Algorithmic Bias and Corporate Responsibility: How companies hide behind the false veil of the technological imperative

Kirsten Martin, PhD

University of Notre Dame

Forthcoming in *Ethics of Data and Analytics,* Taylor & Francis

In this chapter, I argue that acknowledging the value-laden biases of algorithms as inscribed in design allows us to identify the associated responsibility of corporations that design, develop, and deploy algorithms. Put another way, claiming algorithms are neutral or that the design decisions of computer scientists are neutral obscures the morally important decisions of computer and data scientists. I focus on the implications of making technological imperative arguments: framing algorithms as evolving under their own inertia, as providing more efficient, accurate decisions, and as outside the realm of any critical examination or moral evaluation. I argue specifically that judging AI on efficiency and pretending algorithms are inscrutable produces a veil of the technological imperative which shields corporations from being held accountable for the value-laden decisions made in the design, development and deployment of algorithms. While there is always more to be researched and understood, we know quite a lot about testing algorithms. I then outline how the development of algorithms should be critically examined to elucidate the value-laden biases encoded in design and development. The moral examination of AI pierces the (false) veil of the technological imperative.

Introduction

Consider a recent example of companies using AI:

> An Amazon driver "spent almost four years racing around Phoenix delivering packages as a contract driver for Amazon.com Inc. Then one day, he received an automated email. The algorithms tracking him had decided he wasn't doing his job properly." Drivers understood that their performance was being monitored – how they drove their route, where they put packages on the porch, etc – and would sometimes receive emails with a rating from fantastic to at risk. However, "Amazon knew delegating work to machines would lead to mistakes and damaging headlines, these former managers said, but decided it was cheaper to trust the algorithms than pay people to investigate mistaken firings so long as the drivers could be replaced easily."[1]

[1] Soper, Spencer. 2021. "Fired by Bot at Amazon: 'It's You Against the Machine'. Bloomberg. June 28, 2021. https://www.bloomberg.com/news/features/2021-06-28/fired-by-bot-amazon-turns-to-machine-managers-and-workers-are-losing-out

For the drivers, they were fired by an algorithm. However, the discussion about whether algorithms *does* things or *has* a bias emanates from a concern as to who is then responsible for good and bad outcomes. In other words, when an organization or individual uses an algorithm, who is responsible for the decisions? The arguments traditionally fall into two camps: those that focus on the algorithm as the actor that 'does' things and is at fault (technological determinists) and those that focus on the users of that algorithm as determining the outcome (social determinists).

However, in this chapter, I argue that we need to acknowledge the value-laden biases of technology – including algorithms – while preserving the ability of humans to control the design, development, and deployment of technology.[2] This is important so that we appropriately attribute responsibility for how algorithms perform and better interrogate those that design and develop algorithms to augment our decisions. In other words, only by acknowledging the value-laden biases of algorithms can we begin to ask how companies inscribed those biases during design and development.

While the general argument that technology has biases and that individuals and companies design those biases is not new (Bijker, 1995; Friedman & Nissenbaum, 1996; D. G. Johnson, 2015; Latour, 1992; Winner, 1980), here I make more explicit the implications for corporate accountability. In this chapter, I also examine the implications of making technological imperative arguments – framing algorithms as evolving under their own inertia, providing more efficient, accurate decisions, and outside the realm of interrogation. I argue specifically that judging AI on efficiency and pretending algorithms are inscrutable produces a veil of the technological imperative which shields corporations from being held accountable for the value-laden decisions made in the design, development and deployment of algorithms. Importantly, claims of algorithms being inscrutable and efficient provide a shield for corporations who are these making value-laden decisions. Finally, I offer how AI and algorithms should be interrogated given what we know currently about the value-laden decisions of computer and data scientists.

[2] Biases are value-laden design features with moral implications in use.

AI and Value-laden Bias. What does that mean?

Determinist Arguments

For *social determinists*, society is the main actor of the story in constructing technology and determining the outcome. If an algorithm is not performing correctly, e.g., violating rules, undermining rights, harming others, producing unjust results, etc, then a social determinist would point to the many ways that people created that technology and then decided how it would be used. For social determinists, what matters is not technology itself but the social or economic system in which it is embedded. In regards to algorithms, social determinists blame the *use* of the program rather than the design of the program or acknowledge that the *data* may be flawed ("it's just the data") and that society needs to get better data for the algorithm. For the medical triage algorithm, one would focus on the disparities that existed in the data as a reflection of historical discrimination and not focus on the decisions around the design of the algorithm (and the decision to use that data!).

One attraction of arguing that algorithms are neutral and society is to blame is the ability to avoid any form of technological determinism: in attributing values or biases to algorithms, scholars are concerned we would also attribute control to technology and thereby remove the ability of society to influence technology. Even further, identifying any form of materiality in algorithms could lead to a form of worship, where an algorithms' preferences are deemed unassailable and humans are left subservient to the whims of the algorithm (Desai & Kroll, 2017).

For *technological determinists*, technology is the primary actor of the story. As such, technology is to 'blame' for the outcome. Strident technological determinists frequently see technology as having an internal dynamic that leads the best technology to survive in the market. This faction argues that the ethical evaluation of technology is not appropriate since it may curtail development. For data analytics, we hear technological determinist arguments when the algorithm or program is the main actor in the paragraph or the sentence. For example, "The algorithm decided…" or "the program categorized…" For example, the algorithm decided to fire Amazon's drivers.

Figure 1, adopted from Martin and Freeman (2004), shows two versions of determinism at either end of a spectrum where social determinism means framing technology as a blank slate, neutral, and socially controlled and where technological determinism means framing technology as value-laden and controlling society.

Social Determinism
Technology as a blank slate;
Neutral and socially controlled
Technological Determinism
Technology as determining society;
Value-laden and outside society's control

Figure 1: Traditional approaches to technology (adopted from Martin and Freeman 2004).

Determinists and (a lack of) Accountability

This tension – between social determinists and technological determinists – is not merely an academic exercise but is important to assigning responsibility because who is 'acting' is normally who we look to hold responsible for those acts. For social determinists, an algorithm is neutral, a blank slate, and society is then responsible for how the algorithm is used. Algorithms-as-a-blank-slate would suggest minimal responsibility for the developers who craft the algorithm (Martin & Freeman, 2004). For technological determinists, algorithms *do* things, but these inherent biases are then outside the influence of society, designers, and developers. The algorithm-as-autonomous-agent narrative suggests the users have no say or accountability in how algorithms make decisions.

This false tension – algorithms as objective, neutral blank slates versus deterministic, autonomous agents – has implications for whether and how firms are responsible for the algorithms they develop, sell, and use. Both mistakenly absolve developers – computer scientists, data analysts, and corporations – of their responsibility. Whether you hold the users of the algorithm responsible (social determinism) or the algorithm itself (technological determinism), you are not holding responsible the systems of power -- the government or company -- that designed, developed, and implemented the program (Martin, 2022b).

This deterministic conversation about algorithms absolves firms of responsibility for the development or use of algorithms. Developers argue that their algorithms are neutral and thrust into fallible contexts of biased data and improper use by society. Users claim algorithms are difficult to identify let alone understand, therefore excluding users of any culpability for the ethical implications in use (Martin, 2019).

Algorithmic Biases

Reality "is a far messier mix of technical and human curating" (Dwork & Mulligan, 2013, p. 35). Those who fall into these deterministic arguments conflate two ideas: whether or not a technology is value-laden and who controls the technology. Martin and Freeman argue these two mechanisms are independent and see technology as simultaneously value-laden yet under social control (Martin & Freeman, 2004), where one need not claim technology as neutral to maintain control over it. Similarly, and focused on algorithms, Mittelstadt et al note that algorithms are value-laden with biases that are "specified by developers and configured by users with desired outcomes in mind that privilege some values and interests over others"(2016). This approach acknowledges the materiality of algorithms without allocating control to the technology. Figure 2, adopted from Martin and Freeman (2004), illustrates how one can acknowledge the materiality of technology in that algorithms can have value-laden biases, discriminate, create and destroy value, cause harm, respect or violate ethical norms, diminish rights, reinforce virtues, etc without forgoing control over technology.

In other words, these deterministic arguments are making two separate assumptions that are not required to move in tandem (Martin & Freeman, 2004). First, whether a technology is value-laden can be separated from who controls the technology as in Figure 2. While technological determinists assume value-laden technology with technology also being autonomous, and social determinists assume a neutral technology with society in control, many scholars have embraced technology as value-laden while acknowledging the control of society in the design, development, and deployment of algorithms.

		Technology as …	
		Neutral	**Value-laden w/Moral Implications**
Who is in control	**Technology**	II. Technological Imperative *Efficient AI that should not be questioned*	I. Technological Determinism *Biased AI that is beyond our control*
	Society	III. Social Determinism *Efficient AI corrupted by society's messiness.*	IV. Value-Laden Biases *Biased AI with value-laden decisions in design, development, deployment*

Figure 2 Technology's bias and control.

These tensions and assumptions about technology are not new and many of the theories attempting to break free of deterministic approaches dealt with trains, bikes, dams, scallops, door groomers, etc. In other words, AI and algorithms is not the first time we have questioned how technology is value-laden while we also are responsibility for the design, development and deployment of technology. For example, Wiebe Bijker explores bicycles, lightbulbs, and plastics (Bijker, 1995); Latour examines seatbelts and door groomers (Latour, 1992); Winner uses bridges, tomato harvesters, and ships (Winner, 2010).

Importantly, these authors and others acknowledge the materiality of algorithms has biases that are value laden (Friedman & Nissenbaum, 1996; Johnson, 2004) or have politics (Winner, 1980) while also identifying how individuals, corporations, and society control that same technology. In this way, Latour notes that technology – including algorithms – is anthropomorphic: "first, it has been made by humans; second, it substitutes for the actions of people and is a delegate that permanently occupies the position of a human; and third, it shapes human action by prescribing back" what humans should do (p. 160). These scholars both identify the materiality of technology while also maintaining the responsibility of individuals who design, develop, and use algorithms. In fact, obliterating the materiality of technology, treating algorithms as if they are a blank-slate and value-neutral, absolves computer scientists and companies of their moral decisions in developing the algorithm.

We can think of algorithms as having biases in three ways. First, algorithms *are value-laden* in that algorithms are biased and designed for a preferred set of actions. Algorithms create moral consequences, reinforce or undercut ethical principles, and enable or diminish stakeholder rights and dignity (Martin, 2019). However, we should broaden this to include algorithms as fair, just, abiding by virtue ethics (Vallor, 2016), expressing or violating an ethics of care (Villegas-Galaviz, 2022), reinforcing racist systems and policies (Benjamin, 2019; Eubanks, 2018; Gebru, 2019; Poole et al., 2020), reinforcing or undermining misogyny (D'Ignazio & Klein, 2020). Figure 3 illustrates the value-laden-ness of AI systems, including the development of the specific algorithm, as well as the types of ethical issues we find around outcomes, algorithms, and data.

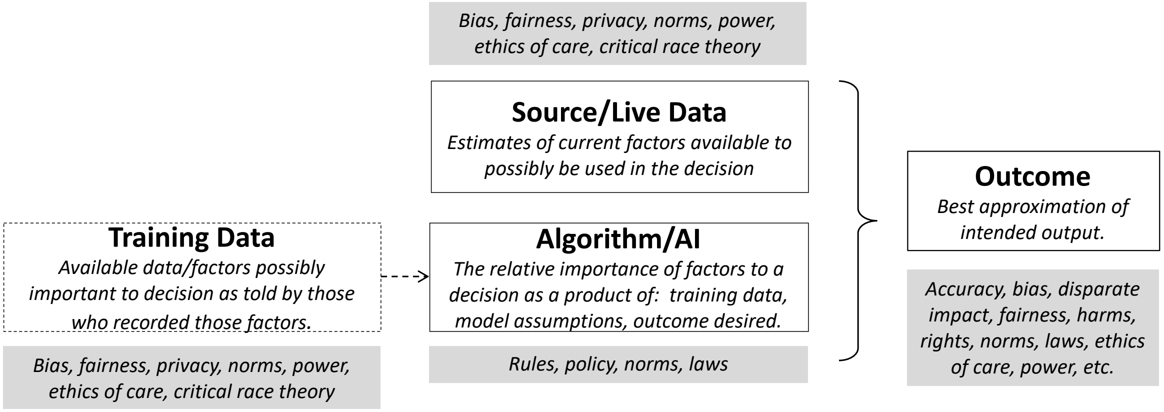
Bias, fairness, privacy, norms, power, ethics of care, critical race theory
Source/Live Data
Estimates of current factors available to possibly be used in the decision
Outcome
Best approximation of intended output.
Training Data
Available data/factors possibly important to decision as told by those who recorded those factors.
Algorithm/AI
The relative importance of factors to a decision as a product of: training data, model assumptions, outcome desired.
Accuracy, bias, disparate impact, fairness, harms, rights, norms, laws, ethics of care, power, etc.
Bias, fairness, privacy, norms, power, ethics of care, critical race theory
Rules, policy, norms, laws

Figure 3 Value-laden-ness of AI

Second, algorithms can be seen as *policy at scale.* In other words, in creating the algorithm, developers are taking a stand on ethical issues and "expressing a view on how things ought to be or not to be, or what is good or bad, or desirable or undesirable" (Kraemer et al., 2011, p. 252). Algorithms act like design-based regulation (Yeung 2016) where algorithms can be used for the consistent application of legal and regulatory regimes (Thornton, 2016, p. 1826). Algorithms can enforce morality while still being designed and used by individuals (Diakopoulos, 2013). In the case of automated decision making, algorithms combine adjudication with rule making in setting policy (Citron, 2007). For Amazon, imagine developing a policy for drivers that each warehouse manager had to implement that included criteria for their performance and, if necessary, criteria and process for their termination or promotion. Instead, Amazon implements policy through how the AI is developed, and then makes decisions on behalf of the company.

STS scholar Madeleine Akrich suggests the following thought experiment which is of particular importance for algorithms:

> "…How can the prescriptions encoded in the mechanism be brought out in words? By replacing them by strings of sentences (often in the imperative) that are uttered (silently and continuously) by the mechanisms for the benefit of those who are mechanized: do this, do that, behave this way, don't go that way, you may do so, be allowed to go there. Such sentences look very much like a programming language."

As Latour rightly summarizes, "every time you want to know what a nonhuman does, simply imagine what other humans or other nonhumans would have to do were this character not present" (p. 155). The technology's script answers who matters, which group is important, who counts, which race/ethnicity is included and delineated. What factors should be considered, what is the outcome that is important, how 'bad' are outliers, what is the criteria for working.

Finally, *algorithms influence who-does-what* in a decision system. In addition to the design of value-laden algorithms, developers make a moral choice as to the delegation of who-does-what between algorithms and individuals when in use**.** At a minimum, technologies alleviate the need for others to do a task. In Latour's classic case of seatbelt, making the seat belt automatic – attaching the seatbelt to the door so that it is in place automatically – relieves the driver from the responsibility to ensure the seatbelt is used. A bell may be added to remind the driver. (Martin, 2019). However, when developers design the algorithm to be used in a decision, they also design how accountability is delegated within the decision. Sometimes algorithms *are designed* to absorb the work and associated responsibility by precluding users from taking on roles and responsibilities within the decision system. Inscrutable algorithms that are designed to minimize the role of individuals in the decision take on more accountability for the decision.

Technological Imperative

I turn now to revisit the quadrant II in Figure 2, not yet addressed, where algorithms are framed as value-neutral and outside the control of society. The technological imperative frames technologies as almost inevitable and outside our control (See Chandler, 2012) -- a technological determinist who also believes that the technology always is correct.

According to the technological imperative, algorithms are detached from the imperfections of the 'real world.' You hear this when computer and data scientists believe their work does not include dealing with the imperfections and ambiguity of the world or that algorithms are merely efficient, accurate machines that are better than alternatives. Algorithms, according to the technological imperative, are not worthy of being questioned not only because any imperfections arise from the messiness of the world but also because algorithms are seen as inscrutable and not subject to interrogation.

Importantly, for the technological imperative, technology should be adopted *without much question*.[3] This extreme view is becoming *more* common in how we talk about machine learning in the news and in research. For example, in an argument against researchers who have highlighted the dangers of using artificial intelligence and predictive analytics without regard to their biases or moral implications, Alex Miller, in "Want Less-Biased Decisions? Use Algorithms," lists the ways AI *could* be an improvement because humans are bad at decisions (true – we are not great[4]). His argument joins a common refrain that technology, because it can be an improvement if designed properly, is then always improvement (Miller, 2018).

Algorithms in quadrant II are not critically examined both (1) because they are deemed to be inscrutable but also (2) because there is no point since they are framed to be neutral – seen as efficient and accurate.

[3] Hans Jonas, in "Toward a Philosophy of Technology" (1979), envisions a restless technology moving forward under the pressure of competition.

[4] For example, we continue to face racism in lending decisions and bias against women and minorities in hiring decisions. (Kessler et al., 2019; Moss-Racusin et al., 2012; Perry, 2019; Quadlin, 2018) (Kessler et al found " employers hiring in STEM fields penalized résumés with minority or female names. The effect was big: These candidates were penalized by the equivalent of 0.25 GPA points, based solely on the name at the top of the résumé. That meant such a candidate needed a 4.0 GPA to get the same rating as a white male with a 3.75." https://www.latimes.com/opinion/story/2020-07-24/employment-hiring-bias-racism-resumes?_amp=true

These two assumptions reinforce the false idea of the technological imperative in our current conversations about algorithms. First, claims that algorithms are inscrutable frame algorithms as so complicated that they are impossible to explain or question, *even by computer and data scientists* (!) (Barocas et al., 2013; Desai & Kroll, 2017; Introna, 2016; Ziewitz, 2016). In fact, algorithms are seen as so difficult to explain that assigning responsibility to the developer or the user is deemed inefficient and even impossible. Previously I have argued that the Inscrutable Defense ("It's too complicated to explain") does not absolve a firm from responsibility, otherwise firms would have an incentive to create complicated systems to avoid accountability (Martin, 2019). Here I am arguing that claiming that algorithms are inscrutable *even to computer and data scientists* produces a veil behind which companies can hide the value-laden judgements made in the design and development of algorithms.

Figure 4 illustrates all that is hidden when we falsely claim that algorithms are inscrutable and cannot be critically examined. Claims of inscrutably allow computer and data scientists to make value-laden decisions without being questioned from those inside and outside the organization.

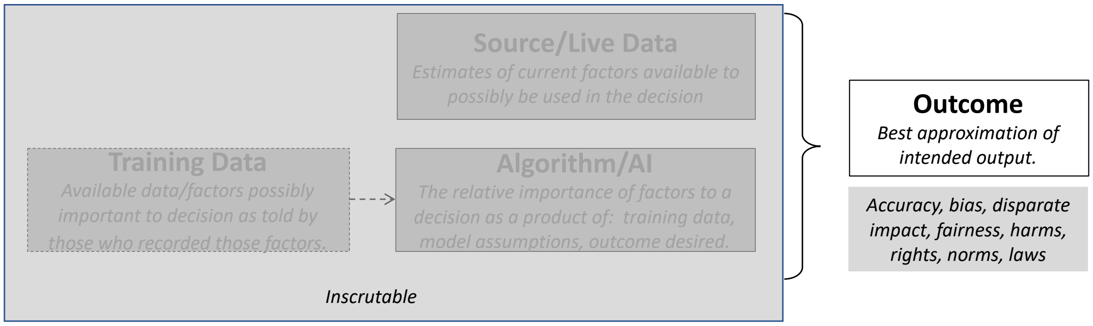
Source/Live Data
Estimates of current factors available to possibly be used in the decision
Training Data
Available data/factors possibly important to decision as told by those who recorded those factors.
Algorithm/AI
The relative importance of factors to a decision as a product of: training data, model assumptions, outcome desired.
Outcome
Best approximation of intended output.
Accuracy, bias, disparate impact, fairness, harms, rights, norms, laws
Inscrutable

Figure 4 Inscrutable algorithms with value-laden outcomes.

In an essay, "The Fallacy of Inscrutability," Joshua Kroll argues that the common argument that algorithms are inscrutable or black boxes is a fallacy. In fact, Kroll argues that "inscrutability is not a result of technical complexity but rather of power dynamics in the choice of how to use those tools" (Kroll, 2018). A company may want to claim a program is inscrutable even if it is not (Pasquale, 2015). The company may want to protect a program they believe is a competitive advantage: this could be because it *is* a competitive advantage or due to shear inertia (they always like to say things are a competitive advantage). The organization also may not like the answer they would need to provide, or the people asked may not know enough to answer. Importantly, Professor Kroll argues that not understanding a system is just bad practice. As Kroll summarizes: "rather than discounting systems which cause bad outcomes as fundamentally inscrutable and therefore uncontrollable, we should simply label the application of inadequate technology what it is: malpractice, committed by a system's controller" (Kroll, 2018, p. 5).

The second assumption that feeds the false narrative of the technological impetrative is when algorithms are framed as more accurate or more efficient without interrogating the program. Further, that efficiency and accuracy are neutral concepts *and outside ethical considerations*. In an examination of top machine learning and AI conference papers, Birhane et al find the dominant values to be "performance, accuracy, state-of-the-art (SOTA), quantitative results, generalization, efficiency, building on previous work, and novelty" (Birhane et al., 2021).[5] Current attempts to include any ethical consideration of the development and impact of algorithms in research has similarly been met with claims that the moral evaluation of algorithms are outside the scope of the field.[6] This approach has infected management as well. For example, "machine learning is significantly faster and seemingly unconstrained by human cognitive limitations and inflexibility" (Balasubramanian et al., 2020). Even Google's recent introduction of a new large language model for search was focused on performance without any mention of all the additional outcomes of such models (Bender et al., 2021).

Figure 5 illustrates the focus on the efficiency of outcomes. While value-laden decisions may occur in design and development, the deployment of the algorithms is assumed to be better than alternatives in terms of efficiency and accuracy. Further efficiency and accuracy are assumed to be value-neutral designations and *not subject to moral evaluation*.

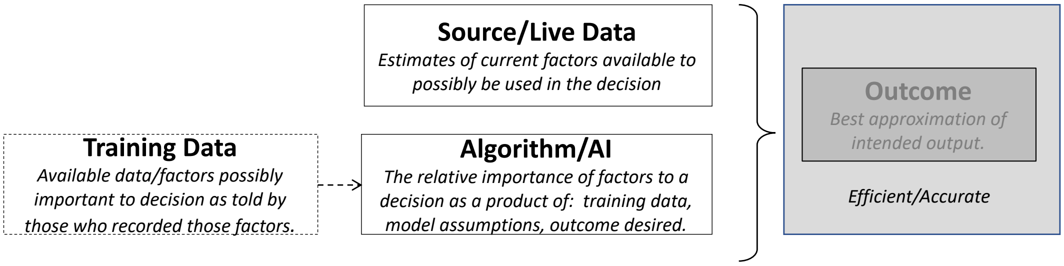
Source/Live Data
Estimates of current factors available to possibly be used in the decision
Training Data
Available data/factors possibly important to decision as told by those who recorded those factors.
Algorithm/AI
The relative importance of factors to a decision as a product of: training data, model assumptions, outcome desired.
Outcome
Best approximation of intended output.
Efficient/Accurate

Figure 5 Value-laden algorithms with efficient and accurate outcomes.

[5] Further, these values are defined as when they benefit the corporation deploying the AI program by default. In other words, efficiency is defined as that which is efficient for one actor, the corporation deploying the algorithm, but not the efficiency for all actors and overall.

[6] In other words, computer science engineering, as opposed to any other engineering discipline, would not be required to critically examine and identify the morally implications of their research projects. Castelvecchi, Davide. 2020. "Prestigious AI meeting takes steps to improve ethics of research". *Nature*. https://www.nature.com/articles/d41586-020-03611-8 E.g., https://www.geekwire.com/2020/retired-uw-computer-science-professor-embroiled-twitter-spat-ai-ethics-cancel-culture/

This assumption is actually two mini-assumptions in one: (a) assuming that AI is efficient and (b) therefore neutral. First, one must ask “efficient for whom?” The program may be efficient for the organization based on a narrow metric that the decision is made faster than a human. However, the implementation may create work for subjects of the AI program and make the process longer, less efficient for those that fix problems, hear appeals, investigate problems. Second, efficiency is one metric but not the entire measurement of whether an AI decision ‘works.’ Organizations implement programs and decisions and policies with more goals than ‘efficiency: fairness, treating people with dignity, creating long term value for stakeholders, following rules, etc.

Finally, Gabbrielle Johnson makes a compelling case that the concepts of efficiency, simplicity, and accuracy are value-laden. Her argument is longer, and worth reading, that alternatives such as novelty, complexity of interaction, diffusion of power, etc are also compelling criteria for AI (and science more generally) to be judged. Importantly, we are making value judgments when we decide to judge AI using efficiency as a criteria and alternatives exist. So, for Johnson, even if one uses efficiency as a criteria, we cannot claim that is in any way ‘neutral.’(Johnson, n.d.)

Figure 6 combines these two assumptions ---- efficient outcomes and inscrutable algorithms – which serve to hide the value-laden biases, decisions, and outcomes in the design and development of algorithms and halt the moral evaluation of algorithms. Figure 6 illustrates how we currently are unnecessarily and inappropriately constructing the technological imperative in our approach to algorithms.

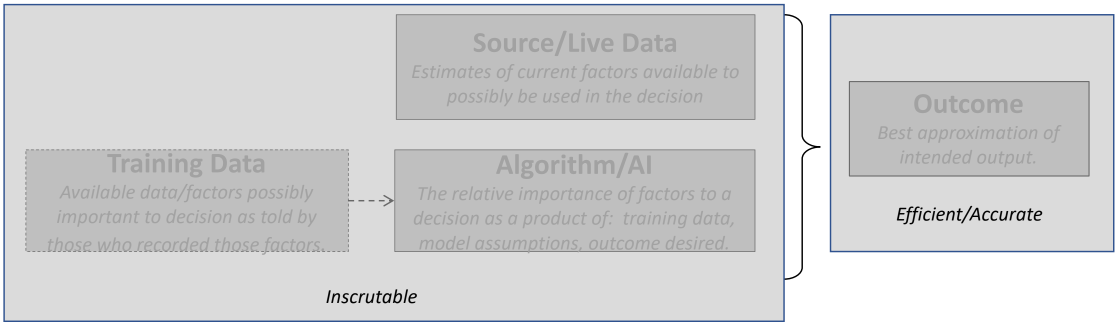
Source/Live Data
Estimates of current factors available to possibly be used in the decision
Training Data
Available data/factors possibly important to decision as told by those who recorded those factors.
Algorithm/AI
The relative importance of factors to a decision as a product of: training data, model assumptions, outcome desired.
Inscrutable
Outcome
Best approximation of intended output.
Efficient/Accurate

Figure 6: Constructing the technological imperative – inscrutable algorithms creating efficient outcomes.

Critical Evaluation of Algorithmic Bias.

Acknowledging the value-laden biases of algorithms – inscribed in design and development and with moral implications in use – opens the door to questioning how and why algorithms are developed and used. Algorithms are developed with design choices that "can affect the relative distribution of power, authority, privilege in a community" (Winner, 1980). Friedman and Nissenbaum offer a framework to identify sources of bias in computer systems, where bias is defined as the tendency of a computer system to "systematically and unfairly discriminate" against certain individuals. Friedman and Nissenbaum identify three types of biases based on how the bias emerges: Amazon's firing AI program could have biases (a) preexisting in the data, then (b) embedded in the design of the algorithm, and (c) emergent in how the program is then deployed on live data. Similarly, Barocas and Selbst outline how algorithms have a discriminatory disparate impact through design choices such as defining the target variable and adding labels, choosing training data, selecting features, including proxies for protected classes, or purposefully designing an algorithm to have a specific discriminatory bias (Barocas & Selbst, 2016).

I extend this conversation on the value-laden decisions of computer and data scientists focused on the development of the algorithm. In terms of Friedman and Nissenbaum, I focus here on how biases are embedded during design and development of algorithms. Or, as Johnson says, "values are constitutive of the very operation of algorithmic decision-making" (Johnson, n.d.). As noted by Kearns and Roth in *The Ethical Algorithm,* "The designer may have had a good understanding of the algorithm that was used to *find* the decision-making model, but not the model itself." In other worlds, computer and data scientists do understand the assumptions used to create the machine learning algorithms based on training data even if the model created is harder to explain. Kearns and Roth suggest to ensure that these models respect societal norms, we need to learn how to design with these ethical norms in mind. For Rudner, we need to better manage value-laden decisions of computer and data scientists by first acknowledging that these value-laden decisions are occurring, otherwise we will make them unconsciously, haphazardly, and "leave an essential aspect of scientific method scientifically out of control" (1953, p. 6). Here, I turn to offer a framework to critically examine the development of algorithms to illuminate the value-laden biased designed in the AI program.

1. Moral Evaluation of Outcomes

The first step of any program is to create an outcome variable that measures something they are interested in – but the outcome is actually constructed based on the phenomenon of interest (what we are worried about or interested in the world) as well as what is easily measured and labeled by a program (Martin, 2022a). In addition, *the* outcome variable chosen has implications as to what the organization thinks is important and whose interests are prioritized in the design of the algorithm. How the outcome is constructed and how changing the outcome impacts how the algorithm is developed are important assumptions to critically examine Some questions to interrogate the outcome variable chosen:

- What is the phenomenon we are trying to represent with this outcome?
- How was this outcome chosen? What other outcomes did you consider and what was the impact?
- For whom does this outcome variable represent the phenomenon of interest and for whom does this outcome *not* represent the phenomenon of interest?
- For whom is value created and who is disempowered in choosing this outcome?

2. Moral Evaluation of the Criteria for Whether the Algorithms Works.

The criteria for an algorithm 'working' is usually chosen by the company developing the algorithm. This is handy (for the developer or researcher!) because the criteria chosen for success can be tailored to the algorithm in question. Importantly, whether or not the use of an algorithm is an improvement should be critically evaluated. One measure, accuracy, is used frequently: the percent of cheaters caught with facial recognition or the percent of patients correctly identified as high priority. In fact, much has been researched and explained as to measuring accuracy, or more importantly, how to judge that an algorithm *works*. However, the percent of true positives identified is only one measure. For Amazon's firing algorithm, we would also care about the number of false positive (falsely identified as a bad driver when they are not) and whether the accuracy rates are consistent across protected classes or type of subjects.[7] Some questions for critically evaluating whether an algorithm works:

- What measurements did you use to test if the algorithm worked? Which did you decide not to use or report?

[7] Protected classes are designations protected in certain laws from discrimination: race, ethnicity, gender, nationality, religion, etc. However, an HR algorithm used to read resumes may have a true positive rate for people from Indiana that is better than the true positive rate for people from South Dakota. We would find that unfair, unethical, and inappropriate even though being from South Dakota is not a protected status.

- For whom is this algorithm accurate? For whom is it not accurate?
- What is the rate of all mistakes (false positives, false negatives)? Is the rate of mistake consistent across demographic groups?
- What measurements for fairness did you use? What benchmarks did you use to compare the algorithm to?
- What is the impact on effectiveness or efficiency for the subjects of the algorithm?
- How was the model tested for overfitting to the training data? How did you ensure the algorithm was not optimized for data in training but less useful for live data?
- What are the effects on individuals' rights with the use of this algorithm? Are any ethical norms violated with the use of this algorithm?
- Who benefits from the use of this algorithm? Are they the designers of the algorithm?

3. Moral Evaluation of Data Choices

The ethical questions about both training data and the data used when an algorithm is deployed can be thought of as covering (a) whether the data is appropriate to use (or not) *regardless* of the results and (b) whether the use of data is appropriate to use (or not) because of the associated results. In regards to the former, training data could be gathered in violation to users' privacy (Barocas & Nissenbaum, 2014) or the mere use of the data would be considered unfair for that particular context (Barocas & Selbst, 2016) or the use of that data may be in violation of a law or norm (Gebru et al., 2018). In each case, the use of the data would be inappropriate not matter how 'well' the algorithms works. In addition, the use of data may lead to adverse outcomes for subjects including disparate impact and harms. The computer and data scientist decides what data is appropriate to use and makes the value-judgment not only as to the use of a data set, but also which factors in the data set to include. *Not all data should be used in training a model* and the use of data is a judgment call. We should be concerned if a computer or data scientist did not exclude some data from the creation of the model. Some questions for critically evaluating data choices:

- What data or factors did you not include because doing so would be a privacy violation, unfair, or inappropriate for this decision?
- What data or factors did you not include because doing so resulted in outcomes that were a privacy violation, unfair, or inappropriate?

4. Moral Evaluation of Assumptions in Developing the Model

Perhaps the least discussed portion of the value-laden decisions in development is around the assumptions that computer and data scientists make in the design of algorithms. Consider an example from Kearns and Roth (2019), an algorithm is developed based on historical university data to identify acceptances based on students' SAT and GPAs. The algorithm falsely rejects qualified black applicants more often than qualified white applicants. "Why? Because the designer didn't anticipate it. She didn't tell the algorithm to try to equalize the false rejection rates between the two groups so it didn't. …machine learning won't give you anything 'for free'" (p. 10).

In general, each assumption that a computer scientist makes about the data – how it is distributed, what missing data means and should be treated, how much to punish outliers or those that do not 'fit' the algorithm – is an assumption about the subjects in the data. When outliers are treated as not a big deal, the computer scientists are saying *it is not morally important if a person is not adequately represented by this algorithm.* When missing data is ignored, the computer scientists are saying *it is ethically appropriate for individuals to be punished when they are not well covered by this data set.* For example, when an algorithms was used to predict whether consumers will pay back credit-card debt, the predictions skewed to favor wealthier white applicants because "minority and low-income groups have less data in their credit histories" (Heaven, 2021). The lack of data just made the prediction less precise, and the lack of precision downgraded the scores. Similarly with outliers: computer scientists need to decide whether outliers are 'punished' in the development of the algorithm. Some modeling assumptions square the distance to the outlier while others just take the absolute value (others cube the distance, etc). Squaring the distance to the outlier means that the performance is 'hurt' by more outliers that are further out. Importantly, while computer and data scientists make these assumptions *about outliers in the data*, these outliers are *about people represented in the data* that are not well characterized by the algorithm as it is being developed.

Work has been done to better understand fairness in terms of the design, development, and testing of algorithms (Barocas et al., 2018; Chouldechova & Roth, 2018; Hardt, 2014; Mitchell et al., 2020; Narayanan, 2018). Here, I broaden the types of questions to include the general moral evaluation of the many assumptions computer and data scientists make:

- Data: What assumptions were made about how the data is distributed? What does that assumption mean in terms of people?
- Missing Data: What assumptions were made about *missing* data and how missing data should be treated? Should people be punished for not being well represented in this particular data set?

- Outliers: How are outliers defined? Which data points are *well characterized* by the algorithm and which are *not recognized* by the algorithm? What assumptions are made about the distribution of the outliers (e.g., is it assumed to be random)?
- Outliers: How are outliers treated in the development of the algorithm? How are large versus small outliers treated? Is it morally important to not have large outliers? What assumptions are made and what happens when different assumptions are made?
- Fitting: How do you test the performance of the algorithm? Do you have results for training, validation, and testing data? (This checks to see if the developer overfit the algorithm to the training data which would make it less useful with live data).

5. Moral Evaluation of Plans for Resiliency

Computer and data scientists should ensure algorithms support good decisions – including managing the inevitable mistakes. This requires developers to *expect* mistakes to occur and designing algorithms should include planning the ability to identify, judge, and correct mistakes. In this way, computer and data scientists are designing for the resiliency of algorithms (Martin & Parmar, Forthcoming.). While mistakes may be unintentional, ignoring or even fostering mistakes is unethical. Ethical algorithms plan for the identification, judgment, and correction of mistakes, whereas unethical programs allow mistakes to go unnoticed and perpetuate mistakes. In interrogating an algorithm, one would examine if the program (a) created mistakes and if the type of mistakes was appropriate for the decision context and (b) allows for the identification, judgment, and correction of mistakes:

- How was this algorithm designed to identify mistakes in use?
- How was this algorithm designed to fix mistakes in use?

Conclusion

In this article, I argue that acknowledging the value-laden biases of algorithms as inscribed in design allows us to identify the associated responsibility of corporations that design, develop, and deploy algorithms. Claiming algorithms are neutral or that the design decisions of computer scientists are neutral obscures the morally important decisions of computer and data scientists. I also examine the danger in framing algorithms as evolving under their own inertia, providing more efficient, accurate decisions, and outside the realm of interrogation. I argue specifically that judging AI on efficiency and pretending algorithms are inscrutable produces a veil of the technological imperative which shields corporations from being held accountable for the value-laden decisions made in the design, development and deployment of algorithms. While there is always more to be researched and understood, we know quite a lot about the value-laden decisions in development and how to morally evaluate algorithmic biases.

REFERENCES

Balasubramanian, N., Ye, Y., & Xu, M. (2020). Substituting human decision-making with machine learning: Implications for organizational learning. *Academy of Management Review*, *ja*.

Barocas, S., Hardt, M., & Narayanan, A. (2018). *Fairness and machine learning: Limitations and Opportunities*.

Barocas, S., Hood, S., & Ziewitz, M. (2013). *Governing algorithms: A provocation piece*. http://dx.doi.org/10.2139/ssrn.2245322

Barocas, S., & Nissenbaum, H. (2014). Big data's end run around anonymity and consent. In J. Lane, V. Stodden, S. Bender, & H. Nissenbaum (Eds.), *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. Cambridge University Press.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* 🦜. 610–623.

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons.

Bijker, W. (1995). *Of bicycles, bakelite, and bulbs: Towards a theory of sociological change*. MIT Press.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2021). *The Values Encoded in Machine Learning Research*. https://arxiv.org/abs/2106.15590

Chandler, J. A. (2012). "Obligatory Technologies" Explaining Why People Feel Compelled to Use Certain Technologies. *Bulletin of Science, Technology & Society*, *32*(4), 255–264.

Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *ArXiv Preprint ArXiv:1810.08810*.

Citron, D. K. (2007). Technological due process. *Washington University Law Review*, *85*, 1249.

Desai, D. R., & Kroll, J. A. (2017). Trust But Verify: A Guide to Algorithms and the Law. *Harvard Journal of Law and Technology*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2959472

Diakopoulos, N. (2013, August 2). Sex, Violence, and Autocomplete Algorithms. *Slate*. http://www.slate.com/articles/technology/future_tense/2013/08/words_banned_from_bing_and_google_s_autocomplete_algorithms.html

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. Mit Press.

Dwork, C., & Mulligan, D. K. (2013). It's not privacy, and it's not fair. *Stan. L. Rev. Online*, *66*, 35.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin’s Press.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330–347.

Gebru, T. (2019). Oxford Handbook on AI Ethics Book Chapter on Race and Gender. *ArXiv Preprint ArXiv:1908.06165*.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *ArXiv Preprint ArXiv:1803.09010*.

Hardt, M. (2014, September 26). *How big data is unfair: Understanding unintended sources of unfairness in data driven decision making*. Medium. https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de#.lyynedan6

Heaven, W. D. (2021, June 17). Bias isn’t the only problem with credit scores—And no, AI can’t help. *MIT Technology Review*. https://www.technologyreview.com/2021/06/17/1026519/racial-bias-noisy-data-credit-scores-mortgage-loans-fairness-machine-learning/

Introna, L. D. (2016). Algorithms, governance, and governmentality: On governing academic writing. *Science, Technology, & Human Values*, *41*(1), 17–49.

Johnson, D. G. (2004). Is the Global Information Infrastructure a Democratic Technology? *Readings in Cyberethics*, *18*, 121.

Johnson, D. G. (2015). Technology with No Human Responsibility? *Journal of Business Ethics*, *127*(4), 707.

Johnson, G. (n.d.). Are algorithms value-free? Feminist theoretical virtues in machine learning. *Journal Moral Philosophy*.

Kearns, M., & Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Kessler, J. B., Low, C., & Sullivan, C. D. (2019). Incentivized resume rating: Eliciting employer preferences without deception. *American Economic Review*, *109*(11), 3713–3744.

Kraemer, F., Van Overveld, K., & Peterson, M. (2011). Is there an ethics of algorithms? *Ethics and Information Technology*, *13*(3), 251–260.

Kroll, J. A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180084.

Latour, B. (1992). Where are the Missing Masses? The Sociology of a Few Mundane Artifacts. In W. Bijker & J. Law (Eds.), *Shaping Technology/Building Society: Studies in Sociotechnical Change* (pp. 225–258). MIT Press.

Martin, K. (2019). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics*, *160*(4), 835–850.

Martin, K. (2022a). Creating Accuracy and Predictive Analytics. In *Ethics of Data and Analytics*. Taylor & Francis.

Martin, K. (2022b). Value-laden Biases in Data Analytics. In *Ethics of Data and Analytics*. Taylor & Francis.

Martin, K., & Freeman, R. E. (2004). The separation of technology and ethics in business ethics. *Journal of Business Ethics*, *53*(4), 353–364.

Martin, K., & Parmar, B. (Forthcoming.). Designing Ethical AI: Anticipating ethical lapses and building for resilience. *Sloan Management Review*.

Miller, A. (2018). What Less-Biased Decisions? Use Algorithms. *Harvard Business Review*.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2020). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, *8*.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, *3*(2), 1–21.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*(41), 16474–16479.

Narayanan, A. (2018). *Translation tutorial: 21 fairness definitions and their politics*. *2*(3), 6–2.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Perry, V. G. (2019). A loan at last? Race and racism in mortgage lending. In *Race in the Marketplace* (pp. 173–192). Springer.

Poole, S., Grier, S., Thomas, F., Sobande, F., Ekpo, A., Torres, L., Addington, L., Henderson, G., & Weekes-Laidlow, M. (2020). Operationalizing critical race theory (CRT) in the marketplace. *Journal of Public Policy and Marketing*.

Quadlin, N. (2018). The mark of a woman's record: Gender and academic performance in hiring. *American Sociological Review*, *83*(2), 331–360.

Thornton, J. (2016). Cost, Accuracy, and Subjective Fairness in Legal Information Technology: A Response to Technological Due Process Critics. *New York University Law Review*, *91*, 1821–1949.

Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.

Villegas-Galaviz, C. (2022). Ethics Of Care As Moral Grounding For AI. In K. Martin (Ed.), *Ethics of Data and Analytics*. Taylor & Francis.

Winner, L. (1980). Do artifacts have politics? *Daedalus*, *109*(1), 121–136.

Winner, L. (2010). *The whale and the reactor: A search for limits in an age of high technology*. University of Chicago Press.

Ziewitz, M. (2016). Governing Algorithms Myth, Mess, and Methods. *Science, Technology & Human Values*, *41*(1), 3–16.