Public Trust, Institutional Legitimacy, and the Use of Algorithms in Criminal Justice

When contempt and mistrust too often characterize public attitudes toward lawful authority, all—young and old, private citizens and public officials—suffer the consequences.
–Lyndon B. Johnson[1]

Abstract:

A common criticism of the use of algorithms in criminal justice is that algorithms and their determinations are in some sense 'opaque'—that is, difficult or impossible to understand, whether because of their complexity or because of intellectual property protections. Scholars have noted some key problems with opacity, including that opacity can mask unfair treatment and threaten public accountability. In this paper, we explore a different but related concern with algorithmic opacity, which centers on the role of public trust in grounding the legitimacy of criminal justice institutions. We argue that algorithmic opacity threatens the trustworthiness of criminal justice institutions, which in turn threatens their legitimacy. We first offer an account of institutional trustworthiness before showing how opacity threatens to undermine an institution's trustworthiness. We then explore how threats to trustworthiness affect institutional legitimacy. Finally, we offer some policy recommendations to mitigate the threat to trustworthiness posed by the opacity problem.

1: Introduction

In 2013, Eric Loomis was convicted of eluding the police, a charge stemming from his role in a drive-by shooting incident in La Crosse, Wisconsin. During sentencing, the judge consulted a computer program called the Correctional Offender Management Profile for Alternative Sanctions (usually called by its acronym "COMPAS") , which calculates the risk that a given defendant will reoffend. COMPAS found that Loomis presented a high risk of reoffending. Having consulted this finding, the judge sentenced Loomis to six years in prison. Loomis appealed his sentence on the grounds that because COMPAS's methodology for generating risk assessments is a trade secret, using COMPAS violated his right to due process. In particular, Loomis claimed that he was denied his right to an individualized sentence and to be sentenced by an accurate and reliable decision-making process (*State v. Loomis* 2016). Loomis's appeal made its way to the Wisconsin Supreme Court, which upheld his sentence.

[1] Lyndon B. Johnson, "Special Message to the Congress on Crime in America," February 6, 1967, as quoted in Matt Stroud. 2019. *Thin Blue Lie: The Failure of High-TechPolicing* Metropolitan Books: New York.

Courts and law enforcement agencies across the U.S., as well as certain other countries, employ predictive algorithms for a variety of purposes, including risk assessment in sentencing, predicting locations where crime is likely to occur, and identifying those likely to commit criminal acts in the near future. Much of the focus on the use of predictive algorithms has concerned their legality. This, after all, was the basis of Loomis's appeal to the Wisconsin Supreme Court. But predictive algorithms have also faced ethical criticism as well. The two most widespread criticisms are that the algorithms are biased against or unfair to people of color (Barocas and Selbst 2016; Lum and Isaac 2016; Selbst 2017) and that the algorithmic predictions are for various reasons "opaque," unintelligible, or incomprehensible to human subjects (Burrell 2016; O'Neil 2016). There is no consensus definition of algorithmic opacity, but we understand an algorithm to count as opaque to some degree just in case (a) facts about the contribution of any single feature of the world to the algorithm's final determination cannot be easily accessed, either by the human decision-maker or by persons affected by the determination, or (b) the way that the algorithmic determination figures in decision-making cannot be easily understood. We are not here concerned with providing a complete conceptual analysis of ALGORITHMIC OPACITY. We offer this definition because it is attractively broad, because it allows for diverse sources of opacity (several of which we outline below in Section 3), and because it is consonant with the definitions used by other scholars in the field (de Laat 2018; Watson and Floridi 2020). A person might be ignorant of the contribution that a feature of the world makes to an algorithmic system's determinations because of the system's innate complexity, because the person lacks specialist knowledge, or because the source code is protected as intellectual property as a trade secret. And even when an individual knows how an algorithmic system works, they might remain ignorant of the role its determinations play in institutional decision-making if that role is not disclosed by decision-makers. We are interested in all of these sources of opacity, and our definition captures all of them.

Two moral and legal concerns with opacity are well established. Opacity can mask unfair treatment, and it can threaten public accountability. If we cannot know *how* some decision about a defendant's pretrial detention has been made, because the algorithm's mechanism for generating a risk score is opaque, then we cannot know whether we were treated unfairly in the rendering of that decision (Barocas and Selbst 2016). It is also challenging to hold the algorithm or its designers "accountable" for its determinations if an explanation is impossible to secure (Vestby and Vestby 2019; Wachter et al. 2018).

In this paper we raise a novel moral concern with algorithmic opacity, one that arises from the relationship between algorithmic opacity and the trustworthiness of criminal justice institutions. In particular, we argue that **algorithmic opacity can undermine the trustworthiness of criminal justice institutions, which threatens their legitimacy**. This problem can persist even when predictive algorithms satisfy reasonable standards of fairness and when they are more accurate than humans in the same domain.

Section 2 offers a discussion of the nature of trust and trustworthiness, both as these concepts figure in interpersonal relationships and as they figure in the relationship between citizens and the institution of law enforcement. We offer an account of institutional trustworthiness according to which trustworthy institutions are competent to do what they are entrusted with doing, but they are also arranged *so as to ensure* that the institution will function as it has been entrusted to do. We further suggest that a fully trustworthy institution will *signal* its competence and responsiveness to the needs of those who count on it.

In section 3 we describe several sources of algorithmic opacity, andin section 4 we argue that algorithmic opacity undermines the trustworthiness of criminal justice institutions along several dimensions. In section 5 we show that a threat to trustworthiness is morally problematic because there is a morally significant link between trust in and the trustworthiness of a law enforcement institution and its legitimacy: as trust and trustworthiness erodes, that institution's claim to legitimacy weakens. Algorithmic opacity, then, threatens the very legitimacy of the institution of law enforcement.[2]

2: Trust and Trustworthiness

Trust and distrust are attitudes, whereas trustworthiness is the property that makes an attitude of trust rational or fitting. Typically, trustworthiness is a property attributed to individuals. We might trust our spouses, our teachers, our priests, or our physicians, and we trust them because we believe they are trustworthy. In developing an account of trust and trustworthiness that applies to the relationship between individuals and institutions, it can therefore be instructive to first explore plausible accounts of trust and trustworthiness concerning relationships between individuals.

[2] It is worth flagging here that an apparent implication of our view is that an institution can be in fact trustworthy, and yet citizens do not trust it (for whatever reason), and therefore its legitimacy is undermined. Thus, the same institution could be both trustworthy and illegitimate. Some might find this to be an odd conclusion; however, we think this is correct, and offer the resources for understanding why in section 5 below..

Annette Baier, in her influential article "Trust and Antitrust" characterizes trust between individuals as accepted dependence on the good will of another. She writes, "When I trust another, I depend on her good will toward me…Trust then, on this first approximation, is accepted vulnerability to another's possible but not expected ill will (or lack of good will) toward one" (Baier 1986, p. 235). Much more can be said to unpack Baier's characterization of trust, but it is plausible enough on its face. What is most insightful about Baier's analysis is that trust or distrust are attitudes that figure in relationships of vulnerability on the part of one individual to the choices of another. But Baier does not provide conditions that make trust rational or fitting. Those conditions are given by an account of trustworthiness.

Karen Jones, expanding on Annette Baier's earlier work, develops a "three-place" account of trustworthiness.

> **Three-place trustworthiness***:* B is trustworthy with respect to A in domain of interaction D, if and only if she is**competent** with respect to that domain, and she would take the fact that **A is counting on her**, were A to do so in this domain, **to be a compelling reason for acting** as counted on (Jones 2012, pp. 70–71).[3]

On Jones's account, when one person counts on--i.e., accepts their vulnerability to--another person in some domain, the party being counted on is trustworthy if and only if they are competent to do what they are being counted on to do and they possess certain motivations to act as counted on. Returning to Baier's initial quote, counting on or depending on someone in a domain entails that one is vulnerable to the individual's choices in that domain. A professor's students count on them to grade their work impartially. This does not entail that students trust their professors to grade impartially. One might be unhappily vulnerable to someone's choices if they do not trust them. Trustworthiness requires competence. A teacher would be a less trustworthy evaluator of their students' work in a philosophy course if their graduate degree was in an unrelated field, because they would lack competence to do the job. And finally Jones thinks that being trustworthy requires being motivated to act as counted on by the fact that one is being counted on to act that way. Trustworthiness is therefore incompatible with, say, a person acting as counted on merely out of fear of punishment. A person would not be a fully trustworthy teacher if their motivation to grade their students' work impartially was that they feared being caught. If, on the other hand, one is motivated

[3] Emphasis added.

by good will toward one's students or conscientiousness concerning their pedagogical duties, they are a more trustworthy evaluator. What all trustworthy motivational states share, according to Jones, is that they involve seeing the fact of someone else's dependence on you as a direct and compelling reason to act on their behalf. Something like Jones's account is widely accepted among trust theorists.[4]

Jones provides an attractive set of necessary and sufficient conditions for trustworthiness in reciprocal interpersonal relationships (e.g. friends, romantic partners), certain fiduciary or advisory roles (e.g. personal advisors, spiritual leaders), and so forth. But does it make sense to apply an account of interpersonal trustworthiness to institutions, including the criminal justice institutions that are the focus of this paper? To make progress on this question, we will consider whether it is sensible to devise an institutional analogue of each of Jones's three conditions—the idea of counting on another, the competence requirement, and the motivational requirement—to criminal justice institutions.

2.1 The Dependence Requirement

For Jones's account to be suitable for application to the relationship between citizens and criminal justice institutions, there must first be some domain in which citizens "count on" or depend on those institutions. What do citizens count on law enforcement and the courts to do?

It will suffice here to note that there is at least one domain in which citizens count on criminal justice institutions in virtue of those institutions' possessing a monopoly on the domestic use of force. This monopoly comprises a number of different methods, from the use of lethal force in threatening scenarios to the restriction of liberty via detainment, compelled presence and disclosure, and incarceration. We assume for the purposes of the paper that this monopoly on force, and the methods that comprise it, exist for the purpose of promoting peace, security, and justice.[5] But whatever its purpose, the fact that this monopoly exists means that citizens count on criminal justice institutions for peace, security, and justice; they have no feasible option other than to rely on

[4] Influential trust scholar Russell Hardin and coauthors defend a similar account (Cook et al. 2005).

[5] By "justice" we mean at least equal treatment of citizens regardless of one's constitutionally protected characteristics like race, sex, and religion. Aside from this, we do not assume any substantive account of justice. By making this assumption we are therefore setting aside recent structural criticisms of law enforcement according to which the function of law enforcement in the United States is to surveil and oppress poor people of color. If that criticism is sound, we are confident that our account of trustworthiness will yield the correct conclusion that American law enforcement is not trustworthy, because it is not responsive to the needs of citizens who are counting on it.

criminal justice institutions to secure those ends on their behalf. Taking security into one's own hands is not a feasible alternative to the rule of law, administered by the state.

2.2 The Competence Requirement

The competence requirement of Jones's account of trustworthiness also seems to be applicable to institutions. Criminal justice institutions can be more or less competent at promoting the ends of peace, security, and justice. The competence of our criminal justice institutions at achieving these aims will vary along a number of dimensions, including the efficacy of investigational tools, the soundness of strategic decision-making, the talent of the individuals acting in their official capacity on behalf of the institutions, the efficacy of mechanisms for ensuring compliance with institutional norms, the reliability of the procedures for the gathering and evaluation of evidence, and institutional superiority with respect to the threat of force. Some will be tempted by a deflationary explanation of institutional competence according to which the competence of the institution is nothing more than the competence of its officials. We resist this view on the grounds that some of the dimensions of institutional competence do not lend themselves to the deflationary explanation. For example, institutional mechanisms designed to ensure compliance with institutional norms and policies can be more or less effective at ensuring compliance. In one of the most shocking examples of law enforcement corruption in the past 40 years, The Los Angeles Police Department's Rampart scandal involved as many as 70 sworn officers being implicated in various illegal activities including stealing and selling cocaine from evidence lockers. A key finding of the LAPD's internal inquiry into the scandal was that inadequate supervisory mechanisms existed to ensure compliance with departmental policies (Parks 2000). The Rampart Scandal was not a failure of *individual* competence but rather a failure of the *institutional* arrangements for ensuring that officers acted as they were being counted on to act. The inadequacy of these institutional arrangements would have undermined the LAPD's institutional competence even if the Rampart scandal had not occurred.

2.3 The Motivational Requirement

The motivational requirement for three-place trustworthiness poses the greatest challenge to adapting Jones's account of interpersonal trustworthiness to institutions. If being trustworthy requires taking the fact of another's dependence to be a *reason* to act as counted on, then being trustworthy seems to require the possession of a capacity to respond to facts as reasons.

It is at best unclear whether institutions possess such a capacity and hence whether they can take the fact of someone else's dependency as a direct and compelling reason to act. *Individual members* of an institution possess this capacity, and so can be more or less trustworthy, but, perhaps, institutions cannot.[6] Karen Cook, Russell Hardin, and Margaret Levi in their book *Cooperation Without Trust?* deny that citizens can reasonably trust institutions precisely because institutions do not possess motivations, and citizens lack sufficient familiarity with members of institutions to have access to their motivations. They write,

> [To secure certain goals] we might need only the sporadic services of professionals, business representatives, scientists, and many others, but these are people whom we could not trust [...] because we cannot monitor them and do not have repeated interactions with them...Hence, we want devices that are de facto alternatives to trustworthiness to align their interests with our own. (Cook et al. 2005, p. 104)

This quote suggests that no institutions are worthy of our trust because: (a) trusting an institution is nothing more than trusting its members, and (b) we cannot trust the members of an institution because we are not directly familiar with their behavior and motives. This skeptical account is in tension with our ordinary practices, which suggest that institutions can be more or less trustworthy, apart from the motivations of their members. It is commonplace to speak of "trusting" or "distrusting" institutions, especially medicine, media, academia, and law enforcement in spite of the fact that most people have little access to the motivations of the members of the institutions (Jackson and Bradford 2010; Tyler and Huo 2002).

Ordinary practice might be mistaken, but, ideally, an account of institutional trustworthiness could vindicate these ordinary attributions of trustworthiness to institutions. With some simple modifications, Jones's three-place account can be adapted to institutions, and this can be done without attributing to institutions the capacity to take facts as reasons. We propose the following account:

> **Institutional three-place trustworthiness***:* An institution I is trustworthy with respect to a subject S in domain of interaction D, if and only if I is competent with respect to D, and I is

[6] What we say about institutions having motivations will depend on complex issues of group agency, which we do not have the space or need to discuss. See (Pettit 2009; Pettit and List 2011; Tuomela 2013) for important work in this area.

non-accidentally responsive to the fact that S is counting on I, were S to do so in this domain, such that I functions as counted on.

What does it mean for an institution to be *non-accidentally responsive* to the fact that a person is counting on it in some domain? A view suggested by the Cook, Hardin, and Levi quote is that institutional responsiveness is entirely determined by the responsiveness of its constituent members. For an institution to be responsive to the fact of a subject's dependence is just for its members to be responsive to the fact of the subject's dependence. But we have already resisted that view above. For it seems that an institution can be trustworthy—and trusted—even when the subjects of its authority have little knowledge of the motivations of its members. Instead, we propose that an institution is non-accidentally responsive to the fact that others are counting on it in some domain if the institution's mechanisms, operations, and incentive structure have been successfully designed for the purpose of ensuring that, to some satisfactory degree, representatives of the institution will act as counted on *qua* representatives of the institution.[7] Institutional responsiveness to others' dependence, then, is non-accidental insofar as it is responsive *by design*.

An institution's trustworthiness is not merely a function of the aggregate trustworthiness of its individual members; the structures in place within the institution, which may exist independently of our assessments of any one individual's trustworthiness, are also a significant factor in the trustworthiness of the institution. Our analysis of the Rampart Scandal above illustrates this. An institution that creates an abundance of opportunities for non-compliance among its officials is less trustworthy, even if none of its officials takes advantage of those opportunities. A police department can bolster trustworthiness by supporting a robust internal affairs division, one that reliably identifies and addresses officer misconduct. A city commission might take this a step further by making internal affairs a separate agency, external to the police department, and thereby avoiding the potential conflicts of interest posed by an in-house program. These measures make the law enforcement more responsive by design by ensuring that officers act as counted on by citizens without requiring any fundamental change in officers' motives, and hence without any fundamental change in the trustworthiness of officers.

If one wishes to maintain that trustworthiness is not a feature that institutions can possess, then call the concept we describe here "quasi-trustworthiness." As we attempt to establish below,

[7] This last latin term is included so as to make clear that breaches of trust in the merely personal conduct of institutional representatives cannot undermine the institution's trustworthiness.

whether we call this feature of institutions "trustworthiness" proper, or something else, the feature is relevant to state and institutional legitimacy, and it is threatened by the use of opaque algorithms in criminal justice contexts.

2.4 The Evidence Requirement

But we do not yet have a full picture of what it is for an institution to be fully trustworthy. Being fully trustworthy requires more than competence and responsiveness--to be worthy of others' trust requires that those depending on you have adequate reason to *believe* that you are competent and responsive to their needs. Thus Baier writes, *"*'Trust me!' is for most of us an invitation which we cannot accept at will—either we do already trust the one who says it, in which case it serves at best as reassurance, or it is properly responded with, 'Why should and how can I, until I have cause to?"' (Baier 1986, p. 244).

Jones formalizes what we might call the "evidence requirement" for full trustworthiness as follows:

> B is richly trustworthy with respect to A just in case (i) B is willing and able reliably to signal to A those domains in which B is competent and will take the fact that A is counting on her, were A to do so, to be a compelling reason for acting as counted on and (ii) there are at least some domains in which B will be responsive to the fact of A's dependency in the manner specified in i (Jones 2012, p. 74).

Jones emphasizes the ability and willingness to *signal* responsiveness and competence. But we suspect that signaling is but one way to provide evidence that one is responsive and competent. What is key to achieving full or "rich" trustworthiness is that those who are depending on you have adequate reason to believe you are competent and responsive to the fact of their dependence. Signaling is simply the most obvious way to provide those reasons. To continue with the teaching analogy, one way to give students reason to believe that one is grading their work impartially is by providing adequate feedback on their assignments. In addition to acting as guidance for improvement, providing feedback signals to students that their teacher takes students' dependence on them seriously; that the evaluation of their work is not determined by shifting whims and moods. Failure to signal in this way undermines a person's trustworthiness as a teacher.

If signaling is typically needed to count as fully trustworthy in the interpersonal case, we see no reason that it is any less important for institutional trustworthiness. Like individuals, trustworthy institutions must give citizens reasons to believe that they are competent and responsive to dependence in the domain in which they are being counted on. Indeed, we will see in section 4 that signaling competence and responsiveness by criminal justice institutions is particularly important for the well-being of the citizens these institutions were designed to protect; peace and security break down when citizens lack confidence that the police and courts will respond competently and earnestly to their calls for assistance, follow through with investigations of criminal wrongdoing regardless of a person's class or race, and judge the merits of their case impartially and in light of full evidence. We are now in a position to give our full account of institutional trustworthiness:

> **Rich institutional three-place trustworthiness***:* An institution I is richly trustworthy with respect to a subject S in domain of interaction D, if and only if I is competent with respect to D, I is non-accidentally responsive to the fact that S is counting on I, were S to do so in D, such that I functions as counted on, and I provides adequate reason for S to believe that I is competent with respect to D and non-accidentally responsive to the fact that S is counting on I in D.

3: Forms of Opacity

Section 2 put in place a framework for understanding the nature of trust and trustworthiness as it figures in relationships between citizens and public institutions. In section 3, we describe several sources of opacity and the ways in which the opacity of predictive algorithms can undermine the trustworthiness of state institutions.

Recall our earlier definition of opacity according to which an algorithm counts as opaque when (a) the contribution of any single feature of the world to the algorithm's final determination cannot be easily accessed, either by the human decision-maker or by persons affected by the determination, or (b) the way that the algorithmic determination figures in institutional decision-making cannot be easily understood. What is key for opacity, then, is that something about the algorithm itself, or about the context in which the algorithm is implemented, makes it extremely difficult for an ordinary person to understand how or why it arrives at its determinations, and how those determinations figure in institutional decision-making. Algorithmic opacity can have several sources:

Proprietary Opacity: the algorithm's code may not be made publicly available because of intellectual property protections and concerns about competitive advantage.

Technical Opacity: understanding programming languages is a specialized skill, and few non-programmers are computationally literate in ways that would allow them to understand why an algorithm makes the determinations that it does.

Fundamental Opacity: the decision procedures of machine learning algorithms, which work by a mathematical process of iterative statistical optimization, resist interpretation in terms comprehensible to any human (Burrell 2016).

Implementation Opacity: algorithmic systems are often shrouded in secrecy, either on the grounds that secrecy is important for strategic advantage or because of concerns about public attitudes toward them.

3.1 Proprietary Opacity

Many algorithmic systems in use in the criminal justice system are created and maintained by private technology companies. For example, until very recently, the LAPD used Palantir's data collection and analysis program to generate "Chronic Offenders Bulletins" to alert officers to the most dangerous criminals in a community. The LAPD also made use of software provided by PredPol, another private firm, to identify crime hotspots. Plenty of technology in use in public institutions is used this way. What makes criminal risk assessment and predictive policing technologies distinctive, however, is that the algorithms on which these systems rely to generate their predictions are often inaccessible to those wishing to scrutinize them. As Sarah Brayne puts it:

> Private vendors can hide behind trade secrecy and nondisclosure agreements, ultimately circumventing typical public-sector transparency requirements and lowering police accountability by making it harder for scholars to study, regulators to regulate, and activists to mobilize for or against specific practices (Brayne 2020, p. 140).

Subjects of criminal justice algorithms have sought to scrutinize the algorithms that were used against them as part of their legal defense; yet, the companies that own the technologies are protected by intellectual property laws, and the courts are unable to grant the accused access to the algorithms (Wexler 2018, p. 1346). What's more, even when defendants have reason to believe there was an error in certain relevant and significant inputs to the program, they are not able to access the algorithm to prove that it materially affected the resulting assessment. In the case of one New York inmate, this meant that due to a non-trivial error in the input data, he was denied parole (Wexler 2018, p. 1354). When algorithms used by the police or courts are protected from scrutiny on the grounds that the data constitutes proprietary information, the result is a kind of opacity that not only frustrates the accused's ability to verify that their treatment was fair and appropriate, but also weakens the general public's ability to have faith in the deployment of such methods in society more broadly. This is acute when there are substantial doubts concerning the role of certain variables, as was true in the case just noted of the inmate denied parole despite an error in his input data.

Unlike certain other sources of opacity to be discussed below, proprietary opacity is not inherent in the algorithmic system: regulators could simply require that technology companies waive intellectual property rights when their products are being used by public institutions, at least to allow the accused and oversight groups to access the algorithms.

3.2 Technical Opacity

Even with full access to an algorithmic system, however, few non-specialists have the statistics or computer science background needed to understand why a predictive algorithmic system makes the determinations that it does. Of course, those with education and experience programming algorithms could, to some extent, understand the decision-making process employed by the algorithmic system. But the average person—and, thus, many of those directly impacted by these systems—will be incapable of understanding how the systems work. And this problem would likely remain even if those with the relevant expertise attempted to explain or teach those with less understanding. Thus, technical opacity serves as a barrier to many individuals' ability to understand the treatment to which they are subject.

For example, take predictive policing algorithms. Even the officers who use them do not pretend to understand how or why they make the predictions they do. Sarah Brayne witnessed this ignorance by command officers first-hand while embedded with the LAPD. She writes, "Asked to explain how PredPol works, one captain replied that it 'involves a mathematical equation I know

nothing about"' (Brayne 2020, p. 87). For this reason, many officers are skeptical of the algorithms' utility (Brayne 2020, p. 87). If these systems are technically opaque to the officers using them, they certainly are opaque to the average citizen.[8]

3.3 Fundamental Opacity

While technical opacity concerns the average person's inability to comprehend how the algorithm reaches its assessment, there is a further, deeper kind of opacity, which we call *fundamental opacity*. This kind of opacity results from the fact that the decision procedures of machine learning algorithms, which work by a mathematical process of iterative statistical optimization, are extraordinarily complex (Burrell 2016). The resulting systems are the product of many different individuals working relatively independently, using a variety of technical skills and programming techniques. Thus, it is not only that the average person cannot understand such algorithms; even the best trained experts cannot, on their own, understand them.

What's more, as these systems learn and incorporate new data over time, they tend to become increasingly opaque. This is a function of any machine learning system, which is programmed not only on initial data sets, but also how to learn from mistakes, new data, and various other commands. The promise of such systems is that they will improve over time; however, with this improvement comes an ever-increasing opacity. Beyond a certain point, the 'mind' of such an algorithm can evolve to be so extraordinarily complex that even the best human minds together—even those responsible for its initial programming—could not understand it.

3.4 Implementation Opacity

In addition to the three sources of opacity stemming from features of predictive policing systems themselves, the way that certain criminal justice institutions—most notably, police departments—*implement those systems* can be hidden from the public. *Implementation opacity*, then, concerns the relative inability of individuals subject to algorithmic assessments to understand the role that these assessments play in decisions that affect their lives. Brayne reports encountering several instances where LAPD officers would deny her access to information about how an

[8] To be sure, this sort of technical opacity is pervasive throughout our lives. Most of us do not understand, and are not capable of understanding, how MRI machines, GPS, encryption software, pharmaceuticals, or even computers work. But in many such cases, the technical opacity of such resources does not pose a problem for our treatment by criminal justice institutions, where fair treatment is a constitutive aim and the right to due process a central feature. Thus, while technical opacity is not unique to the use of algorithms in criminal justice, it does pose a special and urgent problem in that domain.

algorithmic system works on this basis (Brayne 2020, p. 93). In the courts system, many defendants are surprised to learn that their sentences will be influenced by an algorithm. In other areas of government, citizens are shocked and dismayed to learn that their well being, access to social services, and livelihoods are determined in part by algorithms (Eubanks 2017). Like proprietary opacity, implementation opacity can be remedied through public disclosure, yet it remains pervasive and persistent.

We have surveyed four sources of opacity arising from the use of computer algorithms in criminal justice. We will now illustrate some of the ways that these sources of opacity can undermine the trustworthiness of criminal justice institutions.

4. Opacity and Trustworthiness

4.1 Signaling Responsiveness and Masking Discrimination:

To understand how opacity undermines trustworthiness, we must first describe a separate concern that some scholars and activists have raised for predictive policing. According to this concern predictive policing systems will be (or are being) used to both *mask* and reinforce discriminatory policing practices. Solon Barocas and Andrew Selbst put the general worry about masking succinctly:

> [D]ecision makers could knowingly and purposefully bias the collection of data to ensure that mining suggests rules that are less favorable to members of protected classes. They could likewise attempt to preserve the known effects of prejudice in prior decision making by insisting that such decisions constitute a reliable and impartial set of examples from which to induce a decision-making rule. (Barocas and Selbst 2016, p. 692).

When data collection and labeling practices are not disclosed by criminal justice agencies, the prospect of biased data collection looms large. Few algorithms used in decision-making by police or the courts use race as a factor in making their determinations. The determinations of those algorithms therefore appear to be race neutral. But appearances can be misleading. An agency that uses that system might knowingly or unknowingly make classifications based on data tainted by prior discriminatory behavior. For example, arrest data appear race-neutral but might be racially skewed by explicit or implicit prejudice in decision-making by police officers. Those who intend to engage in

racially discriminatory practices can thus mask biases in the algorithmic system by appealing to its "race-blind" decision-making procedure.

Now suppose that a watchdog group wanted to determine whether, say, a given predictive policing system was being used in this way. Proprietary, technical, fundamental, and implementation opacity converge to make it next to impossible for the watchdog group to investigate the data collection and labeling methods involved in the training of the system. It is not merely the absence of publicly available training or output data that makes investigation impossible. It is that conducting such an investigation requires knowledge of advanced statistical techniques that few possess. Thus, opacity enables discrimination masking.

In a case like this, one moral concern is discrimination itself. But the mere *prospect* of discrimination masking, made possible by opaque algorithmic systems, also undermines the trustworthiness of law enforcement agencies. Remember that a key feature of institutional trustworthiness is the ability and willingness of an institution to *signal* its responsiveness to the fact that citizens are counting on it. By making discrimination masking a live possibility, the opacity of algorithmic systems calls into question our criminal justice institutions' responsiveness to the needs of some citizens. At the same time, the opacity of algorithmic systems compromises the ability of criminal justice institutions to *signal* responsiveness to need by demonstrating that no discrimination masking is taking place; indeed, it suggests not just an inability but an *unwillingness* of the institution to properly signal its trustworthiness. Therefore, the very prospect of discrimination masking, made possible by opacity, undermines the trustworthiness of criminal justice institutions. Trustworthiness is especially compromised for Black Americans who have antecedent reasons to suspect discriminatory conduct by police agencies. Importantly, this loss of trustworthiness can occur even if masking is not in fact taking place. To signal responsiveness, and to preserve trustworthiness in the face of algorithmic opacity, law enforcement agencies need a mechanism by which they can "show their work" with respect to the operation of predictive algorithmic systems and the role these systems play in strategic operations. More on this in the concluding section.

At this point one might contend that access to the inner workings of the algorithmic system is not needed to determine whether using the system is discriminatory or otherwise unfair. Rather, it might be sufficient simply to inspect the outputs of such a system. For instance, if you want to know whether a predictive policing system is discriminatory you can see what the outcomes are across various relevant racial groups. And none of this requires any particular knowledge of the inner workings of the algorithm itself.

If discrimination or unfairness can be addressed without overcoming opacity, then we have overstated the threat to trustworthiness that opacity poses. But addressing discrimination without overcoming opacity assumes that we can identify sources of unfairness simply by evaluating the outcomes an algorithmic system produces. This works well enough when the fairness at stake is statistical in the sense that it concerns the distribution of beneficial and burdensome outputs across the population. But other fairness concerns cannot be addressed without knowing more about the algorithmic system. Discrimination, at least in part, depends upon *treatment*. That is, in addition to whether outcomes are unequal or suggestive of discrimination, it also matters whether individuals' interests were not equally considered, discounted, or other relevant constraints were violated in the process leading to a determination. And, crucially, this is something that cannot be understood merely by looking at outcomes: we need to understand *how* the outcomes were produced, which requires access to the inner workings of the algorithm. For example, Deborah Hellman has recently argued that the use of certain factors by an algorithmic system in the rendering of an output can constitute "compounded injustice" for those affected by the output. Hellman describes the concern very clearly,

> Accurate data on base rate differences may result from prior injustice. For example, suppose that low educational attainment is predictive of recidivism. And suppose that blacks are more likely to have left school early because the schools they attended were inferior. If an algorithm uses educational attainment to predict recidivism, it may use the fact that blacks were unfairly treated in the past to justify treating them worse today. This is the problem I term "*compounding injustice*"(Hellman 2020, p. 841)

Compounding injustice seems to occur as a result of a certain causal relationship between prior injustice and present mistreatment. It is therefore very difficult to know whether an algorithmic system is compounding injustice unless we know some detail about the factors used in rendering the decision. According to Hellman's proposal , if those factors are both the product of, and aggravating of, past injustice, then their use is unfair on the grounds that it compounds a prior injustice. Opacity poses a serious informational obstacle to those who wish to know whether an algorithmic system compounds injustice.

4.2 Opacity, Competence, and Signaling:

Algorithmic opacity also threatens institutional trustworthiness by eroding the competence of both the law enforcement institution itself and those officials tasked with carrying out its mandate by reducing the expertise of these officials. Increasingly, officials are encouraged or required to rely on and comply with the assessments of algorithmic systems, in lieu of relying on their own professional expertise. Indeed, the algorithmic systems are largely intended to *override* whatever expertise the individual official has. What's more, these systems do not easily allow their methods or findings to be incorporated into a given official's expertise or skill set. Thus, rather than supplementing officials' expertise, these systems are an obstacle to expertise.

One might claim that diminished expertise on the part of officials is counterbalanced by the increase in competence at the institutional level. That is, the institution of law enforcement (or the criminal justice system more broadly) is more competent on balance thanks to the use of algorithmic systems, even if individual officers are less competent. The promise of such systems, after all, is that they decrease bias and error and increase the ability of officials to anticipate future criminal activity. It stands to reason, then, that this could produce a net gain in institutional competence. Relatedly, it might also be claimed that the implementation of algorithmic systems displays a higher-order institutional competence, which involves recognizing the limitations of its officials and adopting alternative methods to satisfy institutional goals. This is ostensibly what the shift to algorithmic systems aims to achieve.

However, this line of reasoning rests on several questionable assumptions. First, it is not obvious that algorithms *have* increased, on balance, institutional competence. In fact, there is hardly any evidence that supports the alleged efficacy of these programs (Boba Santos 2020). Indeed, the reverse may well be true: while algorithmic systems promise to eliminate bias, there is some indication that they further entrench biases, and do so in a way that precludes easy resolution (Angwin et al. 2016; Heaven 2020; McGrory and Bedi 2020; Selbst 2017). Further, there is little evidence that law enforcement agencies, or the private firms from which they purchase their predictive policing systems, have sought such evidence—for example, by funding or otherwise supporting independent research initiatives. This highlights another way institutional trustworthiness is further eroded by opacity—namely, through the institution's failure to robustly signal their competence and responsiveness by embracing only those methods that are empirically sound, and abandoning those methods that fail to live up to their expectations.

Moreover, for institutions to have a plausible claim to higher-order competence, they must have clear procedures in place to address the failures of the algorithmic systems on which they rely.

This is also essential for signaling institutional competence: without an ability to demonstrate that mistakes will be ameliorated, those subject to them will have little reason to trust these institutions. Thus far, however, there has been no clear indication that the institutions deploying these systems have sought to cultivate robust programs for addressing systematic errors. One reason for this is the proprietary opacity mentioned above: there is little that police departments can do to investigate the prospect that an algorithmic system compounds injustice without access to the source code. And again, even if they *did* have access, technical and fundamental opacity present an obstacle: most police departments, for example, do not have the staff or resources, let alone the sophisticated knowledge and expertise, to understand these errors and to identify ways of avoiding them.

The foregoing points generate a practical dilemma for many of the institutions that rely on this technology. On the one hand, to sustain or improve their trustworthiness, these institutions ought to seek out ways to mitigate those areas in which their competence is lacking, and may even opt for as yet unproven but plausible methods when other available alternatives are unlikely to fare better. Thus, predictive policing and risk assessment algorithms appear to be viable approaches for promoting trustworthiness. In practice, however, the circumstances in which these methods are often employed are likely to erode trustworthiness for the many reasons listed above.

5: Trust, Trustworthiness, and Legitimacy

So far, we have argued that algorithmic opacity undermines the trustworthiness of criminal justice institutions. But why exactly is a lack of trustworthiness a *moral* problem for these institutions? In this section, we argue that the answer to this question lies in the relationship between institutional trustworthiness and institutional legitimacy. In making this argument we appeal to two distinct but related conceptions of legitimacy. The first, *descriptive* conception emerges from the social science literature, and holds that a decline in the trustworthiness of an institution can cause citizens to perceive the institution as less legitimate, which reduces compliance and cooperation with the law, which in turn hurts citizens. On a second, *normative* conception of legitimacy, diminished trustworthiness compromises the normative grounds of these institutions' authority to coercively enforce the law.

5.1 Descriptive Legitimacy

According to the conception of legitimacy commonly deployed among social scientists, the legitimacy of criminal justice institutions is nothing over and above the attitudes or judgments of those subject to it (Tyler 2006; Tyler and Huo 2002; Tyler and Jackson 2014). We call accounts of legitimacy that adopt this conception of legitimacy *descriptive*, because they are not concerned with providing a moral justification of the authority of criminal justice institutions. Rather, their aims are to specify conditions that cause individuals to have certain attitudes toward criminal justice institutions and then to investigate how those attitudes influence behavior. One version of such an account holds that legitimacy is just "the *belief* that authorities, institutions, and social arrangements are appropriate, proper, and just" (Tyler 2006, p. 376). A closely related approach takes institutional legitimacy to be a function of individual perceptions of the institution's trustworthiness, or individuals' subjective attitudes of trust toward the institution (Tyler and Huo 2002, p. 104). In other words, an institution is legitimate just in case, and to the extent that, it is trusted, or viewed as trustworthy, by those subject to it.

Tom Tyler has demonstrated that judgments of an institution's legitimacy correlate with both reported compliance and help-seeking behavior. He finds that "if people generally viewed legal authorities as legitimate, they were more likely to indicate that they followed the law in their everyday lives. They were also more likely to indicate that they sought help from legal authorities in a variety of situations" (Tyler and Huo 2002, p. 106). In other words, an institution that is legitimate—i.e., that is trusted, or seen as trustworthy, and thus legitimate—will both secure greater compliance with its commands, and will be sought out by citizens for assistance, allowing it to better fulfill some of its core institutional aims.[9]

The flipside of this, however, is also true: evidence from Tyler's extensive work on descriptive legitimacy supports the conclusion that to the extent that an institution is not trusted—and thus, illegitimate in the descriptive sense—there is a similar decline in individuals' compliance and help-seeking behavior (Tyler and Jackson 2014). Notice that this can (and surely does) generate the following feedback loop: decreasing legitimacy within a given population causes lower levels of compliance and help-seeking behavior by those in that population. The inability to secure voluntary compliance means it is less capable of achieving desired outcomes and delivering on its distinctive mandate. This failure gives rise to a further decrease in legitimacy, as individuals

[9] But see (Hawdon 2008) who questions the direct link between perceptions of trustworthiness by the public and perceptions of legitimacy. Hawdon's skepticism is, in part, due to the assumption that trust is an attitude that takes individuals as its object, and so it is a category mistake to talk of trusting institutions. In section 2 we argued that this assumption is mistaken.

find themselves less trusting of a less capable institution. And the cycle starts anew. Declining descriptive legitimacy therefore negatively affects the institution's ability to achieve the aims that define the institution, which in turn erodes trust and confidence.

Declining descriptive institutional legitimacy can lead to a decline in the trustworthiness of the institution. This result may seem paradoxical, but it follows naturally from the account of institutional trustworthiness that we described in section 2. A failure of institutional legitimacy erodes compliance and help-seeking behavior, and this undermines an institution's trustworthiness by diminishing its *competence* with respect to its mandate. The competence of law enforcement institutions requires compliant, cooperative subjects. Criminal case closure rates, for example, are dramatically affected by witnesses' willingness to come forward with suspect descriptions. The successful investigation of domestic battery offenses requires that victims are willing to report those crimes to police. Thus, when trust is absent, this can erode trust*worthiness* by eroding the competence of law enforcement. In other words, while the attitudes of those subject to an institution's authority are not constitutive of the institution's trustworthiness, they are instrumentally important because they affect behaviors that can, in many cases, hinder the institution's competence, and thus, a decline in its trustworthiness.

As we argued above, opaque criminal justice algorithms undermine the ability of law enforcement and the courts to signal to citizens that they are competent and responsive to their needs. For example, an opaque predictive policing system can make it difficult or impossible for a police department to demonstrate that it is making racially fair choices about where to allocate police on patrol. In this case, signaling competence and responsiveness requires transparency. This prospect is especially troubling in light of the litany of recent high-profile incidents of police brutality and misconduct in the U.S. that have caused a decline in public trust in policing, most significantly within the Black community. A recent poll found that 48% of African-Americans report having very little or no confidence at all in local police to treat Blacks and whites equally; only 12% of whites felt the same way (Santhanam 2020). Furthermore, despite the promise of diminished bias in policing, the use of algorithms by law enforcement has also resulted in many troubling instances, ranging from mistaken facial recognition to harassment and abuse by police officers (Angwin et al. 2016; McGrory and Bedi 2020). The limited empirical research on public attitudes toward criminal risk assessment algorithms finds that 61% of Blacks and 49% of whites believe that these systems are not fair to people up for parole (Smith 2018). Thus, the continued and increased use of opaque algorithmic tools risks fostering a deeper distrust of law enforcement at a time when its

trustworthiness is perilously thin. So long as these algorithms remain opaque to those who are already distrusting of police officers, the institutional legitimacy of law enforcement will continue to erode.

In sum, declining or compromised trustworthiness of law enforcement is morally problematic as both a symptom and further cause of a decline in (descriptive) legitimacy. A decline in descriptive legitimacy produces socially undesirable outcomes, including lower compliance from citizens and decreased help-seeking behavior. This can cause crimes to go unreported, a decline in compliance with the law, and a significant disruption in social order and public safety. Insofar as algorithmic opacity contributes to these outcomes, it is morally problematic.

5.2 Normative Legitimacy

Descriptive legitimacy, however, cannot be the whole story of institutional legitimacy: on its own, it cannot provide the moral ground of law enforcement institutions' authority to enforce the laws. A severely corrupt and unjust state or institution might enjoy descriptive legitimacy if it secures trust from the public through deception concerning its motives and operations. A second conception of legitimacy, popular among philosophers and political theorists, points to a more fundamental problem with declining trustworthiness. Call this *normative legitimacy*. An institution's normative legitimacy concerns the justification of its distinctive authority. The question of normative legitimacy asks what are the grounds of an institution's claim to authority, where authority includes a moral claim to issue demands to those subject to its authority and to coercively enforce those demands. In the case of law enforcement, this authority includes its particular coercive powers, its monopoly on the use of force, its broader role in criminal justice, and so on. In contrast to descriptive legitimacy, an institution's normative legitimacy is not reducible to the beliefs or judgments of those subject to the institution's authority (though, as we will see, these can sometimes play a role in the overall justification). Rather, it is rooted in certain objective conditions that justify that authority. When an institution's normative legitimacy is diminished or compromised, so is its claim to authority.

Rather than defending any particular view of normative legitimacy, we will briefly discuss three general approaches to grounding institutional legitimacy and show how, on each of them, trustworthiness is essential for legitimacy. On perhaps the simplest and most straightforward approach, one might claim that trustworthiness is a *constitutive* feature of legitimate institutions—particularly those, like law enforcement and the courts, that exercise extraordinary power. That is, part of what it is for an institution to enjoy a claim to authority--to be justified in the

exercise of coercive powers in enforcing its demands--is for it to be competent in the relevant domain and non-accidentally responsive to the fact that it is counted on by citizens in that domain. As an institution's competence or responsiveness declines, so too does its claim to authority; beyond a certain point, the institution ceases to be legitimate altogether. Again, the teaching analogy is useful here: if someone (a) fails to possess the skills needed to assess their students' work, (b) lacks the proper motivation to assess their work impartially, or (c) lacks the inclination to signal to students that one is competent and motivated in the right ways, then one's claim weakens to be a legitimate evaluator of students' academic work. Possessing these qualities is part of what it*is* to enjoy legitimate authority to judge the merits of students' work. Similarly, diminishing trustworthiness erodes an institution's claims to (normative) legitimacy. Setting aside whatever harmful consequences might eventuate, the loss of constitutive legitimacy poses a moral problem insofar as criminal justice institutions cannot justifiably issue and coercively enforce demands of citizens without normative legitimacy.

A social contract approach to legitimacy can also support the conclusion that trustworthiness is essential for institutional normative legitimacy. At the risk of oversimplifying, social contract theories generally ground institutional (or state) legitimacy in the consent, agreement, or acceptance of the authority, whether tacitly or explicitly, by those subject to it. The connection between consent and trustworthiness is typically only implicit in these discussions; however, it is made clearest perhaps in the writings of Thomas Hobbes. Owing to the rampant distrust one finds in his vision of the State of Nature, Hobbes defends the rational shift to life under a sovereign authority, who is "trusted to judge between man and man" (Hobbes 1968). Though Hobbes does not use the language of 'trustworthiness', it is natural to read him as holding that one's placing trust in this judge is rational only insofar as the judge is *trustworthy*. Evan Fox-Decent makes this point when he writes that, on Hobbes's view, "the office of the judge is constituted by the judge's trustworthiness in relation to his role" (Fox-Decent 2019, p. 11) We understand these passages to be expressing the claim that the legitimacy of the state is determined by its trustworthiness, because rational parties to the social contract would only cede legitimate authority to govern to a *trustworthy* state. Parties to the social contract would never agree to cede authority to a state that was not competent and responsive to the fact that citizens were depending on it for the provision of certain vital goods like security in one's person. Because a state's competence and responsiveness depend heavily on the competence and responsiveness of its key institutions, when state institutions fail to be trustworthy, so does the state itself.

According to a third approach to legitimacy, which we will call the *comparative feasibility account*,[10] some institutions have "morally mandatory" aims. These are aims there is a moral imperative to promote. Emotional and physical health is a morally mandatory aim of healthcare institutions. N.P. Adams has argued that the legitimacy of an institution with morally mandatory aims turns on its ability to achieve those aims in comparison with feasible alternatives.[11] Whether an institution of this sort is legitimate is determined in part by whether there are alternatives that would better achieve the morally mandatory aims of the current institution, other things being equal. For example, N.P. Adams writes,

> [A] medical institution in the mid-nineteenth century that did not have any effective cancer treatments could very well be legitimate: it could have the right to carry out its tasks, to serve as a cooperative venture with the aim of providing medical care. If we transplant that medical institution, with all its capacities and equipment intact, to our time, it is illegitimate. The only thing that has changed is that there are now better feasible alternatives because we know much more and have much better technology (Adams 2020).

When superior feasible alternatives exist that better promote an institution's morally mandatory aims, an institution that does not employ them suffers a proportionate erosion of its legitimacy; in some cases, failure to employ such alternatives could render the institution illegitimate altogether.

Let us suppose, contrary to the claims of strict abolitionists, that law enforcement institutions do have morally mandatory aims, including the protection of individuals' rights and the promotion of peace, security, and justice. The relevant question for the comparative feasibility account of legitimacy, then, is whether the practices of existing law enforcement institutions are more effective than feasible alternatives at achieving these aims. As we have already seen, the increasing reliance on algorithmic technologies, like predictive policing, can erode public trust. Further, as we noted earlier in our discussion of descriptive legitimacy, this decline in public trust frustrates law enforcement's ability to achieve the aims of the institution, insofar as public distrust causes a decline in compliance or help-seeking behavior. Especially against a backdrop of racial strife concerning American law enforcement's relationship with black Americans, this decline in law

[10] See: (Adams 2020, pp. 300–301) for one articulation of a view like this, which he attributes to (Pogge 2008, p. 25).
[11] This view applies to institutions with morally required aims as opposed to institutions with merely morally permissible aims, such as voluntary social groups, or those with morally impermissible aims, such as hate groups.

enforcement competence, caused by the use of algorithmic systems, makes alternative models of crime prevention seem more promising in comparison.

Finally, the so-called "public reason" approach to institutional legitimacy has recently seen a resurgence in the academic literature concerned with algorithmic *accountability* (Binns 2018). Accountability is the requirement that, in order for a decision to be legitimate, the institutional decision-maker must provide adequate justification to those affected by the decision. Public reason can be useful as a framework for determining what counts as an adequate justification. The concept of public reason has its roots in the political philosophy of Rousseau, Kant, and later John Rawls and Jurgens Habermas.[12] According to the requirement of public reason, legitimate "laws and political institutions must be justifiable to each of us by reference to some common point of view, despite our deep differences and disagreements"(Quong 2013). The exact scope of the requirement of public reason is contested, but it is broadly agreed that it at least acts as a constraint on legitimate coercion by state institutions. When the public reason requirement is applied to institutional decision-makers such as police or court officials, it requires that their decisions be justifiable at least to those affected, but also to the larger citizenry, by appeal to normative and empirical claims that a reasonable person would accept. Ruben Binns has argued that the requirement of public reason can be marshalled to provide a constraint on algorithmic decision-making. He writes, "public reason could act as a constraint on algorithmic decision-making power by ensuring that decision-makers must be able to account for their system's outputs according to epistemic and normative standards which are acceptable to all reasonable people"(Binns 2018, p. 550). On this account, the legitimacy of an institution's decision-making apparatus is threatened when institutional decision-makers cannot account for the outputs of the algorithmic system that informed their decision according to standards that are acceptable to reasonable people. Does the public reason approach leave any role for trustworthiness in grounding institutional legitimacy? We think it can. The public reason approach imposes a constraint on the justifications that the state can avail itself of when imposing coercive measures. One plausible higher order constraint of public reason is that the decision-making apparatus of state institutions must not diminish the trustworthiness of those institutions. Any decision issuing from a decision-making apparatus that threatens the institution's trustworthiness is in one respect unacceptable for a reasonable person.

6: Conclusion

[12](Rawls 1997)

This essay has identified ways in which the use of opaque algorithms by criminal justice institutions undermines the trustworthiness of criminal justice institutions. It has also identified several moral problems with a failure of institutional trustworthiness. To remedy these issues, certain policy proposals suggest themselves. This section briefly notes some of the various approaches criminal justice institutions might take to bolster trustworthiness.

First, those criminal justice institutions deploying algorithmic systems ought to subject these systems to regular external audits, and those in relevant positions of oversight (e.g., legislators) ought to require such audits as a condition of their use. Algorithmic audits allow an unaffiliated third-party expert or team of experts to examine the source code and its recommendations, and identify any problems (Guszcza et al. 2018). Such audits are typically thought to be an important check against the possible biases encoded in such systems (Friedler and Diakopoulos 2016; Kim 2017). Allowing these audits is an important step for both establishing and signaling the competence of the institutions that use algorithmic systems.

One obstacle to such audits stems from the proprietary opacity we discussed above. That is, the companies that produce such algorithmic systems typically retain broad intellectual property rights that protect them against having to make their source code available for audits or other types of inspection. While a concern to foster innovation speaks in favor of granting companies some measure of intellectual property protection, these protections must be curtailed to ensure that systems used in criminal justice, or in the provision of other public goods, are able to be audited. In short, the law and practice here must change. One possibility is to require, by law, that all algorithmic systems be open for audits by third-parties, while also retaining broad intellectual property rights. Another possibility is to require by law that public institutions, like courts and police departments, only purchase and use algorithmic systems that are able to be audited. This would create clear market incentives for companies making these systems to comply with such audits.

Second, and perhaps surprisingly, police departments can use algorithmic systems to promote trust by signaling competence and responsiveness to need. For example, researchers collaborated with the Charlotte-Mecklenburg Police Department in 2015 and 2016 to develop a machine learning model to predict which police officers are at risk for an adverse event involving a member of the public (Carton et al. 2016). The model developed by the research team increased true positives by ~12% and decreased false negatives by ~32% in comparison with the Charlotte-Mecklenburg PD's existing Early Intervention System. Equally importantly, the developers of the system are transparent about the role that various factors play in generating the system's risk

classifications, and its accuracy was robustly tested, not only in terms of a subset of its training data, but in comparison with existing Early Intervention Systems. By helping to develop this system, and then subsequently implementing it, the Charlotte-Mecklenburg Police Department thereby signaled its responsiveness to the safety concerns of citizens about interacting with members of the police force as well as its competence to address those concerns.

Neither of these remedies is a panacea. For one thing, the problems with algorithms in criminal justice extend far beyond the opacity problem we have discussed here. And even with respect to that problem, the recommendations we provide here are, at best, steps in the right direction.

Bibliography

Adams, N. P. (2020). Legitimacy and Institutional Purpose. *Critical Review of International Social andPolitical Philosophy*, *23*(3), 292–310.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed 15 February 2021

Baier, A. (1986). Trust and Antitrust. *Ethics*, *96*(2), 231–260.

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*(3), 671–732.

Binns, R. (2018). Algorithmic Accountability and Public Reason. *Philosophy & Technology*, *31*(4), 543–556. https://doi.org/10.1007/s13347-017-0263-5

Boba Santos, R. (2020). Predictive policing: Where's the evidence? In D. Weisburd & A. A. Braga (Eds.), *Police Innovation: Contrasting Perspectives* (2nd Edition.). Cambridge: Cambridge University Press.

Brayne, S. (2020). *Predict and Surveil: Data, Discretion, and the Future of Policing* New York: Oxford University Press.

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data and Society*, *3*(1).

Carton, S., Helsby, J., Joseph, K., Mahmud, A., Park, Y., Walsh, J., et al. (2016). Identifying Police Officers at Risk of Adverse Events. In *Proceedings of the 22ndACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 67–76). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939698

Cook, K., Hardin, R., & Levi, M. (2005). *Cooperation Without Trust?* Russell Sage Foundation.

de Laat, P. B. (2018). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy & Technology*, *31*(4), 525–541. https://doi.org/10.1007/s13347-017-0293-z

Eubanks, V. (2017). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Fox-Decent, E. (2019). Trust and Authority. In P. B. Miller & M. Hardin (Eds.), *Fiduciaries and Trust: Ethics, Politics, Economics and Law*. Cambridge: Cambridge University Press.

Friedler, S., & Diakopoulos, N. (2016, November 17). How to Hold Algorithms Accountable. *MIT Technology Review*. Accessed 17 February 2021

Guszcza, J., Rahwan, I., Bible, W., Cebrian, M., & Katyal, V. (2018, November 28). Why We Need to Audit Algorithms. *Harvard Business Review*. https://hbr.org/2018/11/why-we-need-to-audit-algorithms. Accessed 17 February 2021

Hawdon, J. (2008). Legitimacy, Trust, Social Capital, and Policing Styles: A Theoretical Statement. *Police Quarterly*, *11*(2), 182–201. https://doi.org/10.1177/1098611107311852

Heaven, W. D. (2020, July 17). Predictive policing algorithms are racist. They need to be dismantled. *MIT Technology Review*. https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/. Accessed 15 February 2021

Hellman, D. (2020). Measuring Algorithmic Fairness. *Virginia Law Review*, *106*(4), 811–866.

Hobbes, T. (1968). *Leviathan*. Baltimore: Penguin Books.

Jackson, J., & Bradford, B. (2010). What is Trust and Confidence in the Police? *Policing: A Journal of Policy and Practice*, *4*(3), 241–248.

Jones, K. (2012). Trustworthiness. *Ethics*, *123*(1), 61–85.

Kim, P. T. (2017). Auditing Algorithms for Discrimination. *University of Pennsylvania Law Review*, *166*, 189–204.

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, *13*, 14–19.

McGrory, K., & Bedi, N. (2020, September 3). Targeted. *Tampa Bay Times*. https://projects.tampabay.com/projects/2020/investigations/police-pasco-sheriff-targeted/intelligence-led-policing/. Accessed 15 February 2021

O'Neil, C. (2016). *Weapons of Math Destruction: HowBig Data Increases Inequality and Threatens Democracy*. Crown.

Parks, B. (2000). *Board of Inquiry into the RampartArea Corruption Incident* (p. 371). Los Angeles Police Department. http://lapd-assets.lapdonline.org/assets/pdf/boi_pub.pdf

Pettit, P. (2009). The Reality of Group Agents. In C. Mantzavinos (Ed.), *Philosophy of the Social Sciences: Philosophical Theory and Scientific Practice* (pp. 67–91). Cambridge: Cambridge University Press.

Pettit, P., & List, C. (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents* Oxford: Oxford University Press.

Pogge, T. (2008). *World Poverty and Human Rights*. Polity.

Quong, J. (2013). Public Reason. https://stanford.library.sydney.edu.au/entries/public-reason/. Accessed 20 July 2021

Rawls, J. (1997). The Idea of Public Reason Revisited. *University of Chicago Law Review, 64*(3). https://chicagounbound.uchicago.edu/uclrev/vol64/iss3/1

Santhanam, L. (2020, June 5). Two-thirds of black Americans don't trust the police to treat them equally. Most white Americans do. *Pbs.org*. https://www.pbs.org/newshour/politics/two-thirds-of-black-americans-dont-trust-the-police-to-treat-them-equally-most-white-americans-do. Accessed 17 February 2021

Selbst, A. (2017). Disparate Impact in Big Data Policing. *Georgia Law Review*, *52*(109), 109–195.

Smith, A. (2018, November 16). Public Attitudes Toward Computer Algorithms. *Pew Research Center*. https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/. Accessed 2 March 2021

State v. Loomis. , No. 881 N.W.2d 749 (2016).

Tuomela, R. (2013). *Social Ontology: Collective Intentionality and Group Agents*. Oxford: Oxford University Press.

Tyler, T. R. (2006). *Why People Obey the Law*. Princeton: Princeton University Press.

Tyler, T. R., & Huo, Y. (2002). *Trust in the Law: Encouraging Public Cooperation with the Police and Courts*. Russell Sage Foundation.

Tyler, T. R., & Jackson, J. (2014). Popular legitimacy and the exercise of legal authority: Motivating compliance, cooperation, and engagement. *Psychology, Public Policy, and Law*, *20*(1), 78–95. https://doi.org/10.1037/a0034514

Vestby, A., & Vestby, J. (2019). Machine Learning and the Police: Asking the Right Questions. *Policing: A*

Journal of Policy and Practice

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology 31*(2), 841–887.

Watson, D. S., & Floridi, L. (2020). The explanation game: a formal framework for interpretable machine learning. *Synthese*. https://doi.org/10.1007/s11229-020-02629-9

Wexler, R. (2018). Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System. *Stanford Law Review*, *70*, 1343–1430.