HARVARD Kennedy School
CARR CENTER
for Human Rights Policy

FALL 2021

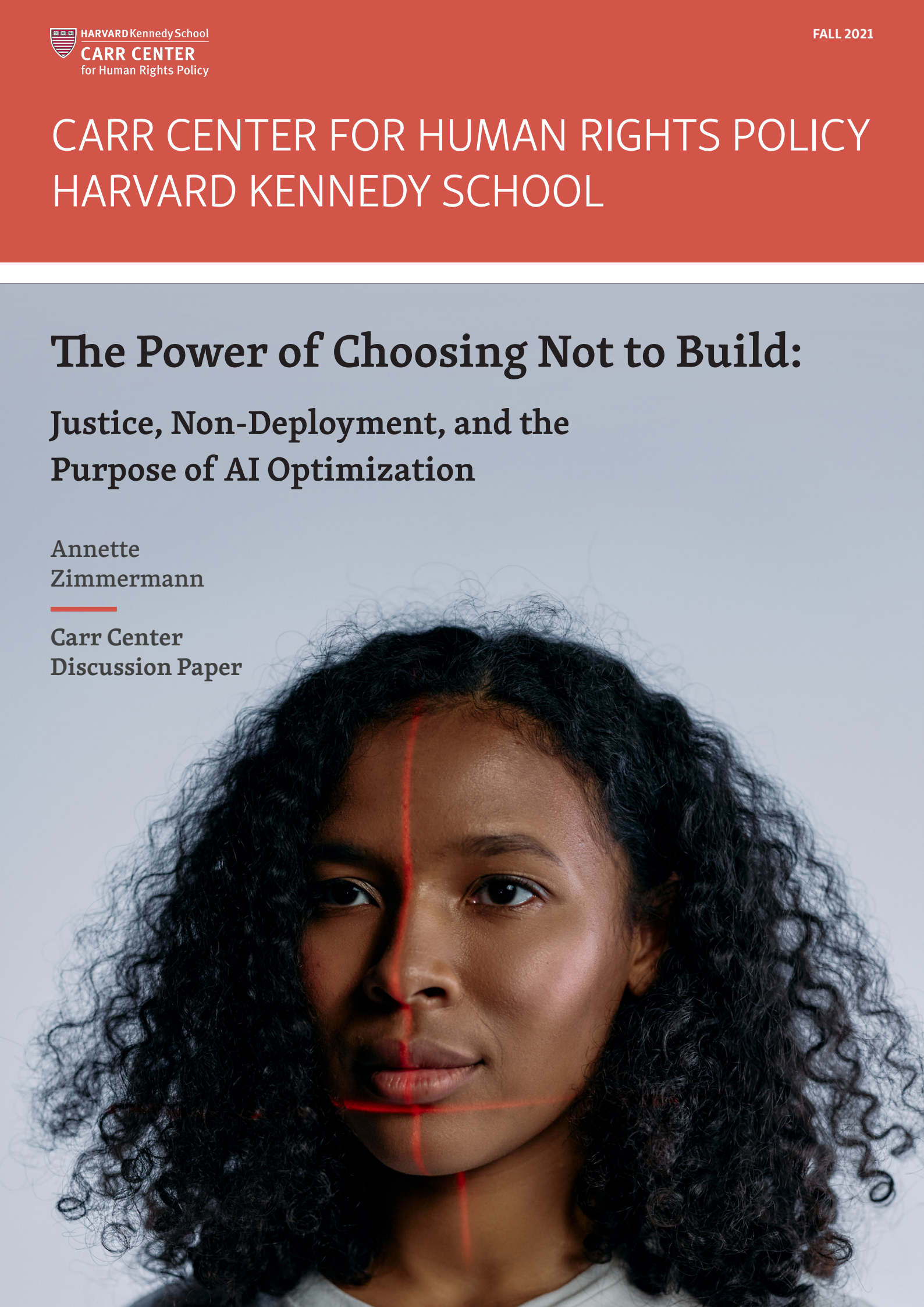
CARR CENTER FOR HUMAN RIGHTS POLICY
HARVARD KENNEDY SCHOOL

The Power of Choosing Not to Build:

Justice, Non-Deployment, and the Purpose of AI Optimization

Annette Zimmermann

Carr Center Discussion Paper



The Power of Choosing Not to Build: Justice, Non-Deployment, and the Purpose of AI Optimization

Carr Center for Human Rights Policy
Harvard Kennedy School, Harvard University
October 5, 2021

Annette Zimmermann
Technology and Human Rights Fellow
Carr Center for Human Rights Policy

Table of Contents

ABSTRACT: Are there any types of AI that should never be built in the first place? The "Non-Deployment Argument"—the claim that some forms of AI should never be deployed, or even built—has been subject to significant controversy recently: non-deployment skeptics fear that it will stifle innovation, and argue that the continued deployment and incremental optimization of AI tools will ultimately benefit everyone in society. However, there are good reasons to subject the view that we should *always* try to build, deploy, and gradually optimize new AI tools to critical scrutiny: in the context of AI, *making things better* is not always *good enough*. In specific cases, there are overriding ethical and political reasons—such as the ongoing presence of entrenched structures of social injustice—why we ought *not* to continue to build, deploy, and optimize particular AI tools for particular tasks. Instead of defaulting to optimization, we have a moral and political duty to critically interrogate and contest the *value and purpose* of using AI in a given domain in the first place.

Consider the following cluster of recently developed AI tools:

(a) *Objective Beauty*: An AI-powered "facial beauty assessment" tool compares users' faces to a supposedly "objective" standard of beauty and offers recommendations for cosmetic surgery.[1] Similar beauty assessment tools have been shown to reinforce racial inequality by entrenching Eurocentric beauty standards that favor whiteness.[2]

(b) *Tone Evaluation*: Amazon's Halo health band claims to be able to evaluate emotions. Female users report receiving feedback from the health band warning them that their tone sounds "condescending" and "dismissive."[3] Given that there exists a set of negative socially salient stereotypes concerning the tone of women's speech, tone feedback of this kind can reinforce those stereotypes because users evaluate such feedback by reference to the wider social *context* in which it is given.

(c) *Gay Facial Recognition*: A facial recognition system is used in order to predict the sexual orientation of a person.[4]

All of these applications of AI raise a fundamental philosophical question: should these tools exist, even if they work reasonably well? This is a controversial issue. Many people—not just technology industry practitioners—think that there are strong reasons not to stifle the development of innovative, powerful forms of AI and machine learning. After all, AI can help us perform many important and complex tasks that humans simply cannot accomplish at the same scale and speed. This much is true: many forms of AI have the potential to be socially beneficial, and not all forms of AI are *inherently* objectionable. In short, there is a strong *general presumption* in favor of continuing to build, deploy, and optimize AI.

Nevertheless, there are important and underappreciated reasons why we ought to question this presumption in several cases. It is not enough to ask whether a given tool is accurate enough, whether it does what it is supposed to do. We also ought to ask a more fundamental question: what is the *purpose* of building a given AI tool—and is it a *good* purpose? Could the pursuit of some purposes simply undermine justice, irrespective of how well AI tools work? And if the goals underpinning a given AI tool—the definition of the decision task itself—are unjust, should such a tool be built at all, no matter how much we can optimize it?

This discussion paper is an extended version of Annette Zimmermann, "Stop Building Bad AI," in *Redesigning AI: Work, Democracy, and Justice in the Age of Automation*, ed. Daron Acemoglu (Boston: Boston Review, Spring 2021). I am indebted to Matt Lord, Joshua Cohen, Sushma Raman, Mathias Risse, and the 2020–21 Technology and Human Rights Fellows of the Carr Center for Human Rights Policy at Harvard University for valuable comments and discussion.

[1] "Facial Assessment Tool," QOVES Studio, accessed August 15, 2021, https://www.qoves.com/facial-assessment-tool/.

[2] For discussion of racial bias in AI-driven beauty assessments—specifically, a tool called beauty.ai—see Ashraf Khalil et al., "Investigating Bias in Facial Analysis Systems: A Systematic Review," *IEEE Access* 8 (June 30, 2020): 130751–61. In response to this problem, QOVES insists on its company website that "we've tried to design our backend algorithm to be as inclusive as possible. Granted, our tool currently only looks for the most basic of facial flaws (superficial skin flaws) which apply to everyone to some degree, we're working on implementing more race-specific analysis, as everyone is different." See QOVES, "Facial Assessment Tool."

[3] Geoffrey A. Fowler and Heather Kelly, "Amazon's new health band is the most invasive tech we've ever tested," *Washington Post*, December 10, 2020, https://www.washingtonpost.com/technology/2020/12/10/amazon-halo-band-review/.

[4] Paul Lewis, " 'I was shocked it was so easy': meet the professor who says facial recognition can tell if you're gay," *The Guardian*, July 7, 2018, https://www.theguardian.com/technology/2018/jul/07/artificial-intelligence-can-tell-your-sexuality-politics-surveillance-paul-lewis.

Mistakes and the Magnifying Glass

Of course, *accuracy* does matter significantly for the question of whether the use of AI in a given domain is just. In many paradigmatic examples of "algorithmic injustice,"[5] accuracy is distributed *unequally* across different socio-demographic groups, which is precisely what renders the outputs of many algorithmic systems unjust.

Though it may seem like AI tools simply, and quite predictably, *reflect* our social world as it is—including its many injustices—it is more accurate to say that AI, if left unchecked, risks *amplifying* and simultaneously *obfuscating* those injustices further. Data is not just a "social mirror"[6]—it is a magnifying glass. Some tools fail differently for different groups of people: algorithmic criminal recidivism risk-prediction tools, for instance, have been shown to fail differently for Black defendants than for white defendants,[7] thus exacerbating existing structures of racial injustice in the criminal justice system. Of course, humans are highly biased, too: much like algorithms, judges may well inaccurately overestimate the likelihood of Black defendants recidivating in the future. But even if algorithms are somewhat less biased than human judges, this does not guarantee just outcomes: humans are significantly less likely to question decision outcomes generated by an automated, quantifiable, and thus seemingly neutral process. This is dangerous: if human decision-makers fail to critically scrutinize algorithmic outputs, any unjust patterns reflected in those outputs will remain, and potentially increase in magnitude.

That said, the problem of injustice-exacerbating AI is not confined to the rather narrow—though important—issue of disparate distributions of error rates in algorithmic systems, coupled with the problem that humans are likely to opt out of the onerous task of reflecting carefully and critically on whether a given algorithmic output is indeed warranted.

Some AI tools work fairly well across the board, by the lights of their own success conditions: inaccuracy as such is not the central problem. But their very use may be morally unjustifiable and politically harmful if they rely on methods of inference that echo unjust historical practices. In such cases, building and using such tools itself constitutes what philosophers call "expressive harm": communicating a demeaning, harmful message, which takes on an unjust social meaning in light of prevalent historical and social norms, assumptions, and experiences.[8]

On Purpose

Consider again the facial recognition tool mentioned above: the one that, according to the team of Stanford researchers who built the tool in 2017, predicted many behaviors, preferences, and characteristics more accurately than humans: sensitive details like a person's sexual orientation, or their propensity for future violent criminality.[9] While humans accurately predicted those features in 61% of cases for facial data from men and 54% of cases for women, the researchers claimed that their tool performed accurately in 91% of cases for men and 83% of cases for women.

The crux is that the tool did this purely by analyzing *images* of people's faces: by measuring the distance between a person's facial features, and by evaluating data about "gender-atypical grooming styles." This tool was met with significant academic and public backlash:[10] making *this particular kind* of inference—an inference from blunt facial measurements to a prediction of a person's characteristics—comes with significant historical baggage. In particular, critics asked, did the method of inferring these characteristics purely from the way a person looks not echo discredited, racist,

[5] Annette Zimmermann, Hochan Sonny Kim, and Elena Di Rosa, "Technology Can't Fix Algorithmic Injustice," *Boston Review*, January 9, 2020, http://bostonreview.net/science-nature-politics/annette-zimmermann-elena-di-rosa-hochan-kim-technology-cant-fix-algorithmic.

[6] Solon Barocas and Andrew S. Selbst, "Big Data's Disparate Impact," *California Law Review* 104 (2016): 671.

[7] Julia Angwin et al., "Machine Bias," *ProPublica*, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[8] Important contributions to this philosophical and legal literature include Catharine A. MacKinnon, *Only Words* (Cambridge, MA: Harvard University Press, 1993); Kimberlé Crenshaw et al., eds., *Words That Wound: Critical Race Theory, Assaultive Speech and the First Amendment* (Boulder: Westview Press, 1993); Elizabeth S. Anderson and Richard H. Pildes, "Expressive Theories of Law: A General Restatement," *University of Pennsylvania Law Review* 148, no. 5 (2000): 1503; Simon Blackburn, "Group Minds and Expressive Harm," *Maryland Law Review* 60, no. 3 (2001); Tarunabh Khaitan, "Dignity as an Expressive Norm: Neither Vacuous Nor a Panacea," *Oxford Journal of Legal Studies* 32, no. 1 (2012): 1–19.

[9] Yilun Wang and Michal Kosinski, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images," *Journal of Personality and Social Psychology* 114, no. 2 (2018): 246–57.

[10] Sahil Chinoy, "The Racist History Behind Facial Recognition," *New York Times*, July 10, 2019, https://www.nytimes.com/2019/07/10/opinion/facial-recognition-race.html.

pseudoscientific methods and ideas like phrenology and physiognomy? Attempts to map out photographs of faces, and to make far-reaching claims about associated "desirable" and "undesirable" personal characteristics on that basis, were favored by eugenicists like Francis Galton. Galton tried (and failed) to use superimposed portraits of convicted criminals, mentally ill people, and Jews to create "composite images" of various "types" of people, before moving on to what he considered the "healthy and talented" classes—Anglican ministers, Westminster schoolboys, Royal Engineers.[11] Using a similar type of inferential method risks communicating and reinforcing the same harmful ideas propagated by Galton and others: that it is not only *possible* to "look like a criminal," or to "look gay," but also that it is *true* that the way a person looks is an apt, accurate basis from which to make inferential judgments about their personal characteristics and future behaviors.

The important question, then, is not *whether* AI can accurately predict if you are gay. The question is *why* we *need* AI to predict if you are gay to begin with. How, if at all, does having a more accurate, data-driven way of determining someone's sexual orientation, without their knowledge and possibly against their will, make the world better, more just? What, apart from the questionable benefit that technologists can now figure out if people are gay without actually asking them, do we as a society stand to gain from this tool? How does a decision model that is committed to the idea that your face is your destiny promote, rather than undermine, justice?

Similarly, we need to deliberate about the *purpose* of using sophisticated technology in other domains of AI innovation, before asking whether using that technology would work well. This includes questions like the following: why do we need AI to tell us which faces are unattractive for not being "perfectly" symmetrical—and why do we need AI to evaluate our tone of voice for the ostensible purpose of monitoring our emotions and making inferences on that basis about our health?

The "Do Not Build" Argument

In principle, a simple choice presents itself: there are two basic possible strategies we might choose to pursue in order to mitigate potential harmful consequences associated with the development and use of contemporary forms of AI. On the one hand, we can try to *incrementally optimize* such tools before and after we build and deploy them, with the aim of making them more accurate, more fair, more transparent, and

Some AI tools work fairly well across the board… But their very use may be morally unjustifiable and politically harmful if they rely on methods of inference that echo unjust historical practices.

so on. We can alter algorithmic decision rules, improve data sets by making them more fine-grained and representative, better measure and operationalize key concepts relevant to the decision task at hand, and so on. 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, for instance in projects by academic and industry research teams. The reasoning underpinning OPTIMIZATION strategies is intuitively appealing: even if technologists cannot fix society, including all of its inequalities, technologists can at attempt to fix what is fixable in the short run: bad technology.

But OPTIMIZATION need not focus narrowly on improving technological tools themselves: more broadly construed, OPTIMIZATION might also mean *improving society* by deliberately using AI tools for the purpose of accurately diagnosing and measuring social disparities. AI, on this picture, functions as a supercharged heuristic device: rather than taking a "fix technology first, fix society later" approach, this strategy hopes to fix society *by* using technology to identify the extent and causes of social injustices. In an influential paper, computer scientist Jon Kleinberg and co-authors argue that "beyond making it much easier to detect discrimination, the introduction of an algorithm into the decision loop now makes it also much more feasible to identify and scale useful fixes."[12] The idea is that algorithmic systems might well be *better suited* to that heuristic purpose than human decision-makers, given the messy complexity and comparative opacity of our own human reasoning processes: contrary to popular perception, then, humans seem like the real black-box, not AI. Therefore, utilizing AI as a diagnostic tool seems to present a promising opportunity for precise, targeted, and equity-oriented technological intervention and policy change: "the risk that algorithms introduce is not from their use per se [...] if we make the necessary adjustments to account for the different world we are in, algorithms have enormous potential to be not just a risk to be managed but actually a force for social good."[13]

[11] Francis Galton, *Composite Portraits of Criminal Types*, 1877, The Galton Archive, University College London, Special Collections, https://www.metmuseum.org/art/collection/search/301897.

[12] J. Kleinberg et al., "Algorithms as Discrimination Detectors," *Proceedings of the National Academy of Sciences* 117, no. 48 (2020): 30096–30100.

[13] Kleinberg et al., "Algorithms as Discrimination Detectors," 30100.

If we can confidently predict that building a specific AI tool would lead to severely harmful outcomes, if we are uncertain about whether any potential risks associated with AI innovation are worth taking at all, or if the purpose of using a particular AI model is unjust or otherwise badly defined, we might choose not to build that tool in the first place.

On the other hand, there may well be cases in which we might choose to stop deploying—and thus, to stop optimizing—a given AI tool, once it becomes apparent that its use leads to unjust or otherwise harmful outcomes. The recent local deployment bans of facial recognition tools used in law enforcement in San Francisco and other US cities are a good example of this. These tools were banned for the time being not only because influential studies, in particular one conducted by Joy Buolamwini and Timnit Gebru, had shown that currently used facial recognition systems are highly inaccurate when classifying people of color and women,[14] but also because their currently still unregulated use was deemed too dangerous even if—and indeed *especially* if—those tools were highly accurate for *everyone*. In light of the fact that current law enforcement practices are characterized by deep systemic racial injustice, tools that simply *support and accelerate* the smooth functioning of such practices do not seem to be the *best* we can do in our pursuit of greater social justice. In fact, the use and continued optimization of such tools may actively *undermine* social justice, if they continue to operate in a social status quo that is itself systemically unjust.

Surprisingly, even private tech corporations agreed that non-deployment was the best solution in this case: Amazon, Microsoft, and IBM all voluntarily imposed temporary non-deployment moratoriums on themselves until the use of facial recognition technology in law enforcement is subject to comprehensive legal regulation.[15] Of course, it is likely that *corporate* endorsements of non-deployment strategies are primarily financially, rather than ethically, motivated: given recent widespread and politically powerful protests against racially unjust policing practices, voluntary moratoriums may be the most efficient way of preventing PR fallouts while minimizing costs. Nevertheless, it is noteworthy that at least in *some* domains of AI deployment, even tech industry practitioners, who would ordinarily be the most persistent critics of NON-DEPLOYMENT strategies, have publicly advocated for the latter. Of course, the second strategy could start at an even earlier juncture. If we can confidently predict that building a specific AI tool would lead to *severely harmful* outcomes, if we are *uncertain* about whether any potential risks associated with AI innovation are worth taking at all, or if the *purpose* of using a particular AI model is unjust or otherwise badly defined, we might choose not to *build* that tool in the first place.

No Blanket Solution

Which of these strategies is better, all things considered? It is impossible to articulate a general argument—an argument that applies to all possible uses of AI—in favor of one strategy over the other. After all, whether a given AI tool yields harmful consequences in a given area of deployment will depend heavily on a large number of empirical, contextual factors, such as how the tool actually works, which problem it is tasked with solving, how the tool interacts with society, and which social structures of advantage and disadvantage are in place in that society. These kinds of facts about what the social world is like are subject to change: political and institutional transformations geared towards mitigating background structures of injustice may change the ways in which people are socially situated, and evolving norms will affect the ways in which people will interact with technology. Since technological models always interact dynamically with the social world, whether or not a given AI tool will in fact lead to unjust or otherwise harmful consequences will depend on the political, social, and institutional context in which it is deployed. As a result, a *blanket* endorsement of OPTIMIZATION—much like a *blanket* endorsement of NON-DEPLOYMENT or the DO-NOT-BUILD argument—is not tenable. Which strategy is more ethically justifiable and politically persuasive will require a careful, context-sensitive human judgment on our part.

Even if it is not possible, then, to articulate a *general* principle about either strategy being *always* preferable to the other, we *can* adopt a principle for how we, as human agents interacting with AI, ought to reason about whether a given AI tool should be deployed and optimized, or not built and deployed at all: "don't move quite so fast—and perhaps you won't break so many things."

Current practice in the tech industry—aside from the specific issue of facial recognition technology used in policing, as well as a limited number of other tools—follows exactly the opposite principle: the idea of moving fast and breaking things still looms large. Former Yahoo! CEO Marissa Mayer is often quoted as saying "with data collection, 'the sooner the better' is always the best answer."[16] Amazon's (in)famous leadership principles list—aside from "be right, a lot"—speed: "Speed matters in business. Many decisions and

[14] Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018): 1–15.

[15] Larry Magid, "IBM, Microsoft And Amazon Not Letting Police Use Their Facial Recognition Technology," *Forbes*, June 12, 2020, https://www.forbes.com/sites/larrymagid/2020/06/12/ibm-microsoft-and-amazon-not-letting-police-use-their-facial-recognition-technology/.

[16] "From the Archives: Google's Marissa Mayer in Vogue," *Vogue*, March 28, 2012, https://www.vogue.com/article/from-the-archives-marissa-mayer-machine-dreams.

actions are reversible and do not need extensive study. We value calculated risk taking."[17] Technology corporations have strong incentives to "chase the State of the Art." In order to achieve or maintain market dominance, tech corporations must therefore build and deploy as many innovative AI tools as possible—as quickly as possible: if it is *possible* to build a tool, the reasoning goes, then one *ought* to build it. Rather than asking "why should this technology exist? Why build this tool?," most technologists are likely to ask: "why *not*?"

Meanwhile, many tech industry practitioners are increasingly worried about—and actively working to counteract—AI's potential ethically and politically harmful impact. Most major technology corporations now have designated teams focusing on "ethical," "trustworthy," or "responsible" AI. This, of course, is in and of itself a rather limited indicator of whether members of tech corporations are genuinely concerned about the moral and political stakes of AI, or instead predominantly concerned about issues of legal compliance and strategic branding. The recent firing of the two co-leads of Google's ethical AI team, Timnit Gebru and Margaret Mitchell,[18] reveals that industry AI ethics efforts can often be restricted and outweighed by competing corporate aims. But employees of major tech corporations are also increasingly organizing *themselves*, often against significant corporate pushback, with the aim of holding their employers accountable for potential harmful uses of AI. The Alphabet Workers Union's mission statement, for example, emphasizes that "we will use our reclaimed power to control what we work on and how it is used. [...] We are responsible for the technology that we bring into the world, and recognize that its implications reach far beyond Alphabet."[19]

It is worth noting that while mission statements like AWU's are in principle compatible with both OPTIMIZATION and NON-DEPLOYMENT, they typically do lean heavily towards OPTIMIZATION, and specifically OPTIMIZATION *within* powerful corporate structures. As of August 2021, AWU's mission page stated that "we will work with those affected by our technology to ensure that it serves the public good. [...] Alphabet can make money without doing evil." Under a section titled "Why We Organized," the union says "help us make Alphabet the best." Statements like this reveal a common attitude in the technology industry: the idea that continued technological optimization across various domains can and should benefit everyone, and that optimizing technology with justice-oriented aims in mind is perfectly reconcilable with the aim of financial profit, concentrated primarily amongst a small number of powerful private corporations.

In order to achieve or maintain market dominance, tech corporations must therefore build and deploy as many innovative AI tools as possible—as quickly as possible: if it is possible to build a tool, the reasoning goes, then one ought to build it. Rather than asking "why should this technology exist? Why build this tool?," most technologists are likely to ask: "why not?"

Skepticism against the NON-DEPLOYMENT strategy, then, need not necessarily be exclusively or even predominantly be motivated by financial gain: many technologists reason that if their team does not build a given tool, someone else will and that the someone else could be a corporation or a government entirely unconcerned with the potentially harmful social impact of AI.[20]

[17] "Leadership Principles," Amazon, accessed August 15, 2021, https://www.amazon.jobs/en/principles.

[18] "Timnit Gebru was fired from Google — then the harassers arrived," *The Verge*, March 5, 2021, https://www.theverge.com/22309962/timnit-gebru-google-harassment-campaign-jeff-dean; "Google fires Margaret Mitchell, another top researcher on its AI ethics team," *The Guardian*, February 19, 2021, https://www.theguardian.com/technology/2021/feb/19/google-fires-margaret-mitchell-ai-ethics-team.

[19] "Mission Statement," Alphabet Workers Union, accessed August 15, 2021, https://alphabetworkersunion.org/principles/mission-statement/.

[20] Indeed, the developer of the aforementioned facial recognition technology used 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" (*supra* note 4). However, there are strong reasons to reject this type of argument. Even if a well-intentioned researcher is the first to develop a given tool, this does not prevent a "bad actor" from repurposing knowledge and technological capabilities for morally bad purposes—and even if bad actors do *not* repurpose knowledge in such a way, and the only available version of the tool remains the one developed by a well-intentioned researcher, the tool might still have morally bad social *effects* despite the developers' original intentions.

Why Not Just Make Things Better?

It is easy to see why, at first blush, there is a strong presumption in favor of AI optimization. After all, who could object to *making things better*? One need not be an innovation evangelist to prefer OPTIMIZATION: even if we do not think that innovation is worth it for the sake of innovation, that innovation is *inherently* valuable, we might very well think that if we can lessen harms associated with the use of a tool by developing a new-and-improved version of that tool, we ought to do so. DO NOT BUILD and NON-DEPLOYMENT, by contrast, seem to amount to nothing more than *giving up*: not even trying to solve problems like algorithmic bias arising in our current status quo, and deliberately refusing to use highly sophisticated technology for the good of all.

The supposedly defeatist attitude underpinning NON-DEPLOYMENT and DO NOT BUILD seems particularly dangerous precisely *because* of the fact that we are often operating under significant uncertainty when it comes to AI: uncertainty about how much we can optimize a given tool if we just try, uncertainty about how exactly such improvements will impact society, and uncertainty about whether a given tool might respond unpredictably to being deployed under a new set of empirical conditions. This might be surprising to those who view algorithmic decision-making as the automated equivalent of a dedicated bureaucrat: a system that predictably, doggedly implements whatever principles its human developers force it to implement. In his path-breaking 1950 article "Computing Machinery and Intelligence," published in the philosophy journal *Mind*, Alan Turing pointedly argued that "the view that machines cannot give rise to surprises is due, I believe, to a fallacy to which philosophers and mathematicians are particularly subject. [...] Machines take me by surprise with great frequency."[21] Of course, such surprises include *positive* surprises: over the course of an OPTIMIZATION effort, we might well find that a given AI tool starts working *remarkably well*, that it is useful for solving highly complex decision tasks that humans would not be able to solve.

Trial-and-error, accidents, and surprises have always been a necessary component of innovation. This much is indisputable: many socially beneficial scientific and technological advances are unforeseeable and unplanned. However, efforts to build and optimize innovative new AI tools need not necessarily bank on fortuitous accidents. More specifically, OPTIMIZATION need not always mean risky deployment on a massive scale: often, we can build new tools, and cautiously deploy them under relatively controlled conditions.

But even though there is a plausible presumption in favor of OPTIMIZATION, there are several good reasons to resist the kind of blanket endorsement of OPTIMIZATION that is now common in the technology industry. Choosing not to build, deploy, and optimize does not necessarily mean making things worse: indeed, it can move us closer to justice.

Bad Goals

Recall our earlier examples: cases in which AI is used for a strange, superfluous, and possibly harmful purposes, including the QOVES's AI-powered beauty assessment tool, which reinforces the damaging idea that there is one "objective" standard of beauty. Beyond this, there is the additional problem that the development of seemingly apolitical AI applications can facilitate the political persecution of vulnerable groups at an unprecedented scale. Consider Megvii, the Chinese company that developed Face++, a tool that is similar to QOVES's tool in that it provides a "beauty score" to users. This was not the only purpose for which Megvii used its expertise in facial recognition technology: it collaborated with Huawei to test a "Uighur alarm" tool, designed to recognize the faces of members of the Uighur minority and alert the police.[22] When nefarious purposes motivate efforts to build and optimize AI, we have weighty moral and political reasons to counteract and critique such efforts.

But the motivations of developers and corporations need not be *explicitly* nefarious in order for the goals pursued via an AI innovation to count as *bad*. A developer team might well think that the continued development of facial recognition technology in a policing context serves justice for all, in the sense that everyone in society has a better chance to receive fair and accurate treatment when interacting with law enforcement. But of course, if the institution of policing as such in its current form, under current social and political conditions, is unjust on a deeper level, the aim of building-and-optimizing is still misguided. Rather than trying to scale up and improve *existing* policing practices further, we should ask: what would it take to create a system that helps us enforce the law in a more just, less violent way? Who should be involved in policing, and which practices should we adopt in this context? Which larger-scale institutional, economic, and legal transformations are needed for this purpose?

[21] Alan M. Turing, "Computing Machinery and Intelligence," *Mind* 59, no. 236 (1950): 433–60.

[22] Drew Harwell and Eva Dou, "Huawei tested AI software that could recognize Uighur minorities and alert police, report says," *Washington Post*, December 8, 2020, https://www.washingtonpost.com/technology/2020/12/08/huawei-tested-ai-software-that-could-recognize-uighur-minorities-alert-police-report-says/.

Answering these kinds of questions may not necessarily involve AI, of course—though it may well do. If so, rethinking and redefining the purpose of using AI in the first place is more important than continuing to build, deploy, and optimize the tools for which the technology industry has already pre-defined a narrowly articulated purpose.

The same reasoning applies to all other areas of AI. In the context of algorithmic hiring, we might ask: would justice require that we continue to build tools that rank applicants depending on how well they fit into a company, and optimize those tools further with respect to company fit? Optimizing for company fit often ends up systematically disadvantaging women and members of marginalized groups. Instead, we should therefore *redefine the goal* of the decision task. We might ask: what does it *mean* to be a "good applicant," and how *should* this be measured? Should AI be used for the purpose of *ranking* applicants in the first place, or could it be better used for a different purpose?

Rethinking and redefining the purpose of using AI in the first place is more important than continuing to build, deploy, and optimize the tools for which the technology industry has already pre-defined a narrowly articulated purpose.

Irreversible Choices

OPTIMIZATION evangelists may reason that all these problems do not in fact undercut a blanket endorsement of OPTIMIZATION. After all, we *always* have the option of non-deployment: if at some point in the future it turns out that the original purpose of a given AI tool was ill-defined, that it carried historical baggage, that it keeps interacting with and replicating an unjust social status quo, and that its use as a heuristic device is limited, we can simply decide to stop using the tool *then*.

The claim that non-deployment is usually a feasible option down the line is, of course, true in many cases: recall the successful local bans on facial recognition tools in policing. But it is not *always* true: sometimes, we cannot *fully* take back an earlier decision to deploy AI tools in a given domain.

In some cases, an earlier decision to automate a process can have irreversible effects on complex social systems: suppose that a company realizes that some task can be completed much more efficiently by an AI system than a human employee, and thus decides to fire and replace all employees in charge of that task. Even if there is backlash against the specific AI tool that the company uses to automate that task, and the company therefore decides to stop deploying that tool, it is not likely that those who have lost their jobs due to automation will be rehired as a result. It is significantly more likely that the company will attempt to identify another way of automating that task: automation in the labor market can be irreversible *de facto*, even though non-deployment technically remains an option, because market actors have no particularly strong incentives to revise an earlier decision to deploy AI.

In fact, there are often strong incentives in the other direction: given that being able to claim to rely on state-of-the-art, powerful AI offers companies a major advantage over competitors, it is in their interest to increasingly create task-driven, low-paid, boring, precarious roles for human employees that invisibly support lucrative AI-driven products and services. Anthropologist Mary L. Gray and computer scientist Siddharth Suri call these types of roles "ghost work": the kind of unrecognized human labor that will become increasingly necessary the more we build and deploy AI.[23]

This can even go as far as companies *pretending* to have built and deployed AI, while actually having human employees pose as an AI system to complete a given task: in 2016, it became public knowledge that personal assistant start-up X.ai was directing hundreds of employees to pretend to be superhuman AI chatbots, performing avalanches of mind-numbing, repetitive tasks like generating auto-reply emails and scheduling appointments.[24] This fake-it-till-you-make-it approach is popular amongst start-ups, because it is a low-cost way of testing out how the market responds to their services before actually having to build and deploy complex, expensive AI tools. It would be misguided to hope that once

[23] Mary L. Gray and Siddharth Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (Boston: Houghton Mifflin Harcourt, 2019).

[24] Ellen Huet, "The Humans Hiding Behind the Chatbots," *Bloomberg News*, April 18, 2016, https://www.bloomberg.com/news/articles/2016-04-18/the-humans-hiding-behind-the-chatbots.

We—all those affected by the use of AI—need meaningful opportunities to deliberate and argue about the question of if (and if so, why) building, deploying, and optimizing AI in a specific area of public or private life will in fact serve a sufficiently demanding idea of justice.

such tools are *actually* built and deployed, the restructured work force would revert back to its pre-ghost work structure: even the *illusion* of AI-powered decision-making, much like AI-powered decision-making itself, can irreversibly damage the structure of social systems.

Of course, building, optimizing, and deploying AI is usually not an all-or-nothing decision: there is often a middle ground between not using a given tool at all, and using it everywhere. It seems reasonable to argue, for example, that when we anticipate that using AI in a specific decision domain risks bringing about unjust or otherwise bad outcomes, we could stop deploying it in the world at large, while continuing to do academic research on it in a secure, carefully constrained setting.

But things are not so simple, as the example of DukeMTMC—a dataset of videos recorded in public spaces on Duke University's campus—demonstrates. The dataset was made publicly available without data protections and used for controversial research on computer vision-based surveillance technology. In response to significant public criticism, this dataset was taken down in June 2019. However, as a team of Princeton University researchers has recently argued, at least 135 research papers utilized that data set after it had been taken down,[25] as well as derived datasets that continue to be publicly available.[26] Even if we build and optimize AI tools for research purposes only, it is not necessarily true that doing so will have no morally and politically harmful effects on society: when unethically collected data can potentially be used for research on how best to accelerate objectionable social practices, the mere fact that only a limited number of people have meaningful access to that data is not a sufficiently strong safeguard against the risks of harm associated with it.

Where Next?

Settling the question of whether a given AI system should be built and optimized, or not built and deployed at all, will have to require careful attention to the socio-political context in which technology operates in each case: there is no blanket solution. This, however, does not mean that a series of one-off, context-specific "non-deployment victories," such as local facial recognition bans and self-imposed corporate moratoria, are going to settle the issue conclusively. Industry-wide regulation in high-stakes AI deployment domains like policing and criminal justice are likely to come—and that would be preferable to one-off (non-)deployment decisions made by a narrow, democratically unaccountable group of technologists. But *democratically legitimate* regulation will require an even deeper shift: we—all those affected by the use of AI—need meaningful opportunities to *deliberate* and *argue* about the question of if (and if so, why) building, deploying, and optimizing AI in a specific area of public or private life will in fact serve a sufficiently demanding idea of justice.[27] Getting clear on the purpose and value of artificial intelligence, then, is more important than the rush to make it better.

[25] Kenny Peng, Arunesh Mathur, and Arvind Narayanan, "Facial recognition datasets are being widely used despite being taken down due to ethical concerns. Here's how.," Freedom to Tinker, October 21, 2020, https://freedom-to-tinker.com/2020/10/21/facial-recognition-datasets-are-being-widely-used-despite-being-taken-down-due-to-ethical-concerns-heres-how/.

[26] "sxzrt/DukeMTMC-reID_evaluation," GitHub, accessed August 15, 2021, https://github.com/sxzrt/DukeMTMC-reID_evaluation.

[27] For a defense of the argument that cautious deliberation about the social effects of algorithmic systems, as well as a cluster of related epistemic duties, are actually a component of *procedural justice* broadly construed, see Annette Zimmermann and Chad Lee-Stronach, "Proceed with Caution," *Canadian Journal of Philosophy* (2021).

Carr Center Discussion Paper Series

Carr Center for Human Rights Policy
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

This publication was published by the Carr Center for Human Rights Policy at the John F. Kennedy School of Government at Harvard University.

79 JFK Street
Cambridge, MA 02138

617.495.5819
https://carrcenter.hks.harvard.edu