights.

In Arkansas, disabled beneficiaries lost half their benefits with no notice and no one was able to explain the algorithm's decision. In DC, a criminal risk assessment tool for juveniles constrained sentencing choices, sometimes only displaying options for treatment in a psychiatric hospital or a secure detention facility, which drastically altered the course of people's lives. In Michigan, an AI system for unemployment benefit "robo-determination" of fraud, adjudicated 22,000 fraud cases with a 93% error rate. Twenty thousand people were subject to the highest-in-the-nation quadruple penalties, amounting to tens of thousands of dollars per person.

The fact these cases were won by the plaintiff now provides precedent. People are beginning to push back on unjust treatment and AI-enabled discrimination. Schultz says, "It takes time for people to 'learn how to get justice' when new technologies are involved. Lawyers take cases on a pro-bono basis, technical experts will also often work for free as people learn and "build on each other's wins." The recent disability austerity cases in Arkansas, Idaho, and Oregon

were won using this exact formula."

The hope is that state and federal governments focus on doing a better job of understanding the implications of taking an engineering mentality to social systems. There is a lack of awareness about how AI systems fail and a lack of training and planning for dealing with technology failure in vulnerable social systems. When human decision makers are removed from the loop, "the probability of harm is very high." Which means that providers of social services should be thinking carefully about whether AI should be used at all.

Schultz questions whether there is any role for AI in the justice system. One of its fundamental principles is the right to a fair judgment as an individual. AI groups, clusters, classifies and uses proxies which may be incompatible with fairness and justice.

Schultz explains that the legal system the US is "in flux." Many foundational legal principles break down in a world powered by intelligent, autonomous systems. But he does see the system responding. "Using tools such as AI Now's Algorithmic Impact Report make me hopeful we can come back from the *Black Mirror*," he says, referring to the dark future depicted by the UK TV series, where AI drives humans into a world of algorithmic autocracy.

Regulation

One approach is to regulate AI. But regulating fairness requires defining it, which means grappling with the tradeoffs between fairness and accuracy, for example.

"The next frontier is a large technology company articulating and justifying the bias that exists in their models," believes Michael Kearns, professor of computer and information science at the University of Pennsylvania. Naturally big tech will resist any calls to regulate AI in ways that would make their models and data more open, because doing so would compromise their intellectual property. But, with the scale and speed of AI, the current reactive regulatory approach falls short. The damage is discovered only long after the harm has occurred.

A move to a proactive regulatory regime would need to look more like what happens in the financial system, with FINRA, the industry's self-regulatory agency which has direct access to highly granular trading data. Kearns points out that similar issues of speed and scale are at work in the finance industry, where regulation has gone "real-time," with sensors placed in data feeds. With direct monitoring of data in specific ways, it would be possible for people to monitor for issues of bias inside of tech companies without needing

to fully understand the models or digest the data at the speed and scale that a tech company operates at.

According to Kearns, there is an incentive for tech companies to allow this form of limited external oversight. It prevents them from being blamed for something that is beyond their control, say where the effect of algorithmic bias crosses an organizational boundary. An example of this is recent research on whether Google's advertising shows gender bias in hiring for science, technology, engineering, and math (STEM) roles, showing more STEM ads to men than women. While on the face of it, Google's model may be biased, at least part of the cause was advertisers being more willing to pay higher prices to get clicks from women for products specifically targeted at women. Ads for STEM roles were outbid by ads for other products.

Technology companies may resist regulation. The other issue, says Kearns, is that while the science of AI has advanced, the regulators themselves are decades behind.

Bias and society

When we choose to delegate physical, emotional, cognitive or ethical work to a machine, we outsource a part of

ourselves. Outsourcing to an AI can make us more efficient but it can have unintended consequences and there are always trade-offs. It can make us more passive, decrease our sense of responsibility, decrease our agency, and make us detached or helpless.

Machines have interacted with humans for a long time but AI raises the stakes. Traditional technology and expert systems were typically developed based on known rules and heuristics. AI is different. Much of how an AI behaves relies on what it encounters "in the wild" and in its interactions with people. AI is "world-creating in a way that other technologies aren't," Jacob Metcalf from Data and Society says.

AI increasingly substitutes for, or enhances, human decision-making. Because AI learns and acts on its own and changes as new data about the world is made available, it can help guide humans in ways that traditional technology does not. This interactivity and shared agency has existential consequences for humans.

Annette Zimmermann, postdoctoral researcher at the Center for Human Values at Princeton, worries that people may reject AI because many AI applications have been shown to be biased in a way that exacerbates social injustice. This, however, does not necessarily mean that all AI applications will be biased and harmful in the same way. Zimmermann

emphasizes that while we should subject AI to critical scrutiny, we shouldn't assume that anything involving AI should be rejected.

What really matters, Zimmermann argues, is that we as a society find a way of implementing more and better democratic processes in this area, so that people can make more informed and direct choices about what AI should and shouldn't be used for. This means people getting involved where AI is used locally—in their cities, schools and communities. It means municipalities creating ways for people to learn about the technology and its social implications—and creating reliable and transparent ways for people to hold governments and corporations accountable when AI deployment leads to unjust outcomes.

AI's role means that these choices can't be a one-time thing. There is a need to constantly evaluate how AI interacts with the world.

Human decision-making works on a geologic time scale compared to how quickly biased and scientifically unsound AI can cause destruction. How can we make space for human-scale reflection and collaborative decision-making?

Entrepreneurs are trying to create that space. Leanne Carroll, a graduate student at the School of Visual Arts in NYC, recognized early on that consumers need a way to

surface bias. She's working on a platform, a "Kickstarter for bias" where people can highlight how they have been affected by an AI system and seek assistance from the AI community and product owners in dealing with it.

Another approach is to use AI and other digital technologies to reflect our unconscious bias back to us. Oregon-based start up, Shift, uses VR and AI to teach people about human unconscious bias and help them make changes to their behavior. Leveraging the scale of AI to train more people and change how we propagate bias is powerful. It borrows an idea from self-driving cars, where new knowledge can be updated and uploaded faster and at scale.

AI can play a role in addressing social problems. It can be used to diagnose and measure them with more precision and clarity than in the past. It can also be used to expose and reframe social problems in new ways, which enables new solutions. AI has an ability to expose bias in ways that make us all more aware and motivated to change.

At its core, AI gives us insight into how human minds work. Humans evolved ways of thinking under constrained resources of energy and time. "Human bias is a necessary consequence of the constraints we are under," according to Tom Griffiths.

But, as a society, we need to be ready, and have the capacity,

to discuss the biases that AI will inevitably reveal.

QUARTZ

The most dangerous AI bias is the bias of the more powerful over the less powerful.

NO JUDGMENT

We can’t address bias in AI without considering power

✦ Member exclusive by Helen Edwards for AI's power problem

Sometimes it takes something unexpected to shift people's perspectives. That's what a group of MIT and Harvard Law School researchers were aiming for when they set out to reframe fairness in AI by studying its use on the powerful rather than the powerless. They presented the results of their research in January at the ACM Conference on Fairness, Accountability and Transparency in Barcelona.

In the US, over half a million people are locked up despite not yet having been convicted or sentenced—a result of pretrial detention policies. Ninety-nine percent of the jail growth since 2002 has been in the pre-trial population, much of this because of an increased reliance on bail money, according to a report by the Prison Policy Initiative. As a result, the report's authors write, "local jails are filled with people who are legally innocent, marginalized, and overwhelmingly poor."

Using theories borrowed from social justice work, the MIT and Harvard research team built a model to test how personal agency affects the accuracy of AI in this context.

"Just as arrest data tells us more about the police than it does about defendants, we wondered if there would be more information in patterns of judges' behavior than in the patterns of defendants in pre-trial evaluation for bail," Chelsea Barabas, one of the authors of the paper, told Quartz.

AI models are all about prediction. The researchers hypothesized that prediction accuracy is a by-product of agency: people who have the power to make their own decisions should be more predictable than those buttressed by countless other complex forces. In the

courtroom, that meant hypothesizing that judges' behavior would be more predictable than defendants'.

That is exactly what they found. The judges' decisions turned out to be more predictable than those of the defendants.

A key outcome for pretrial risk assessment is estimating whether someone will fail to appear for a future court date. The researchers flipped this around, developing an alternative prediction of whether a judge would detain the defendant for more than 48 hours, a measure they dubbed "failure to adhere" and treated as a proxy for imposing unaffordable bail without due process of law. Using the same data, but by interrogating it from a different perspective, their "alt-FTA" achieved an accuracy of 80%. Mainstream "FTA" score models are accurate around 65% of the time. Whether the judges would impose excessive detention was more predictable than whether defendants would show up in court.

The researchers stress that their algorithm is not intended for any practical use. Instead, their intention was to demonstrate that for AI to be fair and unbiased, power matters. They wanted to establish a "counter-narrative, a risk assessment that 'looks up' at judges," and subjects "those in power to the very socio-technical processes which are typically reserved for only the poor and the marginalized."

How might judges feel to see their work reduced to an algorithm—one that didn't paint them in a favorable light or leave much room for individual circumstance? That, of course, is what sentencing algorithms do to less powerful individuals on a daily basis.

That one aspect of judicial decision-making is more predictable than

one aspect of defendants' decision-making doesn't really say much on its own. Different phenomena are easier or harder to predict, and that predictability doesn't always map easily onto power. However, the process of developing the algorithm revealed a dilemma: data that reflect poorly on the powerful may be less likely to see the light of day. And, in fact, the researchers reported varying levels of cooperation from courts throughout their work.

he most dangerous AI bias is the bi of the more powerful over the less powerful. ”

This would introduce another source of bias: a lack of transparency would make it more difficult to question the decisions of those who have the power. As the researchers point out: “Of what value is this access if it is contingent upon refusing to question the unchecked assumptions and premises of the data regime itself?”

This last point is important. The most dangerous AI bias is the bias of the more powerful over the less powerful. Fighting bias, more often than not, involves fighting power. If the data sources and structures of AI are not able to be challenged by those external to the process, then there is no true challenge to power and no way to honestly correct for bias. (This is part of why addressing bias in AI

requires more than technical fixes.)

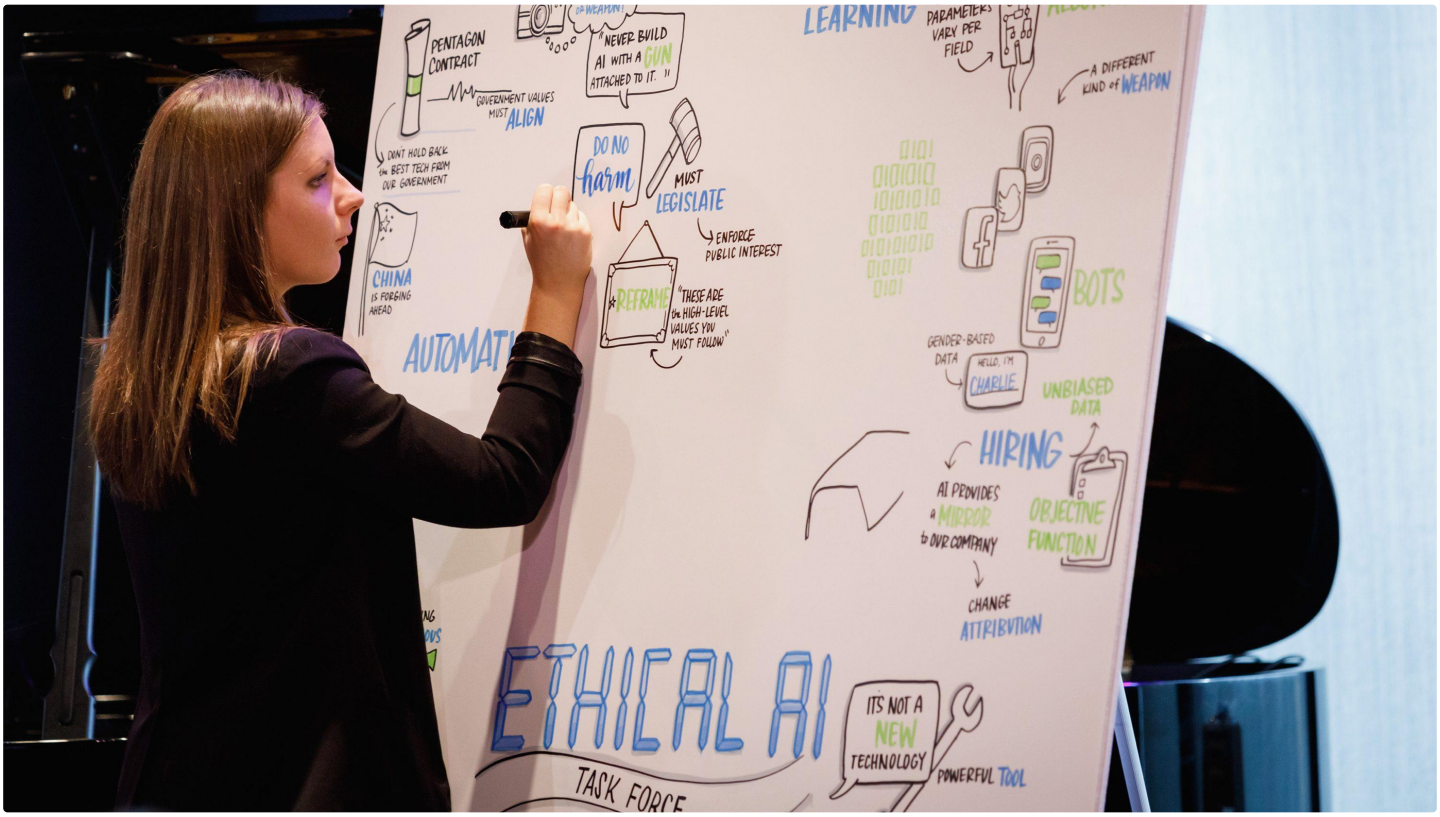
A data scientist doesn't need to be prejudiced to design a prejudiced algorithm. All that is required is that the data scientist conforms to the status quo, including accepting any inherent data bias and any pre-existing imbalance in power structures as an accurate representation of the world. But as the paper's authors point out, "Data and their subsequent analyses are always the by-product of socially contingent processes of meaning making and knowledge production."

For AI to be fair, data scientists will need to include the social context in which the algorithm acts.

The researchers' inspiration came from an unlikely corner: anthropology. In the 1970s, the dominant paradigm in anthropology was to study people at the periphery of western culture. Those who studied had "the relative upper hand," while those who were studied were "the underdog." The challenge, which came from scholar Laura Nadar, was to shift the field to study powerful strata in society rather than simply communities at the margins. As a result, scholarship and practice became more methodologically and ethically complex, uncovering hidden assumptions and delivering richer insights. This, in many ways, is the type of fundamental shift we need today in AI.

Tackling bias in AI is most often thought about in narrow technical terms, where the role of the data scientist is to be neutral and apolitical, where complex social problems yield to the processes of data distillation and "objective truths" are fractionated qualities of AI. But if we want AI to help usher in a fairer society, this narrow view of technology, data, and the role of the data scientist will need

to change in far-reaching ways.



MICHAEL COHEN/GETTY IMAGES FOR THE NEW YORK TIMES

Ethical AI urges technologists to “think slower.”

WORK ETHIC

Are AI ethicists making any difference?

✦ Member exclusive by Helen Edwards for AI's power problem

Tech companies, consulting firms, and even the US military

are rushing to add ethics boards and hire "AI ethicists." They're asking them to think about everything from bias and fairness to the circumstances under which it is acceptable to use autonomous weapons.

It's a welcome acknowledgment of the harm AI can do.

But many people are skeptical about the role of ethicists in technology companies—including both ethicists and technologists. Trained ethicists in particular have bristled at this new role, often because some "AI ethicists" don't have formal training in ethics and moral philosophy.

Josh Lovejoy, a leading AI ethicist and designer himself, says that big tech "loves to break words." Corporate speak—or "garbage language"—changes how we use words: agile used to mean nimble and now it's a software development process. 2020 could be the year that tech breaks "ethics" and "bias," Lovejoy fears, by turning them into corporate jargon.

Google and other tech companies have already been accused of "ethics washing," the term used to describe the weaponization of ethics as a defense against regulation. By employing a group of high-status specialists, tech can claim to have AI ethics under control and thereby deflect questions from activists and policymakers bias and other risks of AI.

That skepticism is often warranted, but it introduces an alternative risk. If the pendulum swings too far the other way, good ethical practices and initiatives may suffer from "ethics bashing," with the result being even less discussion of ethics within tech.

Neither of these extremes are productive. Elettra Bietti, a researcher at Harvard Law School, is an outspoken critic of both the ethics washing and ethics bashing camps. She argues instead that individuals need to see ethics as a mode of inquiry that helps people evaluate competing technology choices—whether they be policy strategies or design choices. Ethics should be measured by how it enables participation because ethics, in practice, often involves redefining boundaries. "It's important to realize that ethics is something that we all do, all the time," she says. "And every action can be embedded within a broader framework of justice."

What this all means for AI ethicists and ethics boards is that they should pay careful attention to how their actions impact key business decisions at key check points. For instance, how is the ethics group constrained by the existing strategy? If an ethics board recommended that the only way to deliver an "ethical AI" was to abandon the entire product line and start again, would they be able to collectively voice this recommendation? Or, if an ethics board made a decision that positively affected the impact of a product in a tiny way,

what would be larger: the positive impact on the users of the product or the positive impact on the reputation of the company, by virtue of the board's presence?

Jacob Metcalf, a researcher at Data and Society, a technology think tank, describes ethics as "the vessel which we use to hold our values." This is a useful idea because it does allow for ethics to be a process. Maria Axente, AI ethics specialist at PwC, is also constantly challenged with measuring ethics: separating the substantive choices that need to be made in AI design from the process that is required to support it. "How do we distill thinking into action, substance into structure that people can practically implement?" Axente asks. A perennial challenge in AI ethics is moving beyond people's personal intuitions to talking about real-life problems to be solved.

Lovejoy argues that ethics need to be seen not as some sort of humane or philosophical add-on, "but as 'just good design' that works when it's been validated in the real world." What's needed, in this view, aren't AI ethicists so much as AI designers trained to incorporate ethics in their process. But, he hastens to add, AI design is a new and specialist practice that doesn't necessarily fit hand-in-glove with a traditional software development process. It's vital that there is a priority put on developing techniques and practices that extend traditional roles—whether as engineer, program manager, designer, or researcher.

Ultimately, the role of the ethical AI designer is to help people work through the complex, sometimes ambiguous, decisions that are required to design, build, test, and manage an ever-changing intelligent product.

The checklist

AI ethics isn't all high-minded discussion of existential conflicts, but abstract principles can be difficult to put into practice. Principles can mask the complexity of ethical decisions, which involve different assumptions, interpretations, personal experiences and biases.

This makes it extraordinarily difficult for people to apply principles consistently in the countless small decisions made everyday in the software development process. Checklists can help with a middle ground, providing a scaffold between high-level principles and granular technical tools.

Researchers have found that the most important role of a checklist in AI ethics is to prompt critical conversations. AI ethics efforts are often the result of ad-hoc processes driven by passionate individual advocates. In companies where the priority is fast-paced development and deployment, there

can be significant social cost by slowing things down to talk about fairness.

It takes time and effort to involve more people in decisions and to wrestle with competing or ambiguous topics. An AI ethics checklist can act as a "value lever" and make it acceptable to reflect on risks, raise red flags, add extra work and escalate decisions.

Do ethicists matter?

How are AI ethicists in technology actually doing their jobs? And does their presence make a difference?

Data and Society researchers recently investigated this by interviewing "ethics owners" from various Silicon Valley tech companies—people employed to "do ethics" for big tech—for their views on the practical progress of ethics in Silicon Valley.

The researchers found a central dilemma: ethics owners have to try to resolve complex social decisions, which are usually framed (and come under challenge) within the logic of Silicon Valley. One of the most prevailing Silicon Valley mindsets is technology solutionism—the idea that the

solution to a “bad” technology outcome is more technology. Because AI is used in social applications, this philosophy is being applied to more and more social problems.

What makes this a unique challenge for ethicists is that these companies can get really big before they are mature. “Ethics owners” become “ethics coordinators,” tasked with making sure ethics doesn’t get in the way of the engineers’ work.

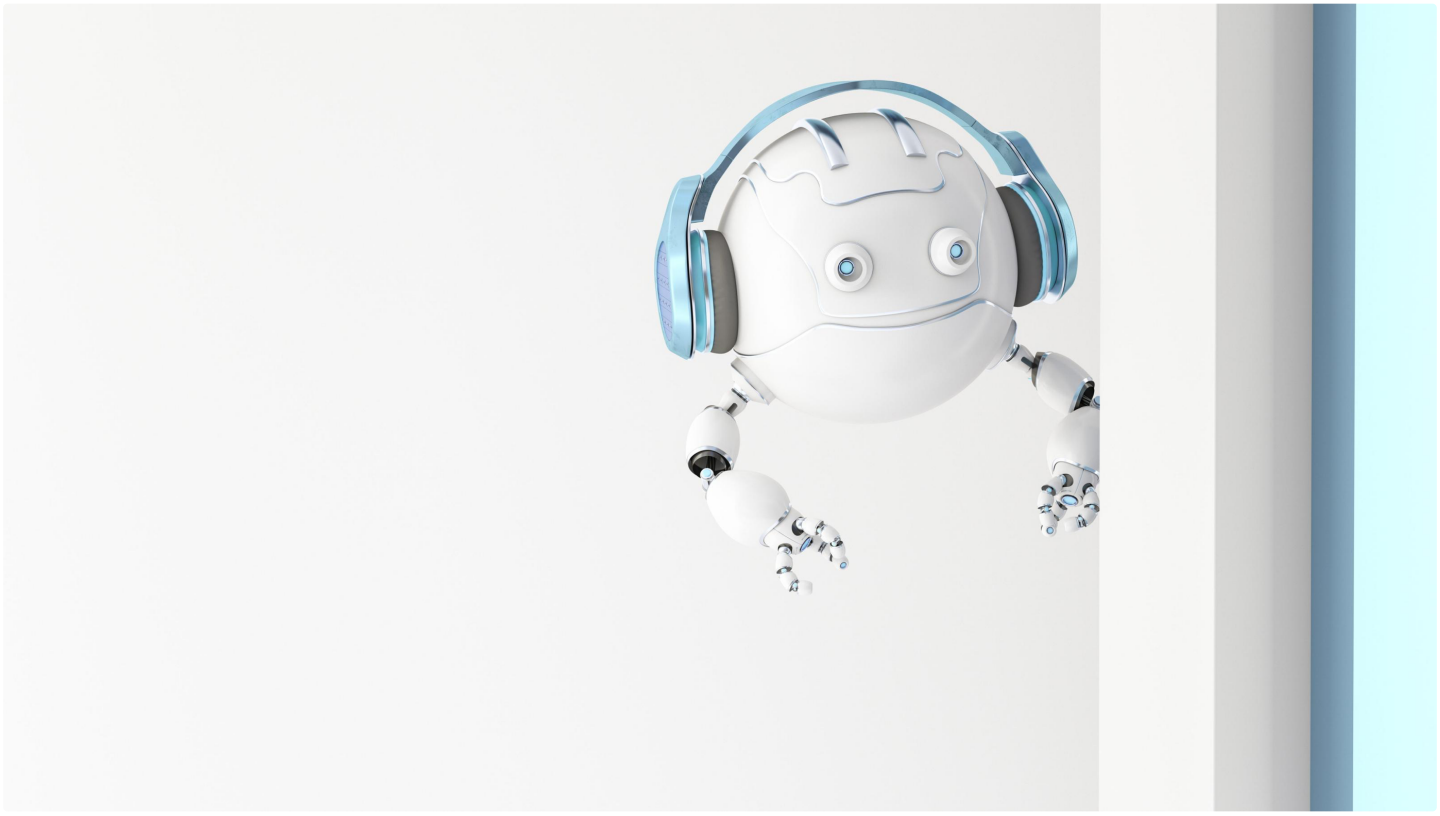
Engineers who make the frontline decisions can position themselves as being able to use their personal judgment. This is one of the primary ways moral judgment gets instantiated inside AI. If engineers are the ones seen to be best positioned to evaluate a hypothetical harm, they also have the power to dismiss the concern as not realistic, not relevant, or not worth bothering about given the probabilities. If engineers lack knowledge of the long term or broader societal consequences, they can lack a sense of accountability for what they do as individuals.

For the ethicists, there is often a disconnect between the technologist’s mind and the moral significance of their work.

The challenge for AI ethicists is to embed values of dignity, agency, and human flourishing in AI design so that these values can be at work up and down the entire development stack and throughout the design and development process.

This will mean not only dealing with biased training data but dealing with how the AI ultimately behaves "in the wild."

The challenge for companies, according to Bietti, is giving their employees space to "think slower, to think with more depth, and more systematically." Developers need to shield their "thinking from pragmatic pressures," let intuitions change and enable people to make sense of other problems. The traditional Silicon Valley approach to "move fast and break things" will not suffice.



For your listening pleasure.

TOOLKIT

The people, podcasts, and papers to check out on AI bias

✦ Member exclusive by Helen Edwards for AI's power problem

AI bias is a rapidly changing field. Below you'll find the

people to follow to stay on top of it, along with the books, papers, podcasts, and other resources you need to get up to speed.

Podcasts

The Machine Ethics Podcast

Innovation for All (particularly this episode and this one)

Exponential View (particularly this episode)

BBC radio on the AI ethics challenge

Your Undivided Attention, from the Center for Humane Technology

People to follow

Josh Lovejoy, head of design, Microsoft Cloud and AI

Meredith Whittaker, AI Now

Dorothea Baur, consultant

Arvind Narayanan, Princeton University

Cathy O’Neil, algorithmic auditor

Safiya Umoja Noble, UCLA

Joanna Bryson, University of Bath

Annette Zimmerman, Princeton University

Kate Crawford, AI Now

Jacob Metcalf, Data & Society

Jason Schultz, NYU, AI Now

John C. Havens, IEEE

Adam Cutler, IBM

Cassie Kozyrkov, Google

Joy Buolamwini, Founder, Algorithmic Justice League, MIT

Maria Axente, PwC UK

Chelsea Barabas, MIT

Michael Kearns, University of Pennsylvania

Papers to read

Discriminating Systems: Gender, Race, and Power in AI, AI Now

Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies, Harvard Journal of Law and Technology

The Intuitive Appeal of Explainable Machines, Fordham Law Review

Proxy Discrimination in the Age of Artificial Intelligence and Big Data, Iowa Law Review

Studying Up: Reorienting the study of algorithmic fairness around issues of power, Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency

From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy, Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency

Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics, Data & Society

Online tools and games

Closing Gaps Ideation Game from the Partnership on AI

Survival Of the Best Fit: a game on bias in hiring

Stealing Ur Feelings interactive documentary from Mozilla

Books

The Ethical Algorithm, Michael Kearns and Aaron Roth.

Algorithms to Live By, Brian Christian and Tom Griffiths

Algorithms of Oppression: How search engines reinforce racism, Safia Umoja Noble

Weapons of Math Destruction: How big data increases inequality and threatens democracy, Cathy O'Neil

Biased: Uncovering the hidden prejudice that shapes what we see, think, and do, Jennifer L. Eberhardt

Other resources

Algorithmic Impact Assessments: a practical framework for public agency accountability, AI Now

Salesforce AI ethics blog resource list

Google's People+AI Guidebook

Microsoft on human-centered AI

World Economic Forum guide for boards on AI Ethics

Machines gone Wrong: a guide for AI practitioners

Gender Shades video from Joy Buolamwini of MIT