AI ETHICS - ESSAY

The Bias We're Building In

How AI systems inherit and amplify the inequalities of the world they learn from -- and why fixing that requires more than better data.

Mandy Hathaway - MA, Ethical Technology & Artificial Intelligence, Metropolitan State University

Adapted from graduate thesis research, Metropolitan State University, 2022.

The Illusion of Objectivity

There is a seductive idea at the heart of how we talk about artificial intelligence: that machines, unlike humans, are neutral. That statistics don't have feelings, and therefore algorithms don't have biases. It's a compelling story. It's also wrong.

Every machine learning system learns from data. And data is not neutral ground -- it is a record of the world as it has been, not the world as it should be. When that world is shot through with racial bias, gender inequity, and economic injustice, the data reflects that. Train a model on it, and the model learns it too. This is not a flaw in the technology. It is a mirror.

This isn't a theoretical concern. It is happening in courtrooms, hospitals, universities, and hiring offices right now. And in most of those cases, the people most harmed have no idea an algorithm was involved -- let alone that it was working from data that was never fair to begin with.

Real People, Real Harm

Consider COMPAS, the risk-assessment algorithm currently used in criminal sentencing across New York, Wisconsin, California, Florida, and other states. Its stated purpose is to predict whether a defendant is likely to reoffend -- a seemingly rational tool for informing sentencing decisions. A 2016 ProPublica investigation found it was 77 percent more likely to flag a Black defendant as high risk for violent reoffending than a white defendant with a comparable record.

The reason is not mysterious once you look at what the system learned from. Black and white Americans use marijuana at approximately the same rates. Black Americans are arrested for it at more than three times the rate of white Americans. When one group is over-policed, you end up with higher crime statistics for that group -- not because they commit more crime, but because the system was already watching them more closely. Feed that history into a machine learning model and you don't get a neutral predictor. You get a mathematical formalization of that disparity, stripped of its context, and handed to a judge as an objective score.

When defendants have challenged these scores, courts have ruled that the underlying algorithm is proprietary and does not need to be disclosed. The bias is real. The harm is measurable in years behind bars. But the mechanism stays invisible -- hidden behind a number and the false authority that number implies. Hiding bias in an algorithm does not eliminate it. It just adds a layer of deniability.

The Problem Isn't Just the Data

It would be convenient if algorithmic bias were simply a data-cleaning problem. Better data in, fairer results out. But the issue runs deeper, and understanding why matters.

Machine learning systems generate their own algorithms from the data they are given. Even when designers carefully remove sensitive variables -- race, gender, zip code -- the model can still learn to use proxies that correlate with those variables in the existing data. Removing the label doesn't remove the pattern. The bias doesn't disappear; it finds a side door.

The Word2Vec project illustrates this clearly. Designed to model statistical relationships between words in the English language, it was trained on a vast collection of human-generated text. What it learned faithfully was that 'man' relates to 'doctor' the same way 'woman' relates to 'nurse.' That the concepts of 'corporate' and 'jewish' appeared in proximity. The data wasn't labeled with those associations. The model found them on its own -- because they were already embedded in the language humans have actually produced.

When designers tried to correct for this by reducing the weight of gendered terms, the model began to believe things could be 'grandmothered in' -- a phrase carrying gendered connotations the fix hadn't caught. Bias in language, like bias in society, doesn't sit in one clearly labeled drawer waiting to be removed. It is woven through the structure of the whole thing.

Outsourcing the Hard Questions

Part of what makes algorithmic bias so persistent is that it aligns with something we are already inclined to do: hand difficult judgment calls to systems that appear more objective than we are. If a computer produced the number, it must be more reliable than a biased human judge. This logic has a critical flaw.

A rational system is not necessarily an ethical one. This is not a subtle distinction -- it is a non-sequitur to assume otherwise. A rational decision may at some point be to let the poor starve to death in order to meet budget requirements, depending on what the system has been designed to optimize for. Most people would recognize that as monstrous. But the algorithm won't.

The U.S. military still requires human authorization for all remote weapons strikes, even where the technology for fully autonomous engagement exists. There is a recognition embedded in that policy that some decisions carry enough moral weight that a human being has to own them. Criminal sentencing -- with its life-altering, often irreversible consequences -- would seem to meet that bar. And yet courts are treating COMPAS scores as routine inputs, and judges are using them as shortcuts. The algorithm's own manufacturer states that staff should 'use their professional judgment' to override it as appropriate. They know the scores are imperfect. The overrides rarely happen.

If a judge declared from the bench, 'you are Black, so I am sending you to jail,' there are legal remedies available, however imperfect. When a system trained on biased data reaches that same conclusion and delivers it as a risk score, there is no recourse. The bias isn't gone. It's just been laundered through an algorithm and handed back to us with a veneer of objectivity.

The Feedback Loop We're Building

There is a long-term dimension to this that rarely gets the attention it deserves. Bias in today's systems doesn't just harm people today -- it generates the data that trains tomorrow's systems. When a biased predictive policing model sends more officers to Black neighborhoods, more arrests happen in those neighborhoods. That data then feeds back into the next model, and the cycle tightens. Each iteration cements the inequity a little deeper into the architecture.

The same dynamic plays out in hiring algorithms, college admissions tools, and social media recommendation engines. Each produces decisions that shape behavior, which generates data, which trains the next model. We are not just automating the present. We are encoding it into the infrastructure of the future.

This is why the stakes of getting this right now are so much higher than they might appear. A flawed system deployed at scale today can become the training foundation for systems a decade from now. Bias, once digitized and normalized, is extraordinarily hard to root out. We are building the world our algorithms will inherit.

What Needs to Happen

None of this means AI is beyond repair. Promising technical tools exist -- adversarial auditing systems, differential privacy frameworks, study pre-registration practices for data scientists. The EU's data protection legislation offers a model for what meaningful legal intervention can look like. Researchers are building systems specifically designed to detect and flag their own biases. These things matter.

But technical fixes alone will not solve a social problem. The deeper issue is that AI development has been, by and large, designed by and for a narrow demographic slice of humanity. When the team building the system doesn't reflect the breadth of people the system will affect, crucial considerations go unexamined -- not out of malice, but out of the blind spots that homogeneity produces. Those blind spots have real consequences.

Until 2011, every crash test dummy used to assess vehicle safety was modeled on the proportions of an average adult white male. The failure wasn't cruelty -- it was the simple fact that women weren't in the room where those decisions were made, so no one thought to ask. The parallel to AI development is not subtle, and the consequences are not abstract.

Ethicists, social scientists, and communities most affected by these systems need seats at the table -- not as consultants to be overruled when their findings conflict with a product roadmap, but as full participants in design from the beginning. Google's 2020 dismissal of AI researcher Timnit Gebru, after she raised concerns about bias in the company's own large language models, is a useful indicator of how far the industry still has to go. When your ethical AI team is fired for doing ethics, the problem is structural.

The Choice We're Making

AI systems don't have values. They have objectives, data, and the statistical patterns they have been trained to find. The values are ours. And right now, we are making choices -- about what data to collect, what to optimize for, who to include in the design process, and how much transparency and accountability to demand -- that will shape these systems for decades.

Allowing AI to perpetuate the behaviors that have resulted in our current social inequities is an indefensible position if the stated goal is to use technology to improve society. We can build systems that reflect the world as it is, bias and all, and call it objectivity. Or we can build systems that take seriously the world as it should be -- and do the unglamorous, difficult, necessary work of holding ourselves accountable to that goal.

The technology is not the obstacle. The will is. And the power to get this right is still within our grasp -- but only if we reach for it.

--

Mandy Hathaway *holds an MA in Ethical Technology & Artificial Intelligence and a BA in Philosophy from Metropolitan State University. She works at the intersection of AI systems, technical communication, and ethics. This essay is adapted from her 2022 graduate thesis, 'An Accessible Introduction to Artificial Intelligence and the Associated Ethical Risks.'*