# STOP BUILDING BAD AI

*Annette Zimmermann*

AN AI-POWERED "facial assessment tool" compares your face to supposedly "objective" standards of beauty and offers an "aesthetics report" with recommendations for cosmetic surgery. Amazon's new Halo health band aspires to recognize emotions and warns women who wear it when their voice sounds too "dismissive" or "condescending." A tool used by Stanford University researchers uses facial recognition technology to predict whether you are gay.

Should these technologies exist? Whether AI can make accurate predictions in these areas is far from clear. But beyond this technical issue, we ought to ask whether we need such tools to begin with. Are the problems they set out to solve worth solving? How does predicting someone's sexual orientation, possibly without their knowledge and against their will, make the world better, or more just? What harms might result from the use of such a tool? We should ask questions about the goals and likely consequences of a particular technology before asking whether it could be made to work well. And when we do so, we need to be open to the possibility that some AI tools should not be built in the first place.

Unfortunately, these questions are not asked often enough about AI. One reason is economic: especially in the absence of robust legal regulation, ethical reflection takes a back seat to the profit motive. Another is cultural: an ongoing wave of renewed AI optimism, following the AI "winters" of the late 1970s and early '90s, often crowds out concerns about its potential harms. Then there is AI exceptionalism, the conceit that AI development is too important or distinctive to be stifled and thus should be exempt from the caution and regulation we apply to other technological innovations. Still another reason is philosophical: the assumption is that AI goes wrong only when it relies on biased data or when it fails to perform well.

Certainly AI can help us perform many important and complex tasks that humans cannot accomplish at the same scale and speed. Many AI projects are worth pursuing, and many developers have good intentions. But that does not license a general norm in favor of building and deploying any AI tool for any purpose, regardless of the social and political context in which it operates. Indeed, there are important reasons why we ought to challenge this presumption in some cases. A just future for AI demands that we think not just about profit or performance, but above all about purpose.

IN PRINCIPLE, there are two basic strategies we might pursue in order to mitigate the harms of a certain technology. On the one hand, we might try to optimize it, with the aim of making it more accurate, fairer, more transparent—better at doing what it is supposed to do. On the other hand, we might refuse to deploy or build it altogether—especially if we judge its goals or likely consequences to be ethically indefensible.

A powerful current within contemporary culture favors the former strategy. After all, who could object to making things better? In this view, there are many mechanisms available for improving flawed AI. We can alter algorithmic decision rules and improve datasets by making them more fine-grained and representative. We can better measure and operationalize key concepts relevant to the given task. We can test AI systems by simulating what would happen if we were to deploy them, and we can deploy them in relatively controlled, constrained ways before implementing them at scale—for instance, in sandboxed projects carried out by academic and industry research teams.

But it is important that we recognize this is not our only option. For tools that have already been deployed, we might choose to stop using them. Recent bans of law enforcement facial recognition tools in several U.S. cities illustrate this approach in action. San Francisco's recent ordinance concerning acquisitions of surveillance technology, for instance, argues as follows: "The propensity for facial recognition technology to endanger civil rights and civil liberties substantially outweighs its purported benefits, and the technology will exacerbate racial injustice and threaten our ability to live free of continuous government monitoring." Even private corporations agreed that non-deployment was the best solution in this case: Amazon, Microsoft, and IBM all voluntarily adopted non-deployment moratoria until facial recognition in policing is subject to top-down regulation. These moves may be motivated more by financial interest—the desire to avoid the costs of PR fallouts—than by ethical commitments. Still, it is noteworthy that even the industry's largest corporations have publicly advocated for non-deployment of technology that has already been built.

Non-deployment efforts in this area have been prompted by influential studies showing that currently used facial recognition systems are highly inaccurate for women and people of color. This is a good reason not to deploy these systems for now, but it is also important to recognize that the unregulated use of such systems might well be politically and morally objectionable even if those tools could be made highly accurate for everyone. Tools that support and accelerate the smooth functioning of ordinary policing practices do not seem to be the best we can do in our pursuit of social justice. In fact, the use and continued optimization of such tools may actively undermine social justice if they operate in a social setting that is itself systemically unjust.

There is one further option, of course. Carrying this logic even further back in the development process, we might decide not just to avoid *deploying* certain AI tools but to avoid building them altogether.

WHICH OF THESE STRATEGIES—optimize, do not deploy, or do not build in the first place—is best? It is impossible to say in general. Whether a particular AI tool warrants development and deployment will depend heavily on a large number of empirical factors: how the tool works, which problem it is tasked with solving, how the technology interacts with social structures already in place. These kinds of facts about the social world are subject to change. Political and institutional transformations may alter the way people are situated socially; evolving norms will affect the way people and institutions interact with technology; technology itself will dynamically reshape the society it is a part of. We thus should not hope for a generic ethical rule, a blanket endorsement one way or another.

Instead, we must develop nuanced, context-specific frameworks for thinking through these issues. This work will entail taking on several obstacles to more robust ethical and political reflection on the strategies at our disposal.

One is the cultural imperative, especially popular in the tech world, to move fast and break things—Facebook's infamous motto until 2014. Former Yahoo! CEO Marissa Mayer is often quoted as saying that "with data collection, 'the sooner the better' is always the best answer." Amazon's leadership principles feature similar language: "Speed matters in business. Many decisions and actions are reversible and do not need extensive study. We value calculated risk taking." In an environment that prioritizes speed above all else, technologists are less likely to ask why or whether a certain technology ought to be built than to think, *why not?*

At the same time, many practitioners are increasingly concerned about—and actively working to mitigate—the harms of AI.



Most major tech companies now have designated teams focusing on "ethical," "trustworthy," or "responsible" AI. But it is unclear whether corporations will empower such teams to intervene in the development and design of new technology. Google's recent firing of Timnit Gebru and Margaret Mitchell, co-leads of the company's Ethical AI team, shows that industry AI ethics efforts are often limited and outweighed by competing corporate goals.

Tech employees, for their part, are also increasingly organizing themselves—often against significant pushback—with the aim of holding their employers accountable. Consider the Alphabet Workers Union. "We will use our reclaimed power to control what we work on and how it is used," its mission statement reads. "We are responsible for the technology that we bring into the world, and recognize that its implications reach far beyond Alphabet." Such statements may be compatible with refusing to build or deploy new technology, but they typically lean heavily toward optimization—specifically, optimization within powerful corporations. "We will work with those affected by our technology to ensure that it serves the public good," the statement continues. "Alphabet can make money without doing evil," it says elsewhere on its website. But whether such justice-oriented optimization is compatible with the pursuit of profit—within a small number of powerful private corporations, to boot—remains to be seen.

A second obstacle we must reckon with is the contention that developing a potentially harmful technology is better than leaving it to bad actors. Many technologists reason, for example, that if *their* team does not build a given tool, someone else will—possibly with more sinister motives. On this view, arguments not

to build or deploy may look like giving up, or even a way of making things worse. The Stanford researcher who used facial recognition technology for predicting sexual orientation, for example, argued that it would have been "morally wrong" not to publish his work:

> This is the inherent paradox of warning people against potentially dangerous technology. . . . I stumbled upon those results, and I was actually close to putting them in a drawer and not publishing—because I had a very good life without this paper being out. But then a colleague asked me if I would be able to look myself in the mirror if, one day, a company or a government deployed a similar technique to hurt people.

But this argument does not stand up to scrutiny. Nothing prevents a bad actor from repurposing knowledge and technological capabilities gained from an AI tool first developed by a well-intentioned researcher, of course. And even tools developed with good intentions can ultimately have damaging effects.

A third obstacle is a too limited conception of the ways AI can be harmful or unjust. In many familiar examples of algorithmic injustice, accuracy is distributed unequally across different demographic groups. Criminal recidivism risk prediction tools, for instance, have been shown to have significantly higher false positive rates for Black defendants than for white defendants. Such examples have elicited significant ethical reflection and controversy, helping to call attention to the risks of AI. But we must also recognize that AI tools can be unjust even if they do not rely on biased training data or suffer from disparate distributions of error rates across demographic groups.



For one thing, even if developers are well intentioned, the *consequences* of implementing a particular algorithmic solution in a specific social context may be unjust, because algorithmic outputs reflect and exacerbate social biases and inequalities. It may also be that the *goal* of an AI tool is simply not just to begin with, regardless—or even indeed because of—the tool's accuracy. Consider Megvii, a Chinese company that used its facial recognition technology in collaboration with Huawei, the tech giant, to test a "Uighur alarm" tool designed to recognize the faces of members of the Uighur minority and alert the police. Here it is the very goal of the technology that fails to be morally legitimate. A related problem is that human decision-makers, prone to automation bias, may fail to scrutinize algorithmic classifications, taking ostensibly neutral and objective algorithmic outputs as a given instead of interrogating them critically. In still other cases, it may be the *logic* of the technology that is objectionable, leading to what philosophers call "expressive harm": the use of particular categories and classifications in AI tools can convey a demeaning, harmful message, which becomes unjust in light of prevalent social norms, assumptions, and experiences. Tools that attempt to deduce sexual orientation or other personality traits from one's physical appearance, for example, may contribute to reinforcing the harmful message not only that it is possible to "look like a criminal," or to "look gay," but also that it is valid to infer personal characteristics and future behavior from the way a person looks. The upshot of these various examples is that the potential harms of AI range far beyond datasets and error rates.

A final obstacle to more robust ethical reflection on AI development is the presumption that we always have the option of

non-deployment. If at some point in the future it turns out that an AI tool is having unacceptably bad consequences, some might say, we can simply decide to stop using the tool *then*.

This may be true in some cases, but it is not clear why we should think it is always possible—especially without industry-wide regulation. The labor effects of automation, for example, may well be effectively irreversible. In current market conditions, it is hard to imagine how a company could take back its decision to replace a human-executed task with an AI-driven, automated process. Should the company face backlash over its AI tool, current incentives make it far likelier that it would seek to find another way to automate the task rather than rehire humans to execute it. The pressure to automate is now so strong in some sectors that some companies are *pretending* to have built and deployed AI. In 2016, for example, Bloomberg News reported that personal assistant startup X.ai was directing employees to simulate the work of AI chatbots, performing avalanches of mind-numbing, repetitive tasks such as generating auto-reply emails and scheduling appointments. It would be naïve to think that once such tools are actually built and deployed, the work force could easily revert to its pre-automated structure.

For another example of the limits of non-deployment, consider DukeMTMC, a dataset of two million video frames recorded in public spaces on Duke University's campus and made publicly available without protecting the identities of the people included in the videos. The data wound up being used for controversial research on computer vision-based surveillance technology, and it was taken down in June 2019 after significant public criticism. But as Princeton University researchers recently pointed out, at least 135 research papers utilized that dataset—and others derived from it—*after* it had been taken down. Non-deployment thus did not make the ethical and political risks associated with this technology disappear.



FOR ALL OF THESE REASONS, we must take the option not to build far more seriously than we do now. Doing so would not only help to make the development of AI more just. It would also lead us to reflect more deeply on the demands of justice more generally.

Return to the example of facial recognition tools used in law enforcement. Rather than trying to scale up and optimize existing policing practices by augmenting them via AI, we could instead ask: What would more just law enforcement look like? Which institutional, economic, and legal transformations are needed for this purpose? The answers to these kinds of questions may not necessarily involve AI—at least not before other sociopolitical changes are made first.

Making these judgments—deciding whether a particular AI system should be built and optimized, or not built or deployed at all—is a task for all of us. One-off non-deployment victories and shifting industry norms are important achievements, but they are not enough. We need systematic regulation and democratic oversight over AI development. We need new frameworks for both national and international governance on these issues. And we need meaningful opportunities to deliberate collectively about whether powerful new forms of technology promote, rather than undermine, social justice.

When asking these kinds of questions, we must resist the tendency to view AI in isolation from the larger history of technological development. Instead we should look for important parallels with the development and regulation of other powerful technologies, from nuclear weapons to gene editing.

As science and technology studies scholar Sheila Jasanoff observes in her 2016 book *The Ethics of Invention*, these urgent forms of engagement will require vigilant public action and continual democratic scrutiny. "Important perspectives that might favor caution or precaution," she notes, "tend to be shunted aside in what feels at times like a heedless rush toward the new." However, history shows that when it comes to technological development, the new is not always just. Getting clear on the purpose and value of AI is more important than the rush to build it or make it better.