8

Near-Term Artificial Intelligence and the Ethical Matrix

Cathy O'Neil and Hanna Gunn

There are several strands of recent work on AI, including a focus on more abstract philosophical problems, among others: Could AI have genuine emotions? Will the singularity be the end of the species? If we can, should we upload our minds? But there is very important research to be done on person-affecting problems raised by the use of AI systems both in the present day and in the near future. In particular, there is a pressing need to recognize and evaluate the ways that structural racism, sexism, classism, and ableism may be embedded in and amplified by these systems. More generally, there are concerns that the adoption of AI ignores the interests and needs of anyone who isn't part of the development or deployment team.

In this paper we take up the issue of near-term artificial intelligence (AI). "Near-term AI" is used to denote artificial intelligence algorithms that are already in place in a variety of public and private sectors, guiding decisions that pertain to advertising, credit ratings, and sentencing in the justice system. Our focus here is to contribute to a critical discussion of the ways that AI is already being widely used in decision-making procedures in these areas. We will argue that developers and deployers of AI systems—in senses to be defined—bear a particular kind of responsibility for the moral consequences of near-term AI. We will present a tool to aid developers and deployers in engaging in the moral reflection we argue is required of them, in order both to help them to meet their moral obligations and to help address the material risks posed by what we take to be the status quo of actual near-term AI development. This chapter can be understood as a contribution to the field of technology assessment, but instead of suggesting policy revisions, we will propose a framework for ethical analysis that can be used to facilitate more robust ethical reflection in AI development and implementation.

We begin in section 8.1 by introducing near-term AI as algorithms designed as expert systems to replace human decision-makers. This is despite many algorithms being designed as complementary to human decision-makers rather than replacements for them. We then proceed to argue that the current status

Cathy O'Neil and Hanna Gunn, *Near-Term Artificial Intelligence and the Ethical Matrix* In: *Ethics of Artificial Intelligence.* Edited by: S. Matthew Liao, Oxford University Press (2020).
DOI: 10.1093/oso/9780190905040.003.0009

quo in designing and implementing near-term AI doesn't meet minimum acceptable ethical standards because the designers of these algorithms fail to consider the interests of a wide enough range of stakeholders—most significantly, those who will actually be evaluated by these AI systems. We will argue that the present norm that establishes who counts as a designer of an AI algorithm is such that typically only the developers (e.g., data scientists or programmers) and the deployers of the algorithm (e.g., a court, a local government) count. We take this to be problematic, as we will argue it is a primary cause of why the interests of wider stakeholders do not make it into the development of the algorithm, for example, the interests of those who are evaluated or judged by the algorithm. In section 8.2 we argue that we need to develop a wider definition of "success" for near-term AI that better reflects the interests of a wider range of stakeholders. In section 8.3 we discuss a case study on the choice to optimize an AI to different definitions of "fairness"; we show how this decision cannot be separated from ethical decision-making, supporting our argument that designers have moral obligations in the development of AI. In section 8.4 we introduce the ethical matrix framework as a tool for intentionally analyzing the ethical consequences of a new technology. The ethical matrix was proposed by Ben Mepham[1] as a guide for analytic ethical reflection by nonethicists; it typically consists of a 3x4 matrix of three ethical concepts (autonomy, well-being, justice) and four stakeholders. To complete a matrix, one considers how each stakeholder will predictably be affected by the new technology with respect to their interests as represented by the ethical concepts. In section 8.5 we present an ethical matrix that incorporates the language of data science and apply this to a case study. We conclude in section 8.6 with a modified version of the ethical matrix to propose a tool that data scientists can build themselves.

8.1. Problems of Near-Term AI

When we hear "artificial intelligence" we typically think of robots and machines capable of thinking and acting like humans, or, alternatively, of robots and machines that are *far more* intelligent than humans. The thought often continues along dystopian lines, so that these superintelligent machines pose a great threat to humans in one way or another. We will call these "futurist" AI systems, with corresponding "futurist" concerns. At the present moment, though, we do not have superintelligent robots plotting against us. This doesn't mean, however, that there is no artificial intelligence around—the problem is that we're not good at recognizing it. Recent public scandals on the data-trawling business models of social media companies, however, have started to redirect some attention to the AI already in play in many products we use and services we rely on.

Futurist concerns of the sort above will likely resonate with persons most familiar with AI from predominantly science fiction, though of course much serious academic work has also discussed the plausibility and risk posed to humans by AI of the future.[2] While there are numerous academic and researched-based initiatives in place to address a range of issues around the AI presently in use, the status quo in *industry* is still not to engage with the ethical concerns that are becoming more widely recognized by academics and researchers. Some examples of these initiatives are the Campaign to Stop Killer Robots (a conglomerate of NGOs, including Human Rights Watch and Amnesty International), the AI Now Institute, the ACM FAT* annual conferences, and numerous AI labs at universities internationally. Nonetheless our general claim about the intuitive concerns posed by AI stands, that industry standards have largely not adopted ethical goals or interests within their design briefs and that many people are not aware of or concerned about many of the algorithms already involved in making decisions in our lives—despite the ways that they or people they know are affected by them.

Many AI algorithms automate a task previously performed by human workers with expertise or specific training. An automated algorithm can crunch larger amounts of data very fast to deliver a result and thus can either replace a human worker or speed up someone's work. We will call these presently existing algorithms near-term AI, "near" because the cases we are concerned with are either already at hand or are in the process of design and/or implementation. The general blindness to presently existing artificial intelligence has consequences: if we're not paying close attention to the artificial intelligence around us, we're hardly likely to be making sure that it is designed and implemented in ways that—at minimum—meet widely recognized moral standards and avoid inflicting great harms.

A word on terminology is required. Whether or not the examples we discuss qualify as AI for an expert machine learner is irrelevant; from the perspective of the targets of these scoring algorithms, they are sophisticated and opaque black box systems that make important decisions about people's lives. There exists a cluster of ethical problems that arise with automated algorithms that warrant discussing them as a general type, even if we can make more fine-grained distinctions between the varieties of machine intelligence presently in existence.[3] For instance, if an automated algorithm denies one's family medical coverage without warning (or without a meaningful warning)[4] because one failed to check a box on a digital form, it makes little difference whether it's ELIZA or Deep Blue behind the scenes. In both cases, we can ask questions about the decision to design and implement an automated algorithm with the power to remove a family's medical coverage without warning. These kinds of decisions around the design of the automated algorithm, including the choice of data sets,

are ethically problematic aspects of currently existing automated algorithms—whether those algorithms are complex lookup tables or neural networks.

In this first section, we want to bring attention to a number of these ethically problematic issues around the design of near-term AI that we will argue stem from a common source: a failure to consider the interests of many persons who will be deeply affected by the algorithm.[5] We will use the term "stakeholder" to denote those deeply affected by an algorithm; these may be the producers of the algorithm, the deployers of the algorithm, those scored or otherwise evaluated by the algorithm, or those companies or communities whose lives or professions will be disrupted by the widespread adoption of the algorithm. This is not an exhaustive list, and we take up the issue of recognizing stakeholders throughout this piece. We take it that, of all the stakeholders there are with respect to a particular algorithm, only two groups are typically taken into consideration in the design of an algorithm, these being the developers and the deployers. We turn now to an example of an algorithm in Indiana to bring out some of the ways that just who gets to contribute to design can determine whose interests are taken into account and the harms that can accrue to those excluded from this process (and excluded because they are not actually involved in design or excluded because their interests are not considered by others who are in the position to do so during design).

Consider an algorithm designed and implemented for the Family and Social Services Administration (FSSA) of Indiana in 2006, which aimed to modernize the provision of welfare benefits, food stamps, and public health insurance (Medicaid).[6] The goals for the new system were to reduce fraud, reduce public spending, and reduce the welfare rolls.[7] The new system replaced individual caseworkers with an automated eligibility-determining process that used a website for applications and a (privately run) centralized call center to replace one-on-one meetings with caseworkers.

One important factor in deciding who gets benefits is a basic decision about whether to err on the side of minimizing false positives, in this case people receiving benefits they do not need, or false negatives, taking benefits from people who do need them. Prior to modernization, the false positive and false negative rates in Indiana for the provision of welfare were consistent with US national averages, at 4.4% and 1.5%, respectively—a trend in erring on the side of giving benefits to those who don't need them.[8] The goals of the new system erred on the side of producing more false negatives—denying benefits to people who are in need—and so the algorithm was designed with this in mind. Overall the combined error rates rose between 2006 and 2008 to 19.4%, with the greatest rise in benefits denied to people who needed them, a false negative rate of 12.2%.[9] One of the striking features of the new system was a "one-size-fits-all" denial notice: "Failure to cooperate in establishing

eligibility." The language itself, while vague, is strikingly confrontational and accusatory, in effect telling applicants that they have been denied benefits because they are being uncooperative by shirking the rules and regulations of the state. Whether the fault in fact lay with the applicant was not a condition for receiving this notice, and it was not accompanied with any further explanation for why an individual's application was denied. The new system did not make use of existing personal records from the previous system, instead requiring all users to resubmit all of their paperwork. This led to very high rates of lost documentation and a denial of benefits for allegedly "failing to comply."

While it is hard to say that there is a single, central fault in the design and implementation of the new FSSA system, we think it is clear that one of the major downfalls of its design and implementation was a failure to properly engage with the impact of the proposed algorithm on a sufficient range of stakeholders. The goals are explicitly those of the local state: save costs, minimize free riders and fraudsters, and reduce welfare rolls. The new system as a whole is geared toward meeting these goals: reducing staff (by utilizing a private, centralized call center and a computer-based checklist), automating applications (via an online application system), and automatically denying benefits to anyone who makes a mistake or misses a deadline.

Whose interests are not taken into account in this design? Most significantly, the new system isn't concerned with prioritizing—or even affording minimal consideration to—the interests of applicants or caseworkers. The benefits that would seem to accrue with the system are to the state (which may save money), the companies that produced the algorithm, and perhaps to the politicians who fulfill campaign pledges. The developers and deployers failed to determine the predictable ethical consequences of their decision to prioritize the reduction of false negatives for those who would be scored by the algorithm. The result? First, the system denies to people who need public assistance the ability to meet their basic human needs and have access to food, healthcare, and money.[10] The new system also provides little to no person-to-person contact, instead requiring applicants to use the online system, a serious problem for blind or deaf persons who rely on public assistance.[11] Second, the system failed to provide any meaningful level of transparency either to applicants or to caseworkers about how decisions were made regarding the distribution of benefits. Third, the system had serious problems with its data quality by failing to make use of an available database and instead requiring new applications from all beneficiaries. Jane Gresham, a long-term employee with the FSSA, described the new system as "de-humanizing" to both employees and clients. As someone who had been a caseworker with FSSA for two decades prior to the modernization, Gresham described the new system as "factory" work, given the new task-based format,

which undermined workers' abilities to actually oversee a particular client and their needs.[12]

This is a paradigm instance of an ethically problematic near-term AI algorithm. We think that a central failure lies in the way the new automated system was designed with such a narrow focus on the interests of the state. Had the developers, IBM and ACS, and the state been required to consider the interests and needs of caseworkers and those actually dependent on the system as well as those of the state, it is hard to see how this modernized replacement would have been the result.[13] A wider consideration of other stakeholders' interests would make it far more likely someone would have reasoned through the consequences of denying poor families access to food stamps and sick persons their healthcare because they missed an automated phone call. The sheer failure of the system to recognize the costs to real persons' lives in the interests of economic and timely efficiency was remarkable.

We argue that developers and deployers of near-term AI have a moral obligation to engage in ethical deliberation about the consequences of the algorithms they design and deploy. In particular, they have an obligation to engage in a process of determining the predictable consequences of their design choices from the perspectives of those who will predictably be deeply affected by those choices, and to then make an informed decision about how to balance competing interests and values against one another. These include choices about what the purpose or goal of the algorithm is to be (e.g., minimizing welfare rolls) and choices about how the algorithm will meet those goals (e.g., by optimizing to false positive rates). As we will continue to establish, developers (again, those who actually write the algorithms) are in a unique position of responsibility over the design of the algorithm as they are typically the only ones in a position to understand how the algorithm functions and are responsible for rendering the design goals into the algorithm. We will argue for a minimal standard for meeting this obligation, and that is to actually engage in a structured reflection on the predictable consequences of the algorithm by using an ethical matrix. We seek here to establish that such structured reflections are necessary and possible, and we provide a framework for engaging in them. This framework does not require specialist training; rather it asks individuals to apply their commonsense intuitions about, for example, what is fair, in combination with empirical data about the predictable consequences of the algorithm's design.

One of our concerns with near-term AI is that, because the developers are concerned primarily (or exclusively) with their own interests (as a company) or those who will deploy the algorithm (e.g., the state), near-term AI is at great risk of exacerbating harms to already marginalized groups because the interests of those groups are not a part of the conversation around design. Take the FSSA's new system that lacks provision for disabled persons—plausibly an oversight

that we think could constitute ableist structural discrimination.[14] In addition, the presumption that errors in an application indicate that an individual is attempting to engage in fraud or is free-riding can, in our view, plausibly be interpreted as part of a pattern of classist discrimination, that is, a pattern of negatively stereotyping those who are dependent on welfare. This indicates a general lack of attention to the broader social context in which many near-term AIs come into existence. If our intuitions are on point, it also indicates a lack of attention to the ways that certain groups consistently fail to have their needs recognized and taken seriously by contributing to a pattern of failing to consider the needs of members of these groups. The failure of AI developers and deployers to actually engage in thinking through the consequences of their design choices thus maintains these discriminatory patterns in new ways.

We've drawn a distinction between near-term AI concerns and futurist concerns, but we take it that a failure to address our near-term concerns will make it more likely that a variety of futurist existential threats will materialize and that they will be made worse if current discriminatory trends are not addressed. First, they are made more likely because a continued failure to pay attention to the widespread adoption of AI that has been developed without an attempt to address the harmful consequences of its design choices increases the chance that a pernicious algorithm is implemented somewhere. Second, if we are faced with artificial superintelligences beyond human control that lead to existential threats for some large portion of the population, then intuitively it is made worse by our present structural prejudices leading to, for example, a racially discriminatory extermination scenario.[15] A distinct harm that arises in nonsingularity cases, but is no less an existential threat to many persons, is the failure to address the ways that algorithms—like the ones employed in the FSSA example—further materially undermine the poor and may lead to a future of even starker material inequality and a lack of due process for those groups.

The consequence, then, is that we run the risk of widely adopted automated algorithms in our society that make poor people poorer, fail to help the sick, homeless, or otherwise needy, and so put persons lives in serious jeopardy. Continuing to allow for the largely unchecked adoption of automated solutions to social problems could present seriously dystopian situations in a future where goods are distributed only on the basis of lists designed to meet economic or political goals, with no consideration for the nuance of individual needs. While near-term AIs do not seem to present us with human-extermination scenarios, it is no great stretch of the imagination to see how they can lead to dystopian futures where one's very ability to access healthcare, shelter, and food might be due to an inscrutable score provided by the black box of a near-term AI.

Such a scenario is easily preventable if we adopt a norm (or better yet, a policy) of demanding ethical reflection on the ways that different interest

groups or stakeholders will be affected by the implementation of near-term AI by those who are primarily responsible for the design of such algorithms. We make no claim that engaging in ethical reflection on these systems is easy. Someone will have to draw a line somewhere that will leave some people in need of food stamps and medical insurance without the ability to fulfill these needs (keeping the present systems fixed without radical changes to public provisions of these goods). That being said, we think that there is important ethical work that can be done here by establishing and requiring processes that engage developers (the people who build the algorithm with technology) and deployers (the people who use the algorithm once it's built) in a process of ethical and statistical reflection.

Before moving on, we should clarify our terminology. When we use the term "designers," we want it to include—at a minimum—both the deployers of an algorithm as well as the developers. Ideally we'd include other stakeholders as well, or representatives of stakeholders. As we indicated earlier, we take it that in the ideal situation, those who stand to be deeply affected by the adoption of the algorithm would be included in the design of the algorithm. This could take the form of actually including those persons in conversations during the design process, though we argue here for something less: that developers and deployers are required to engage in a process of empathetic design that considers the needs and interests of these groups, and that they then make decisions about design with this information at hand.

It's far from obvious that this would be the definition for "designers" because currently the standard model for corporation or government agency use of algorithms is that a third-party data vendor sells its "black box services" through a licensing agreement that typically doesn't allow the deployer to see the source code or even understand the code even at a basic level. The problem with that is it's harder to trace mistakes and to assign accountability. Indeed another standard element of the legal setup is an indemnification contract that assigns costs of legal settlements to the vendor, allowing the deployer even more dangerous moral distance from the algorithm they use for decisions like who to hire or fire.

So when we suggest that those deployers, who are often currently being kept in the dark about the algorithmic design, should be considered part of the design process, they're not automatically a part of algorithmic design, so their inclusion at this level is rare. And yet to accomplish an ethical and accountable process, we'd argue, deployers will have to be considered part of the design process. That said, as we noted very briefly earlier, there are certain issues that are beyond the average deployer's understanding, namely, the code implementation, often written in computer languages that take years of training to write and understand. This raises the question: How can the overall design process

involve deployers and developers and yet remain accountable to a common set of ethics?

One way in which developers and deployers are similarly accountable is with respect to the values and goals that the algorithm should try to meet, for example, to reduce welfare rolls. In this way, the design team includes both developers and deployers because they have to come to a mutual understanding of which values are to be embedded in the algorithm, and how conflicting values must be balanced. The goal for an accountable algorithmic process would be to split the "ethical decisions" from the "translation of those decisions into code." In other words, the ethical decisions would be made by the entire design team first. Then the development team's job, from the perspective of accountability, would be to faithfully translate those decisions into mathematically precise code. Ideally they would also place monitors into the system to confirm over time that these decisions continue to hold.

For example, if it was decided that the disparity in false negative rates and false positive rates in the FSSA algorithm was an important indicator of fairness of the modernized welfare system, there should be a way to keep track of both rates and ensure they are within a chosen window of uncertainty and tolerance in disparity. Choosing that metric as a fairness indicator, as well as what exactly would be the "window of uncertainty and tolerance," would be choices decided by the entire design team. If we require that this decision must be made in an informed way, then we both help to uncover areas of potential risk and harm to a wider range of stakeholders and we have a more transparent design process for monitoring accountability once it is implemented.

This discussion exposes a difference as well as a commonality between human-created and human-run decision-making processes and automated decision-making processes. They have this in common: they need to be carefully considered in terms of their impact on all stakeholders. They have this difference: black box algorithms are inscrutable to most parties involved in the implementation of and interaction with the automated decision-maker. Therefore they must be audited, preferably continually, to make sure they are functioning as designed and within limits. The development team is uniquely capable to make that happen, which makes them importantly responsible for the ethical consequences of the algorithm and grounds part of their particular ethical obligations with respect to this process.

So far, we've identified several problems with the design and implementation of near-term AI and identified a starting point for fixing them: requiring an inclusive design team to engage in structured ethical reflection on the proposed algorithm that incorporates the interests of a wider set of stakeholders. In the next section, we begin with how common understandings of "success" in data science mask ethical decisions that are made in the design of algorithms.

8.2. A Better Definition of Success

Too often, conversations about machine learning focus on cutting-edge algorithms like the newest chess or Go algorithms, which pit "man against machine" and impress us with the computer's superior memory and learning speed, albeit in a narrow and limited way. We think that the tendency to focus on these types of algorithms—beyond impressing the lay audience—gives the impression that there's a well-defined concept of "winning" or "success" for all algorithms, and that the computer can be taught to understand this definition.

Of course, in the context of Go or chess, those things are true. But they are the exception, not the rule. In any larger, more complex, social setting, which most algorithms inhabit, there is no one definition of "winning." Any definition is inexact and relies on proxies, and often computers end up optimizing to truly ludicrous if not perverse definitions of success, to the detriment of their human targets.[16]

Historically speaking, this tendency to think of algorithmic "success" as "winning" is inherited from the toy universes of games, in that we inherit the very language we use when we talk about machine-learning algorithms in general. Our understanding of success, then, is often narrow and insufficient for understanding whether an algorithm "works" in the messy reality of human social interactions partly because of this inheritance. That simplification of our language allows us to pretend, or assume, that there is a simple concept of success, and that it's one that computers can be taught, given enough data and enough time.

What is in fact happening behind the scenes, however, is that we've set up the algorithm to refer to and optimize to a definition of success that is constructed by and for the algorithm's designers, and that typically ignores the algorithm's other stakeholders. We might, in fact, say that the only stakeholder is the data scientist and perhaps the company for whom the data scientist works, and the only concern is accuracy, profit, or efficiency, depending on what kind of algorithmic context the data scientist works in.[17]

In general, when we have two metrics, A and B, which are distinct, and we optimize to A, we necessarily do not optimize to B. Indeed when we optimize directly to A we end up optimizing directly away from B with very high probability. The extent to which metric B suffers when metric A is preferred depends on how different they are, how much the algorithm matters, and how much of a feedback loop is produced by the choice of metric A in the first place.[18]

Just to give an example, let's choose A and B to be rather close. When the *US News & World Report* magazine decided to rank colleges, it chose a rather weak set of proxies for "quality," which included the rate at which students who applied were admitted, the rate at which students who were admitted actually accepted,

and the reputation of the college according to other college administrators.[19] Let's set A to be that *US News* definition of a "quality" college, and let's consider B to be the definition of quality that a typical high school senior might care about, which would include costs, educational and professional opportunities and connections, location, and prestige. Enormous effort at enormous cost has been put into gaming the ranking system by college administrators, running the spectrum from cheating to building outsized luxury gyms to attract elite athletes. That cost has translated to higher tuitions. However, the *US News* ranking system doesn't care about cost. In other words, choosing a proxy to college quality that is blind to cost and optimizing to it has meant that any quality proxy that is sensitive to cost will be directly punished.

Here's the problem. When data scientists develop an AI, they choose a metric, A, to optimize to in order to determine whether the AI is successful or not. This is problematic in and of itself, as we noted, because this is likely to be too simplistic and will tend to be a very narrow and specific metric that best suits a narrow range of interests, such as profit or statistical accuracy. Given the earlier argument, then, when a data scientist chooses a self-serving metric, A, and we have multiple distinct groups who will be affected by the choice of this metric and whose interests are best served by metric B, then choosing to optimize to A will neglect the interests of the other stakeholders. Thus we should expect that near-term AI will consistently fail to meet the needs of other stakeholders so long as success is determined by a narrow set of data scientist–serving metrics. And this isn't to say that other stakeholders in a given context want the algorithm to be inefficient, inaccurate, or unprofitable—this could be in the interest of other stakeholders too. Rather, other stakeholders may have other interests that are not being taken into consideration and that may need to be weighed against those interests of the designer or of the implementer (whose preference may shape the proxies a designer chooses).

To briefly demonstrate how incredibly harmful this can be, consider an algorithm that predicts child abuse. This is intended to help people who work at a hotline, and indeed those workers deserve to have a data-driven system that enables them to rely on more than their own judgment. The problem lies in the multitude of stakeholders and their accompanying concerns with the outcomes of the algorithm. In particular, if the algorithm is simply optimized for efficiency or accuracy, neither of those will take into account the dire consequences for children or for parents who are inaccurately understood by the algorithm.

We need to expand our understanding of what it means for a given algorithm to work well, and it starts with understanding what "success" means for all the stakeholders with respect to a specific algorithm, not simply those in control of the code. In the next section, we discuss a study that shows some ways that data scientists are working on mitigating racial unfairness in their algorithms. Our

discussion of this study helps to highlight how a data scientist's preferred definition of fairness may conflict with commonsense understandings of fairness. We will not come down in favor of a specific definition of fairness as a consequence of this discussion. Instead we will use our discussion to further motivate the need for a new norm that forces us to intentionally engage in ethical discussions to reveal whose interests are being taken into consideration.

8.3. FICO Scores, Profit, and "Fairness"

Of course, it's not merely deciding what will make an AI "successful" that runs us into ethically consequential choices. There are considerable ethical implications for how we decide to achieve "fairness" in our algorithms too. Moritz Hardt, along with Eric Price and Nathan Srebro presented a case study on how fairness measures can be used with FICO credit scores to determine who gets loans.[20]

Here's the messy real-world context. We know that wealth inequality is correlated with racial groups, and we also know that this is very plausibly because of a long history of racial oppression and structurally racist policies.[21] Thus there is an intuitive sense in which the present wealth distribution is unfair: there are many people with wealth because of systematic patterns of race-based privilege and oppression, and there are many people who are without wealth because of systematic patterns of race-based privilege and oppression. We will use "intuitive fairness/unfairness" in this section to refer to the inequality that is present in these cases of race-based privilege and oppression.[22] Historically, FICO scores were introduced to equalize chances for loans among men and women and among races, which was a way of giving loans to people even if they had fewer economic opportunities. Thus FICO scores were introduced, in a sense, to make economic opportunity more fair.

As Hardt et al.[23] note, FICO scores "are complicated proprietary classifiers based on features, like number of bank accounts kept, that could interact with culture—and hence race—in unfair ways." That is to say, FICO scores themselves are partly the result of systematic patterns of race-based privilege and oppression that can unfairly, in our intuitive sense, evaluate some deserving people as undeserving of loans and vice versa. So while they are an attempt to provide a quantifiable measure of someone's deservingness of a loan, they are highly likely to be subject to human bias and to structural problems of racial oppression.

The Hardt et al. case study investigated five scoring methods, which were set up as follows: if a credit company was given FICO scores and the racial information of would-be borrowers, how might it ensure that its business practices are "racially fair" using that information? Hardt et al. restricted themselves to building a decision engine that would look at someone's FICO score and race

and, depending on whether the FICO score was above or below some threshold, the individual would get the loan or not. Each of the five scoring methods had a different constraint for setting this threshold, establishing four technical definitions of fairness.[24]

At one extreme was "Maximum Profit" with no fairness constraint, whereby a loan-affording threshold was chosen to simply optimize the profit gained from each racial group. Without providing the details, maximum profit was gained when a FICO score threshold is chosen for each racial group such that 82% of that group do not default on their loans. At the other extreme was "Equalized Odds," which requires that the loan-affording threshold be set by determining both the fraction of nondefaulters that qualify for loans and the fraction of defaulters that qualify for loans to be constant across racial groups. Hardt et al. kept track of what the profit margins would look like, how the thresholds for loans would change, and what proportions of the populations by race would end up with loans for each definition of fairness. Importantly, they also considered the incentives that a given definition of fairness would give to the credit company itself.

A third option was the "Race Blind" condition. We would like briefly to contrast the Race Blind option with the Maximum Profit option as they have striking consequences for our intuitive sense of fairness. In particular, the contrast of the two demonstrates how an attempt to be racially neutral can actually have unfair (i.e., racially discriminatory) consequences. The "Race Blind" fairness threshold is set by ignoring racial categories altogether and setting a single loan-affording threshold at which 82% of the whole population will not default.[25] There is a long (and controversial) history of arguments to support the idea that the best way to avoid racial discrimination is to ignore race. This Race Blind loan scoring algorithm is premised on arguments of this sort.

When one applies the Race Blind scoring system, one ends up with a very large population with a correspondingly broad range of FICO scores. As a result, one needs to have a fairly high FICO score to be rated by the scoring system as someone who should get a loan. In a society where racial minorities are less wealthy, the Race Blind system will make it harder for members of racial minorities to get loans because they will, on average, have lower FICO scores. This is the case even though many members of these racial groups with low FICO scores are predictably unlikely to default on their loans. Hence, although this scoring system ignores race, it disproportionately benefits members of some races and disproportionately disadvantages other.

By contrast, if the Maximum Profit scoring system is in place, a threshold is chosen for each racial group such that 82% of the members of that group will not default. Thus members of less wealthy racial minorities, with lower average FICO scores, will have a higher chance to get a loan when the company tries to

maximize profit from them. This is because the algorithm is now more sensitive to the defaulting rates of each racial group. This is striking: a self-serving motive is more fair than one that intentionally attempts to be nondiscriminatory. This is not to argue that the fairest route will always be the one that maximizes profit, but instead it shows that if we do not take the time to seriously consider how our technical definitions of "fairness" (or lack thereof) will impact different stakeholder groups, we can unwittingly bring about unintuitive ethical consequences.

What Hardt et al. considered in their paper only skims the surface of ethical questions about the morality of FICO scoring. Hardt et al. do not discuss whether FICO scores themselves are racist (and we've given reason to think they very well may be) or whether fairness can be determined solely by whether someone would have paid back a loan. For example, perhaps we should instead investigate whether, by giving individual people in a certain subgroup loans even when they might not be able to repay them, the economic status of the whole subgroup goes up because of the added opportunities.

Of course, a larger question has been left unaddressed:[26] When can a given technical definition of fairness, which would require companies to sometimes give loans to people they know might not pay them back but is "good for the group," outweigh the problem of lost profit?

Our "morals of the story" from these discussions about defining "success" and "fairness" when designing a near-term AI system show that there's a strong sense in which we can't actually make these design decisions *without also making moral decisions that impact other stakeholders*. Put differently, developers and deployers are making moral decisions already, but they aren't necessarily identifying them as such and are instead treating them merely as engineering decisions about an algorithm. The status quo in present near-term AI design and implementation is frequently to consider only the very narrow interests of near-term AI developers and deployers, with the result of effectively ignoring the interests of other stakeholders. We've argued that developers, in particular, are in an important position with respect to the accountability for the design and consequences of an algorithm because they are often the ones who are capable of understanding or accessing the system. In addition, they are responsible for translating the goals of the wider decision-making process it will be a part of, for example, giving out healthcare benefits, into the algorithm itself in such a way that it is faithful to the design. We've argued that both developers and deployers ought to be required to engage in a process of ethical reflection that allows them to make informed choices with respect to all of the stakeholders who will be deeply affected by the algorithm. This adds an additional layer of accountability to the design process by making decisions around AI design transparent by essentially requiring a conversation about how to weigh relevant stakeholder interests against one another. Through our discussion of several examples, we've shown that the other

relevant stakeholders will typically include at least those individuals who will actually be judged or scored by the algorithm; later we discuss the selection of stakeholders further.

We shall now introduce and motivate a tool that can be used by design teams to intentionally engage in ethical reflection on their design decisions, and thus become informed about how their choices may predictably impact other stakeholders. Our tool aims to provide a framework for ethical reflection that doesn't require an education in philosophy (not that there'd be anything wrong with that!). The tool is a version of the "ethical matrix" proposed by Ben Mepham[27] that we supplement with the common concerns of data scientists to make it easily usable for those familiar with working in an algorithmic space.

8.4. The Mepham Ethical Matrix

The ethical matrix allows us to engage in ethical reflection on a new technology without having to solve deep ethical problems first, and without specialist ethical training.[28] It does so by requiring us to consider the interests of a range of stakeholders with reference to three general types of moral goods: well-being, autonomy, and justice. In this section, we briefly introduce the motivations for and process of using the ethical matrix, and then discuss the stakeholders and three ethical principles in more detail.[29]

Mepham's initial proposal is heavily influenced by Rawls's early work on how one might adjudicate between the competing interests of persons, which eventually led Rawls to his proposal of the "veil of ignorance"[30] procedure. The three ethical principles that Mepham settles on combine the Rawlsian proposal and the widely used principles of biomedical ethics: autonomy, beneficence, nonmaleficence, and justice.[31] Filling out an ethical matrix requires a process similar to Rawls's famous thought experiment: we are asked to imagine ourselves as each of the stakeholders on our matrix and to consider how we might be impacted by the new technology with reference to the ethical concepts represented in the columns of the matrix. This is a task that takes place before the implementation of the new technology, and so the cells are completed by drawing on the best evidence available for how a given stakeholder may be benefited or be put at risk by the new technology. Thus the task engages us in thinking counterfactually about what is likely to happen if we do or do not adopt the new technology with reference to the present state of affairs. In Table 8.1 we have reproduced a matrix provided in Mepham,[32] analyzing the possible outcomes of genetically modified maize that was designed for herbicide, pest, and antibiotic resistance.

So, for example, in the cell for "Producers" and "Well-being" in Table 8.1, the topics of income and quality of life for farmers are raised. It is possible that after

Table 8.1 The Ethical Matrix

Respect for:	**Well-Being**	**Autonomy**	**Justice**
Treated organism	[N/A for maize]	[N/A for maize]	[. . .]
Producers (e.g., farmers)	Adequate income and working conditions	Freedom to adopt or not adopt	Fair treatment in trade and law
Consumers	Availability of safe food; acceptability	Respect for consumer choice (e.g., labeling)	Universal affordability of food
Biota	Protection of the biota	Maintenance of biodiversity	Sustainability of biotic populations

adopting the genetically modified maize, early adopting farmers will be benefited by higher income; it is also possible that farmers may suffer health risks from the use of the herbicides to which the maize is now resistant. Each cell, then, can be used to raise a possible benefit(s) or risk(s), or both, for a given stakeholder.

Who uses the matrix? The ethical matrix is designed so that it can easily be used by nonethicists to engage in an ethical analysis of a new technology. It can be used in a participatory workshop setting, by an individual,[33] by a research team,[34] and so on. It is completed prior to the implementation or development of a new technology in order to try to consider the possible consequences for a chosen set of stakeholders. As we will discuss later, we think it can also productively be used after a conversation about a new technology to try to map the concerns that were (and were not) raised—we will call this a "discussion-led ethical matrix." This can be a helpful tool for future design because it can highlight areas of concern that were not foreseen.

The stakeholders of the matrix are chosen by considering who will be impacted by the adoption of the new technology. In putting a stakeholder on the matrix, one commits to treating that group as a moral patient for the purposes of the matrix, that is, as someone or something deserving of moral consideration and respect. This first step in setting up a matrix, then, can itself be a collaborative task, and there may be disagreements about who (or what) deserves consideration. Mepham, for example, argued that, at least for agricultural applications, "the environment" ought to always be a stakeholder, partly because its interests are typically neglected by people developing new agricultural technologies. In order to limit the number of stakeholders on the matrix, one ought to make sure that each stakeholder presents a unique set of interests or concerns on the matrix with respect to the new technology. We argued earlier that the groups to be

judged, scored, or otherwise had decisions made about them are stakeholders of near-term AI, as well as the developers and deployers.

The three ethical principles of the matrix are chosen as prima facie ethical principles, that is, principles that are good rules of thumb for moral actions (but, as rules of thumb, can be disputed or outweighed by other considerations). The ethical principle "to promote well-being" can be understood as a combination of the principles of beneficence (acting to promote stakeholders' interests) and nonmaleficence (acting so as to avoid causing harm). This principle represents what is commonly regarded as a utilitarian approach to ethics. The well-being column requires us to think about how the interests of each stakeholder will be promoted or undermined by the new technology. The ethical principle of "autonomy" or "dignity" should be understood as respect for the freedom of the stakeholder. This principle represents a traditionally deontological ethics, and the column requires us to consider how the new technology will promote or limit the freedom of others as self-directing beings who should be respected as such ("ends in themselves"). Finally, the principle of "justice" should be understood as a respect for "fairness." This is a Rawlsian concept of justice as fairness, and fair institutions and policies can be understood as those that are not significantly responsive to arbitrary differences between individuals and that do not disadvantage those whom we already recognize as being disadvantaged.[35] Thus this column asks us to think about what is selected as a relevant feature for the algorithm to use as a proxy or a metric, how the consequences of the algorithm's use will be distributed across stakeholders, and whether everyone's interests are being given due weight in design. These principles, much like the flexibility of stakeholder groups, can be tailored more specifically to the particular technology at hand.

Here are some ways we might apply each principle when analyzing a near-term AI system. For well-being, we might ask: How will each stakeholder be benefited by the use of the algorithm (beneficence)? How will each stakeholder be harmed or put at risk by the use of the algorithm (nonmaleficence)? Are there alternative methods or processes that are less risky for each stakeholder to achieve the desired outcome? For autonomy, we might instead ask: Will each stakeholder have a choice to use or be a subject of the use of the algorithm? How will each stakeholder be able to determine how the algorithm is used with respect to themselves or their interests? Can each stakeholder meet informed consent conditions with respect to the use of the algorithm (i.e., understand how it works so that they can meaningfully take responsibility for its use, or for its effect on themselves or others)? What are the costs if they cannot meet a suitable standard of informed consent? Finally, for justice: Does the algorithm unfairly favor the interests of one stakeholder without promoting the interests of others,

or by undermining the interests of others? Do false negatives or positives harm or benefit the interests of one group but not others? How and why?

If we consider the FSSA algorithm's initial design, we could represent it on a 1x3 matrix that puts the only stakeholder, the state government of Indiana, against how well the algorithm met the design goals (fraud reduction, efficiency, affordability), with no explicit consideration of how it might meet our common moral standards. If we had plotted it on even a simple ethical matrix with *one* additional stakeholder, the welfare clients, we might *at least* have had the designers consider Table 8.2.

The matrix does not in the end tell us how to design new technologies or whether to implement them, but it does cater to a serious need for ethical reflection. Importantly, it does this by requiring that designers engage in exactly what we think was missing from the FSSA process of design: a consideration of how this new technology will actually impact a wide range of stakeholders. The matrix is *analytical*; it helps us to understand what is at risk and what the benefits are in adopting a technology so that we can then make an informed decision about the design of the new technology. The robustness of this process of analysis, though, is contingent on the ability of those completing the matrix to charitably and sincerely consider the issue from the perspectives of the relevant stakeholder groups. Hence we see particular value in trying to have multiple matrixes completed by different stakeholders who will be affected by the new technology—alternatively, in completing a series of matrixes as new developments and research come to hand. The ethical matrix, then, is a tool that can be used to ensure that algorithmic design is actually informed design with respect to the interests of a wider range of stakeholders. It also establishes a more transparent and accountable design process by requiring the development of an actual artifact of a conversation about what the relevant harms and risks are to stakeholders, that is, the ethical matrix itself.

Table 8.2 FSSA Simple Ethical Matrix

Respect for:	Well-Being	Autonomy	Justice
Indiana State	Cost; stability of community	Transparency with respect to product	Fair treatment of clients, workers; risk of fraud
Clients	Provision of basic necessities	Transparency with respect to own case; informed care	Fair treatment in law; accessibility

8.5. Applying the Ethical Matrix to Data Science

As we noted, ethical reflection can be complicated and divisive. In addition, it can be a challenge to determine the morally relevant aspects of an action, a law, or a new technology on which one should focus. This applies to the use of the ethical matrix as well: just what should one consider about a near-term AI when reflecting on stakeholder autonomy? In this section, we bring together the language of typical data science with the ethical matrix. This provides us with a laundry list of items for the substructure of the cells—the things that we ought always to try to consider when analyzing a near-term AI with an ethical matrix.

Data scientists already have metrics by which they assess the quality of an algorithm; these can serve as a list of topics to consider for the potential benefits and risks of a near-term AI. Here is a basic list of familiar data science concerns (that we put to work in the case studies in the next section): profit, fairness (which, as we've seen, needs to be specifically defined), false positives and negatives, data quality, proxy quality, efficiency, accuracy efficacy, transparency, and consistency.

Our suggestion, then, is that data scientists (or deployers, or whoever is analyzing an algorithm) do not need to receive other ethical education in order to use the matrix. An introduction to the method and commonsense ethical principles of the columns of the matrix is sufficient. They merely need to consider their familiar concerns from the perspective of a wider range of stakeholders. For example, they should consider what the specific interests of a loan applicant might be when it comes to the data quality that they are using to train their algorithm. The Hardt et al.[36] study demonstrates that this can be important for members of racial minorities because missing data on minorities can mean fewer loans end up being given to members of these groups. This is because credit companies tend to care about optimizing accuracy for the dominant group more than for minority groups. However, when there is less accurate information, credit companies tend to err on the side of caution, which means more false negatives (loans won't be given out even though they would be repaid).

In the following subsection, we provide a case study of the COMPAS recidivism risk model to show how the ethical matrix can be used to analyze a near-term AI system. We make a point of indicating how data science concerns can be mapped onto the ethical framework.

8.5.1. COMPAS Recidivism Risk Model

Recidivism risk models were introduced as a way to help judges and parole boards rely less on their own judgment when deciding how long a sentence a defendant should receive or whether to grant parole to an inmate. Although

actuarial instruments have been in use for decades, the more recent versions are likely to be more mathematically sophisticated. In particular, they are likely to be "black box algorithms" that are opaque both to the targets and the deployers, in this case, the courts. Given that these algorithms are used in such high-stakes circumstances, it's crucial that we think in an expanded way about what it means for a recidivism risk model to "work well." The ethical matrix can shed light on this matter.

In 2016 ProPublica published an audit, including data and source code, of the COMPAS recidivism risk model, a black box recidivism risk tool created by the private company NorthPointe and licensed for use by the court system in Broward County, Florida.[37] They found, among other things, that black male defendants were much more likely than white male defendants to receive high-risk scores, and moreover that the false positive rate was about twice as high for black male defendants as for white male defendants. White male defendants, by contrast, had false negative rates twice as high as black male defendants.

Given that higher risk scores are associated with longer sentences, there's a powerful asymmetry here. In particular, false positives might lead to charges of civil rights violations, whereas false negatives will not. On the other hand, false negatives are a major concern for the court itself and for the judges, who might worry that they are in danger of being accused of not being tough enough on criminals if the criminals end up committing more crime, an issue that is more salient in parole hearings than in sentencing. So here we have a list of interests for courts, the judges, and defendants with respect to false negatives.

What happened next is interesting: the company that built and sold COMPAS to the courts, Northpointe, issued a statement in response to the ProPublica audit, which basically said that they don't define fairness via false positives, and that by their definition of fairness, something they called "predictive parity," which basically means race-blind risk measurement, or "accuracy equity," which is to say similar areas under respective ROC curves, their score was fair.[38] Essentially, Northpointe's response was that their algorithm was fair by their definition of fairness—a definition that is, however, not common to all of the stakeholders. Absent a compelling argument for why Northpointe's understanding of fairness is the one that all stakeholders in the COMPAS algorithm ought to use, this is an unconvincing defense for not taking seriously the concerns that other stakeholders have about false positives.

The first problem we can identify in this case study, then, is that the COMPAS algorithm was not designed by taking seriously the interests of stakeholders who would be deeply affected by the algorithm: the defendants. Rather, it was developed only with a concern for the interests of Northpointe. The second is the problem with their choice of predictive parity as the only viable understanding of

"fairness." As we discussed when introducing the Mepham matrix, there is room for negotiating a specific understanding of "fairness" or "justice" when analyzing a technology. The problem in this case is that Northpointe's choice of predictive parity is one that other stakeholders are highly unlikely to agree with because it predictably fails to allow for serious engagement with their interests. That is, anyone trying to sincerely consider the interests of black and white defendants would not choose this understanding of "fairness."

To illustrate why predicative parity is a poor choice, we can consider the known problem around missing crime data. Typically we consider the records collected by police departments and court systems as indicative of crime: arrests, reported crimes, charges, and convictions. However, there's good reason to think that these data in general, and arrests in particular, are not good proxies for crime because of the influence of structural racism. For example, the racial disparity in marijuana-related arrests nationally is about 4 or 5 to 1, even though blacks and whites smoke pot at around the same rates, by their own admission.[39] If we develop a tool to guide arrest rates based on these police records, we will end up with an algorithm that will disproportionately criminalize blacks. That is to say, if we were optimizing to predictive parity across race, we'd actually be asking for a higher rate of arrests among black criminals than among white criminals, at least for marijuana-related offenses.

Table 8.3 is an ethical matrix that plots our list of data science concerns from the previous section, with a consideration of a broader range of stakeholders in the Northpointe algorithm. We have added in some non–data science risks and benefits in italics. Given known problems in crime data with respect to race, it is a reasonable assumption that different racial groups will have different interests, risks, and benefits and thus should be treated as distinct stakeholders. We note that we have the advantage of being able to have a fairly fine-grained consideration of the different stakeholder groups because this matrix has been completed after the design, implementation, and review of the consequences of the algorithm. However, given that these racial disparities are known issues in crime data, it is not unreasonable for this particular list of stakeholders to have been selected before the implementation of the algorithm.

The ethical matrix is a tool for analysis. It allows us to map out where various features of our algorithm, for example, how accurate it is, will give rise to different kinds of concerns for stakeholders. But it can be useful to add an evaluative dimension to the matrix too, as a way of highlighting areas of concern in particular. This evaluative dimension can be achieved only once a matrix has been completed, mapping out the predictable areas of risk and benefits to each stakeholder. The evaluative content requires, of course, actually engaging in ethical reflection on each cell to assign it moral weight. Thus the evaluative use of the matrix is a way of recording the results of the deliberative process of weighing

Table 8.3 COMPAS Simple Ethical Matrix

	Well-Being	Autonomy	Justice
Court	Efficiency, consistency	Transparency; freedom to adopt/ not adopt AI	Efficiency; false negatives; data quality
Black defendants	Maximizes treatment and rehabilitation; minimizes *confinement and punishment*	False positives; transparency	Discriminatory bias, data quality; predictive parity; *respect for civil rights*
White defendants	Maximizes treatment and rehabilitation; minimizes *confinement and punishment*	Transparency	Discriminatory bias, data quality; respect for civil rights
Public	Stability of community	Transparency	False negatives; false positives; data quality; fairness in law
Northpointe	Creative freedom; economic interests	Protection of intellectual property	Predictive parity as fairness; fairness in trade and law

each item for each stakeholder against another, by discounting some concerns (as unlikely or of low significance) and highlighting or selecting others as items to be addressed (as very likely or as very costly to a stakeholder).

We can, for example, color-code the cells of the ethical matrix to provide this evaluative content. So, for example, we can assign white to mean "Don't worry" or "Benefits the stakeholder," light gray to mean "Don't worry too much," and dark gray to mean "Here's where we should worry first." In a given situation that a dark gray cell might translate to, for example, there's a good change that someone's civil rights will be violated ("justice"), whereas in other situations it might simply represent lost opportunity or accuracy. Color codings are by construction overly simplified and are not intended to replace the full, nuanced, and possibly open-ended conversation that each cell represents, but rather a forced "vote" on the status of the ethical consideration by the group of people who are in the conversation. Different people in different conversations could and would draw ethical matrixes with

different color codings, but the result is a more transparent and informed design process.

Here is how we can read off some of the cells on the matrix using the following key to represent our evaluative colors:

- Indicates respect for the principle (white or light gray).
- Indicates infringement of the principle or a negative impact on the stakeholder (light or dark gray).

For reasons of space, we explain in more detail how a few of the cells of the matrix can be interpreted for three of the stakeholders.

Developer Autonomy (Creative Freedom and Economic Interests)

- Northpointe is not subject to unreasonable or burdensome regulations that make producing their product impossible.
- Northpointe can make an economically viable product, and can patent their product to protect it.

Black Defendants' Autonomy (Transparency and False Positives)

- Persons assessed by COMPAS have little to no transparency about the process, the data quality, or how the data are being used in their cases.
- Black defendants, in particular, are at risk of losing their material freedom and self-determination by being wrongly imprisoned as a consequence of high false positive rates.

Black and White Defendants' Well-Being (Maximizing Treatment, Minimizing Punishment)

- Persons assessed by COMPAS who may be in need of help, particularly white defendants, may not be helped to attain it, and people who are not at risk of reoffending, particularly when they are black, are likely to be scored as risky by the algorithm. This has a sum effect of maximizing punishment for many people who do not need it (causing them harm) and failing to provide support for those who do need it (failing to benefit them).

Public Autonomy (Stability of Community)

- COMPAS fails to maintain a safe and flourishing community by keeping persons unlikely to cause crimes in prison, and releasing those likely to cause further crimes.

As noted, this ethical matrix has the benefit of having been written after the development and implementation of the COMPAS tool. Thus we can be confident about some of the risks and benefits in the cells, and we are passing evaluative judgment on the contents of these cells in accordance with critics of Northpointe. Of course, for a new or developing near-term AI these cells will only be our best guess about the benefits and risks of the AI. It can be incredibly useful to complete an ethical matrix after implementation to map out what has been identified as an area of concern. Empty cells can indicate a failure to engage in a sufficiently deep analysis of the consequences of the AI on specific stakeholders.

In the final section, we present a second modified version of the matrix tool to better optimize it for use by data scientists. We show how one can build a discussion-led and evaluative matrix while a conversation unfolds about an algorithm to find areas of specific concern—or areas of contested rich concepts like "fairness." The color coding we introduced here provides this usable and easily digestible method for evaluating the matrix.

8.6. An AI Ethics Tool Data Scientists Can Build Themselves

Given that data scientists and computer scientists are not ethical experts, we think it is reasonable to modify the construction of the matrix in the following ways. In a "data science ethical matrix," we name the columns by the particular metrics familiar to data scientists, and we color-code (for this chapter, on a gray scale) the cells after considering the issue from the perspective of the relevant stakeholder with respect to Mepham's commonsense ethical frameworks. Table 8.4 presents the schematic, although in general, as we will see, there will be more columns.

The point here is that data scientists are more comfortable thinking through the ethics of their familiar metrics than understanding their metrics through the lens of ethical concepts, for instance, thinking through "efficiency" as it pertains to a particular stakeholder, rather than thinking through "justice"

Table 8.4 Example Data Science Ethical Matrix

	Efficiency	**Profit**	**Accuracy**
Stakeholder 1			
Stakeholder 2			
Stakeholder 3			

Table 8.5

	Efficiency	False Positives	False Negatives
Court			
Black Defendants		[redacted]	
White Defendants			

with respect to how the algorithm will affect a stakeholder. When we begin with the familiar data science concepts and require a consideration of how a stakeholder will predictably be affected by design decisions with respect to the data science concept, we start the ethical matrix process on more familiar conceptual terrain. While these amount to the same thing in the end, that is, thinking through the ethics of data science metrics requires using one's ethical concepts, framing the task in the language of data science makes the task more manageable for data scientists.

This flexible framework leads to good news as well as bad news. On the positive side, it allows us to map conversations that have already arisen, such as the conversations outlined in the previous section associated with the COMPAS recidivism model. The ProPublica matrix might look like Table 8.5.

As seen in the matrix, ProPublica's main point was that the black defendants were being unfairly scored higher, as was exposed by the extremely high rate of false negatives. In other words, the corresponding cell is considered at high risk for unethical or unwanted effects and is therefore color-coded dark gray. Similarly, the court is worried about repeat violent offenders being freed through algorithmic error. Since this concern is mitigated by the fact that judges have discretionary power, the cell is colored light gray. On these simplified matrixes that do not have a cell substructure, the choice of light gray will need to be negotiated; we've used it here to indicate "Don't worry too much," but one might use it to indicate "Undecided" or "Too much uncertainty."

The response by Northpointe can similarly be framed by the data science ethical matrix in Table 8.6, coded white because they didn't see any ethical problems, having defined "fairness" differently.

And finally, when data quality is considered, we might have the expanded data science ethical matrix shown in Table 8.7.

Another useful aspect of this construction is that it's easy to locate the trouble spots, whatever one would focus on first. Having said that, it's clear from our discussion that different conversations with different participants would lead to different trouble spots. Even so, it's a useful way of steering a conversation to focus on priorities and ending up with an ethical matrix that

Table 8.6

	Accuracy	Predictive Parity	Accuracy Equity
Court			
Black Defendants			
White Defendants			
Public			
Northpointe			

Table 8.7

	Efficiency	False Positives	False Negatives	Transparency	Predictive Patity	Consistency	Data Quality
Court							
Black Defendants							
White Defendants							
Public							
Northpointe							

makes that conversation transparent and establishes an artifact of accountability of the design decisions.

Now for the drawbacks of this new construction. For example, and probably most important, it's not clear when one has performed a comprehensive job because the simplified matrix doesn't include as much detail. In other words, when does Mepham's original ethical matrix accomplish more than a data science ethical matrix? What are we at risk of leaving out? As a practical tool that will require being presented with a guide, we recommend that users of the matrix ought to begin with Mepham's original list of stakeholders: consumer, producer, the public, and the environment. We should include the environment—think bitcoin applications—and future generations, which would force us to consider long-term feedback loops. Then the design team—ideally consisting at least of developers, deployers, and those scored or evaluated by the algorithm (or a representative for them)—will need to make a decision about what modifications ought to be made to this list.

We might also want to establish a minimum list of metrics to use as columns, including, for example, privacy metrics, transparency, false positives and negatives, and data quality, which was the fatal flaw for the Northpointe discussion. Depending on the context or industry, those lists of concerns would vary. If, for example, due process rights were enforced for recidivism-risk algorithms, transparency would become a required consideration for such algorithms. We also suggest that the Mepham commonsense ethical principles autonomy, well-being, and justice be presented as a guide for more comprehensive and explicit ethical reflection. It's entirely possible that one might think about ethics only as a matter of weighing positive and negative utility, for instance, and leave off rights as an issue of justice with which particular stakeholders might be concerned. Given that one cannot complete the evaluative use of the ethical matrix without making judgments about just whose interests matter and to what degree, requiring an explicit framework of ethical principles that requires a rich reflection on a variety of standard ethical concerns (well-being, self-determination, and rights) is a way to make this process transparent—and of course, more realistic to the interests of the stakeholders.

The simplest explanation of a data science ethical matrix is that it's an artifact of a conversation related to a specific algorithm used in a specific context. For our last example, we will map Virginia Eubanks's case study of the Allegheny Family Screening Tool (AFST) model, which predicts child abuse, taken from chapter 4 of *Automating Inequality*.[40] This will certainly not contain all the details she provides, but it should give an overview of the issues that are raised in the chapter using a data science ethical matrix.

The AFST algorithm was built and is used by the Allegheny County Office of Children, Youth, and Families (CYF) in Pittsburgh, Pennsylvania, to determine which of the many calls on the child abuse hotline should be followed up with a caseworker. That immediately gives us the following stakeholders: the CYF office, the parents or caregivers of children who may be suspected of abuse, and the children who are at risk of abuse. We may also want to consider the people who made the call concerning suspected abuse, since that group's perception of the system's working or being flawed will probably contribute to its success. We also might want to include the public at large, since the AFST algorithm is intended to protect the safety of children in the community, and we might want to further split these categories, depending on how the conversation proceeds.

Eubanks goes on to point out that the algorithm will act differently on those families that have had higher interaction with the social safety net, such as homeless families and families already in the foster care system. In particular, the more data that are available about a given family, the more likely it is that their score will be sufficiently high to warrant follow-up.[41] That implies we might want to differentiate between "high-touch families" and "low-touch families." It

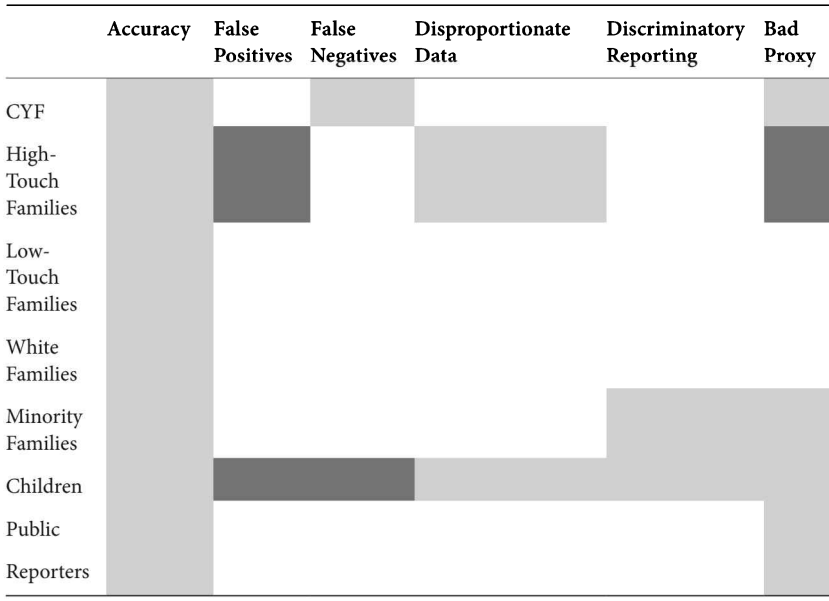
might be easier to simply define "high touch" to mean that a certain threshold of data has been exceeded, or a proxy of low versus high income might make more sense, depending on what information we actually know about the families themselves. The associated concern in the matrix might be characterized as "disproportionate data availability." When we do this, we should keep in mind that one reason there might be disproportionate data availability is that there is a long history of actual abuse, for example. A primary data-driven goal will be to try to distinguish between that "true signal" of abuse and incidental data collection; a secondary data-driven goal, which is famously difficult, would be to try to measure the missing data associated with families that are well-off and that have been historically outside of the system.

Next, Eubanks points out that the way the calls come into the hotline are racialized in general. This suggests that we should further distinguish by race among stakeholders, and that an associated concern would be "discrimination in reporting." We want to make clear here that if the design team had been required to complete an ethical matrix, this would require a process of determining the predictable consequences of the algorithm. In order to do this, one would need to consider the data quality, existing patterns or trends in the population that may be relevant to metric selection, and many other statistical facts relevant to the problem to be solved. That is, proper use of an ethical matrix requires empirical research into what we know about a problem already, and thus is likely to turn up evidence of biased data sets or sampling problems with the target population that the algorithm will be used with. We make these comments to establish that a number of these consequences of the actual design of the AFST algorithm that we can see after implementation may have been predictable had the design team been required to consider them.

Eubanks also makes a salient point about the choice of definition of success for the model; namely, the model doesn't actually train on "substantiated abuse" events but rather either the removal of a child from his or her home or follow-up calls to the hotline. The latter type of event is, once again, known to be racialized, and the former is known to happen to parents who are simply too poor to provide their children with common comforts. In general, according to ambiguous laws, it's difficult to know when to call something neglect and when to simply describe it as poverty. The end result is that we should certainly add a column of concern entitled something along the lines of "target variable imperfect proxy for substantiated abuse," or "bad proxy" for short, and keep in mind that both poor families and minority families are likely to have a bigger problem with this particular column than richer white families.

Altogether we now have a data science matrix that looks something like Table 8.8.

Table 8.8



8.7. Summary

In many cases, the very problems that near-term AIs are being deployed to solve are moral problems: informing decisions about parole and imprisonment, helping to decide who gets loans, determining who is eligible for welfare. Given this fact, it is surprising that ethical reflection is not the norm in algorithm design. In the infamous Trolley thought experiment, an uncontrollable train travels down the tracks on its way to run over and kill five people. You as the observer have the option to divert this train onto an alternative track, where it will instead kill just one person. Some people argue that they can absolve themselves of the situation by refusing to decide whether or not to pull the lever. Perhaps some people developing near-term AIs take this to be their own situation and that their nonengagement in explicit ethical decision-making in AI design absolves them of any responsibility for the consequences that the algorithm has on the lives of the stakeholders assessed, evaluated, or scored by the system. Of course, whether one can really be passive in the Trolley scenario is controversial. We've argued here that design decisions about algorithms are moral decisions; thus we have in a way denied that one can remain on passive or neutral moral ground in AI design and implementation.

As our discussions of a range of case studies show, near-term AI is being implemented in ways that have serious ethical consequences for many persons simply because their interests are not being taken into consideration in the design of these algorithms. We think that this is not only harmful but that it constitutes a widespread failure to meet our ethical obligations not to cause harm to others—a sentiment that is widely regarded as a commonsense ethical platitude. The ethical matrix and the data science ethical matrix are practical tools that can help to fulfill this much-needed space for serious ethical reflection, and we think something of this sort ought to be required of near-term AI design teams. The immediate, and significant, advantage of adopting this tool is that it forces us to consider how our choices affect a range of stakeholders wider than those of the designer and the implementer. Furthermore, it does so in a way that makes AI design processes more transparent to the stakeholders in general, and it assists in making accountability more transparent as well. Even this small step will be a significant improvement for the field of near-term AI.

Notes

1. Ben Mepham, "A Framework for the Ethical Analysis of Novel Foods: The Ethical Matrix," *Journal of Agricultural and Environmental Ethics* 12, no. 2 (2000): 165–76, doi:10.1023/a:1009542714497.
2. On the risks of AI superintelligence and value alignment, see N. Bostrom, "Ethical Issues in Artificial Intelligence," In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, vol. 2, ed. I. Smit et al., 12–17 (International Institute of Advanced Studies in Systems Research and Cybernetics, 2003). On the moral decision-making of autonomous weapons systems, see D. Purves, R. Jenkins, and B. J. Strawser, "Autonomous Machines, Moral Judgment, and Acting for the Right Reasons," *Ethical Theory and Moral Practice* 18, no. 4 (2015): 851–72.
3. We are, in a sense, defining AI by its functional role in our lives—as those automated systems that are typically black boxes to those evaluated by them, and that are included in (or just function as) the decision-making procedure. One could define AI by another metric, for example, the complexity of the processing of the system, whether it uses certain decision-making procedures, whether it is composed of neural networks, and so on. But our concern is with the adoption of algorithmic solutions to decision-making in a wide range of areas.
4. As we will discuss in the coming paragraphs, an automated system adopted in Indiana gave the blanket notice "Failure to cooperate in establishing eligibility" to all persons whose benefits were denied or taken from them.
5. Though it's worth noting, too, that the environment and nonhumans may also have important stakes in the adoption and design of these algorithms. Our scope here will

be to focus primarily on the harms to persons; we do note in later sections where nonpersons might be taken on as stakeholders.

6. Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York: St. Martin's Press, 2018), ch. 2.
7. A reduction in welfare rolls seems to be serving as a proxy for an improvement in the material well-being of those who had depended on them, and also as a proxy for a more efficient (cost-effective) welfare system.
8. Eubanks, *Automating Inequality*, 48.
9. Ibid., 63.
10. As stated in Article 25 of the Universal Declaration of Human Rights, "Everyone has the right to a standard of living adequate for the health and well-being of himself and of his family, including food, clothing, housing and medical care and necessary social services, and the right to security in the event of unemployment, sickness, disability, widowhood, old age or other lack of livelihood in circumstances beyond his control."
11. Eubanks, *Automating Inequality*, 68.
12. Ibid., 62.
13. In 2010 Indiana actually sued IBM for a breach of contract for the high rates of benefit denials to citizens in need. The judge found in favor of IBM, finding that the company had met the goals laid out by the design brief: "The heart of the contract remained intact throughout the project" (ibid., 75). As Eubanks describes the situation, and we strongly agree, "The problem . . . was not that the IBM/ACS coalition failed to deliver, it was that the state and its private partners refused to anticipate or address the system's human needs" (75).
14. Overt ableist discrimination would require that someone intentionally designed the system so that it did not cater to the needs of the blind or deaf. We take it that the design failure was an oversight that arose from not considering differently abled persons' needs. For a discussion of "overt" and "institutional" racism that informs this distinction on overt and institutional ableism, see G. Ezorsky. *Racism and Justice: The Case for Affirmative Action* (Ithaca, NY: Cornell University Press, 1996), ch. 1.
15. Our thanks to Michael P. Lynch for raising this point.
16. To be clear, serious academic research outside of this realm of well-defined success, and toward the project of aligning machine values to human values, is underway and exploding in numbers, conferences, and academic output. See, for example, the FAT conferences mentioned earlier, which are dedicated to issues of fairness, accountability, and transparency in machine learning and other automated systems.
17. Academic researchers tend to be given leeway when it comes to how they define success than, say, a data scientist working at a social media company, where the definition of success is likely strictly tied to business interests and held there by corporate lawyers.
18. This even has a name when applied to business metrics: Goodhart's Law. And even though it's widely known, it's still in effect.
19. C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Broadway Books, 2017), ch. 3.

20. Moritz Hardt, Eric Price, and Nathan Srebro, "Equality of Opportunity in Supervised Learning," *arXiv*, October 7, 2016, https://arxiv.org/abs/1610.02413v1. See also O'Neil, *Weapons of Math Destruction*, ch. 8.
21. See, for example, Mehrsa Baradaran, *The Color of Money: Black Banks and the Racial Wealth Gap* (Cambridge, MA: Belknap Press of Harvard University Press, 2017), for a detailed discussion of the history of the racial wealth gap in the United States.
22. The thought being that by commonsense or intuitive uses of "fair play," this is an unfair situation.
23. Hardt, Price, and Srebro, "Equality of Opportunity in Supervised Learning," 17.
24. The first scoring system, "Maximum Profit," explicitly states that it has no fairness constraint. As we will show, however, despite this, it achieves a result with an intuitive sense of fairness better than one of the constraints that seeks to establish a technical definition of fairness as "racially neutral treatment."
25. "We . . . consider the behavior of a lender who makes money on default rates below this, i.e., for whom false positives (giving loans to people that default on any account) is 82/18 as expensive as false negatives (not giving a loan to people that don't default). The lender thus wants to construct a predictor $\hat{y}$ that is optimal with respect to this asymmetric loss." Hardt, Price, and Srebro, "Equality of Opportunity in Supervised Learning," 17.
26. Although to some extent followed up by Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt, "Delayed Impact of Fair Machine Learning," *arXiv*, April 7, 2018, https://arxiv.org/abs/1803.04383.
27. Mepham, "A Framework for the Ethical Analysis of Novel Foods."
28. D. Schroeder and C. Palmer, "Technology Assessment and the 'Ethical Matrix,'" *Poiesis & Praxis* 1, no. 4 (2003): 295–307, doi:10.1007/s10202-003-0027-4.
29. For a more recent presentation of the matrix and a discussion of the ethical principles used, see Ben Mepham, "Ethical Principles and the Ethical Matrix," In *Practical Ethics for Food Professionals*, ed. J. Peter Clark and Christopher Ritson (Wiley Online, June 7, 2013), 39–56, doi:10.1002/9781118506394.ch3. C. Kermisch and C. Depaus, "The Strength of Ethical Matrixes as a Tool for Normative Analysis Related to Technological Choices: The Case of Geological Disposal for Radioactive Waste," *Science and Engineering Ethics* 24, no. 1 (2017): 29–48, doi:10.1007/s11948-017-9882-6, assess the utility of the ethical matrix framework, which is typically a collective activity, for individual researchers. K. K. Jensen et al., "Facilitating Ethical Reflection among Scientists Using the Ethical Matrix," *Science and Engineering Ethics* 17, no. 3 (2010): 425–45, doi:10.1007/s11948-010-9218-2, present results that show the ethical matrix successfully works as a tool for promoting ethical reflection among scientists, and in particular on the needs of external stakeholders.
30. (Rawls 2000, 167). In brief and leaving aside a lot of interesting detail, the veil of ignorance is a thought experiment that asks us to imagine that we have been tasked with making political decisions about how to structure and regulate our society. However, we must make these decisions imagining that we have no knowledge of who we are within our society: our social or economic status, interests, religion, level

of education, talents, and so on. The thought is that without selfish reasons to guide us, we are more likely to decide on principles and structures that are truly just or fair.

31. T. Beauchamp and J. F. Childress, *Principles of Biomedical Ethics,* 7th ed. (Oxford University Press, 2012).
32. Mepham, "A Framework for the Ethical Analysis of Novel Foods," 170.
33. See, for example, Kermisch and Depaus, "The Strength of Ethical Matrixes as a Tool for Normative Analysis Related to Technological Choices."
34. See, for example, Jensen et al., "Facilitating Ethical Reflection among Scientists Using the Ethical Matrix."
35. Our "intuitive fairness" concept from section 8.2.1 can be understood more precisely with this Rawlsian notion.
36. Hardt, Price, and Srebro, "Equality of Opportunity in Supervised Learning."
37. Julia Angwin et al., "Machine Bias," *ProPublica*, May 23, 2016, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
38. William Dieterich, Christina Mendoza, and Tim Brennan, "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity," Northpointe Inc. Research Department, July 8, 2016, https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html. The area under ROC (receiver operating characteristic) curves measures the extent to which, as we vary thresholds, the model maintains a better-than-random true positive to false positive ratios. Thresholds are the limit values between the "high-risk" category and the "low-risk" category. Note that, once you are actually using an algorithm, you have of course *chosen one* threshold, so you don't actually care about how accurate the model would have been with differently chosen thresholds, just how accurate your model is with the specific threshold you've chosen. This gives you a bit of a flavor as to why this area is mostly irrelevant to the person deploying the model, never mind the target of the model who is worried that they, specifically, represent a false positive or a false negative.
39. Angwin et al., "Machine Bias."
40. Eubanks, *Automating Inequality*.
41. Ibid., 146, 155.

References

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." *ProPublica*, May 23, 2016. www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Baradaran, Mehrsa. (2017). *The Color of Money: Black Banks and the Racial Wealth Gap*. Cambridge, MA: Belknap Press of Harvard University Press, 2017.

Beauchamp, T., and J. F. Childress. J. F. (2012). *Principles of Biomedical Ethics.* 7th ed. Oxford University Press,Newyork 2012.

Bostrom, Nick. (2003). "Ethical Issues in Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, vol. 2,

edited by I. Smit George Eric Lasker, Wendell Wallach, Iva Smit., 12–17. International Institute of Advanced Studies in Systems Research and Cybernetics,Windsor 2003.

Dieterich, William, Christina Mendoza, and Tim Brennan. "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity." *Northpointe Inc. Research Department*, July 8, 2016. https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html.

Eubanks, Virginia. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* New York: St. Martin's Press, 2018.

Ezorsky, G. (1996). *Racism and Justice: The Case for Affirmative Action. Ithaca, NY*: Cornell University Press, 1996.

Hardt, Moritz, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning." *arXiv*, October 7, 2016. https://arxiv.org/abs/1610.02413v1.

Jensen, K. K., E. Forsberg, C. Gamborg, K. Millar, and P. Sandøe. "Facilitating Ethical Reflection among Scientists Using the Ethical Matrix." *Science and Engineering Ethics* 17, no. 3 (2010): 425–45. doi:10.1007/s11948-010-9218-2.

Kermisch, C., and C. Depaus. "The Strength of Ethical Matrixes as a Tool for Normative Analysis Related to Technological Choices: The Case of Geological Disposal for Radioactive Waste." *Science and Engineering Ethics* 24, no. 1 (2017): 29–48. doi:10.1007/s11948-017-9882-6.

Liu, Lydia T., Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. "Delayed Impact of Fair Machine Learning." *arXiv*, April 7, 2018. https://arxiv.org/abs/1803.04383.

Mepham, Ben. "Ethical Principles and the Ethical Matrix." In *Practical Ethics for Food Professionals*, edited by J. Peter Clark and Christopher Ritson, 39–56. Wiley Online, June 7, 2013. doi:10.1002/9781118506394.ch3.

Mepham, Ben. "A Framework for the Ethical Analysis of Novel Foods: The Ethical Matrix." *Journal of Agricultural and Environmental Ethics* 12, no. 2 (2000): 165–76. doi:10.1023/a:1009542714497.

Mepham, Ben, M. Kaiser, E. Thortensen, S. Tomkins, and K. Millar. "Ethical Matrix Manual: Agricultural and Forest Meteorology." *Agricultural and Forest Meteorology* (January 2006). https://www.researchgate.net/publication/254833030_Ethical_Matrix_Manual.

O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York: Broadway Books, 2017.

Purves, D., R. Jenkins, and B. J. Strawser. "Autonomous Machines, Moral Judgment, and Acting for the Right Reasons." *Ethical Theory and Moral Practice* 18, no. 4 (2015): 851–72.

Schroeder, D., and C. Palmer. "Technology Assessment and the 'Ethical Matrix.'" *Poiesis & Praxis* 1, no. 4 (2003): 295–307. doi:10.1007/s10202-003-0027-4.