The Flaws of Policies Requiring Human Oversight of Government Algorithms

Ben Green
bzgreen@umich.edu
Michigan Society of Fellows
Gerald R. Ford School of Public Policy

Abstract

Policymakers around the world are increasingly considering how to prevent government uses of algorithms from producing injustices. One mechanism that has become a centerpiece of global efforts to regulate government algorithms is to require human oversight of algorithmic decisions. However, the functional quality of this regulatory approach has not been thoroughly interrogated. In this article, I survey 40 policies that prescribe human oversight of government algorithms and find that they suffer from two significant flaws. First, evidence suggests that people are unable to perform the desired oversight functions. Second, human oversight policies legitimize government use of flawed and controversial algorithms without addressing the fundamental issues with these tools. Thus, rather than protect against the potential harms of algorithmic decision-making in government, human oversight policies provide a false sense of security in adopting algorithms and enable vendors and agencies to shirk accountability for algorithmic harms. In light of these flaws, I propose a more rigorous approach for determining whether and how to incorporate algorithms into government decision-making. First, policymakers must critically consider whether it is appropriate to use an algorithm at all in a specific context. Second, before deploying an algorithm alongside human oversight, vendors or agencies must conduct preliminary evaluations of whether people can effectively oversee the algorithm.

Table of Contents

1 Introduction

Governments across the world are increasingly turning to automated decision-making systems, often described as algorithms, as tools to make or inform consequential decisions (Calo & Citron, 2021; Eubanks, 2018; Green, 2019; Henley & Booth, 2020). These applications of algorithms have raised significant debate about when and how governments should adopt algorithms. On the one hand, algorithms bring the promise of making decisions more accurately, fairly, and consistently than public servants (Kleinberg et al., 2018; Kleinberg et al., 2015). On the other hand, the use of algorithms by governments has been a source of numerous injustices (Calo & Citron, 2021; Eubanks, 2018; Green, 2019). The algorithms used in practice tend to be rife with errors and biases, leading to decisions that are based on incorrect information and that exacerbate inequities. Furthermore, making decisions via the rigid, rule-based logic of algorithms violates the strongly principle that government decisions should be responsive to the circumstances of individual people. These competing hopes and fears about algorithmic decision-making have led many regulatory bodies to propose mechanisms for how governments can attain the benefits of algorithms while avoiding the risks of algorithms.

One mechanism that has become a centerpiece of global efforts to regulate government algorithms is to require human oversight of the decisions rendered by algorithms. These human oversight policies enable governments to use algorithms so long as a human decision-maker has some form of oversight over the final decision.[1] This is akin to the notion that algorithms may be used to assist human decision-makers but may not be used to render final judgments on their own. Human oversight provisions have proliferated in recent years, joining other mechanisms such as task forces (Richardson, 2019) and algorithmic impact assessments (Government of Canada, 2021) as central to regulating government algorithms. These policies are intended to ensure that a human is in a position to protect against mistaken or biased algorithmic predictions. They are also intended to protect human dignity by keeping a "human in the loop" of automated decision-making (Jones, 2017; Wagner, 2019). In theory, adopting algorithms while ensuring human oversight could enable governments to obtain the best of both worlds: the accuracy, objectivity, and consistency of algorithmic decision-making paired with the individualized and contextual discretion of human decision-making.

Recent legislation positions human oversight as a distinguishing factor that makes government use of algorithms permissible. For instance, the European Union's General Data Protection Regulation (GDPR) strictly limits significant decisions "based solely on automated processing" (European Parliament & Council of the European Union, 2016). The Government of Canada mandates that federal agencies may use high-risk AI systems only with human intervention and a human making

[1] Throughout this paper, "human oversight" refers to human judgment at the moment an algorithm renders a decision about a specific individual. An example of this form of human oversight involves a judge deciding whether to follow an algorithm's recommendation to release a defendant before trial. The discussion of human oversight in this paper does not pertain to more structural forms of human oversight of algorithms, such as people's councils (McQuillan, 2018) and public task forces (Richardson, 2019).

the final decision (Government of Canada, 2021). Washington State allows state and local agencies to use facial recognition in certain instances, but only if high-impact decisions “are subject to meaningful human review” (Washington State Legislature, 2020). By requiring human oversight as a condition for governments to use algorithms when making consequential decisions, these and other policies suggest that human oversight is able to protect against the most severe harms of government algorithms. As an explicit example of such reasoning, European Commission noted in an explanatory memo accompanying its proposed AI Act that human oversight (along with other mechanisms) is “strictly necessary to mitigate the risks to fundamental rights and safety posed by AI” (European Commission, 2021).

Despite the emphasis that legislators have placed on human oversight as a mechanism to mitigate the risks of government algorithms, the functional quality of these policies has not been thoroughly interrogated. Policymakers calling for human oversight occasionally invoke values (such as human rights) as a motivation for these policies, but rarely reference empirical evidence demonstrating that people are able to oversee algorithms in the intended manner. In fact, when policies and policy guidance do reference empirical evidence about human-algorithm interactions, they usually express reservations about the limits of human oversight, particularly related to people over-relying on algorithmic advice (Engstrom et al., 2020; European Commission, 2021; UK Information Commissioner’s Office, 2020). This lack of affirmative evidence calls into question the efficacy of human oversight policies. Given that human oversight is being enacted into policies across the world as a central safeguard against the risks that algorithms pose, it is vital to ensure that these policies actually provide the desired protections. If regulations are not grounded in evidence supporting their efficacy, they may have the perverse effect of alleviating scrutiny of government algorithms without actually addressing the underlying concerns. Ensuring the efficacy of human oversight policies is particularly pressing given the failures of prior efforts to regulate human-automation interactions (Jones, 2015).

This article interrogates the efficacy and impacts of human oversight policies. It proceeds in four parts. The first two parts lay out the context of my analysis. Section 2 provides background on the tensions and challenges raised by the use of algorithms in government decision-making. Section 3 describes the current landscape of human oversight policies. In this section, I survey 40 policy documents from across the world that provide some form of official mandate or guidance regarding human oversight of public sector algorithms, finding three categories of human oversight that are described.

Section 4 provides the primary evaluative analysis of the article. In this section, I evaluate the three forms of human oversight prescribed by the documents surveyed in Section 3. To accomplish this evaluation, I draw on the growing body of evidence regarding how people interact with algorithms in experimental settings and how government decision-makers use algorithms in practice. I find that existing human oversight policies suffer from two significant flaws. First, human oversight policies have a meager basis in empirical evidence: the vast majority of research suggests that people cannot reliably perform any of the desired oversight functions. This first flaw leads to a second flaw: human oversight policies legitimize the use of flawed and unaccountable algorithms in government. Human oversight is being adopted as a remedy for fundamental concerns about the quality and legitimacy of government algorithms yet is unable to actually address the underlying issues that generate these concerns. Thus, rather than protect against the potential harms of algorithmic decision-making in government, human oversight policies provide a false sense of security in adopting algorithms and enable vendors and agencies to foist accountability for algorithmic harms onto lower-level human operators.

Section 5 provides the primary prescriptive analysis of the article. In this section, I consider how to adapt regulation of government algorithms in light of the two flaws to human oversight policies described in Section 4. Given the critical role that human oversight plays in current regulations, updating regulations requires considering more than just the form and role of human oversight. It is clear that policymakers must stop relying uncritically on human oversight as a remedy for the potential harms of algorithms. However, the correct response is not to simply abandon human oversight, leaving governments to depend on autonomous algorithmic judgments. Nor should regulators prohibit governments from ever using algorithms. Instead, legislators must develop more rigorous governance approaches.

Drawing on the lessons from both flaws of human oversight policies, I propose a two-stage approach for determining whether and in what form governments should be permitted to incorporate algorithms into decision-making. First, rather than assume that human oversight can address fundamental concerns about algorithmic decision-making, policymakers should consider whether it is actually appropriate to use an algorithm in a specific context. If policymakers are not comfortable with an algorithm significantly impacting a particular decision, then they should not allow the algorithm to be used for that decision, even with human oversight. The second stage concerns settings where policymakers envision a potential role for algorithms alongside human judgment. Rather than taking for granted that human oversight is effective, policymakers should require that vendors or agencies conduct preliminary evaluations of whether people can collaborate with the algorithm in a desirable manner. If there is not sufficient evidence demonstrating that human oversight is effective and that the algorithm improves human decision-making, then the algorithm should not be adopted.

Compared to the status quo of blanket rules that enable governments to use algorithms as long as a human provides oversight, this proposed regulatory approach will promote more responsible decisions about how governments use algorithms and will help to prevent human oversight from operating as a superficial salve for the injustices associated with algorithmic decision-making.

2 Discretion and Algorithms in Street-Level Bureaucracies

The introduction of algorithms into government decision-making has been a source of significant controversy (Angwin et al., 2016; Calo & Citron, 2021; Eubanks, 2018; Green, 2019; Henley & Booth, 2020). Debates about the proper role for algorithms in government are grounded in the competing demands placed on public policy in the settings where algorithms are used. Many of the most consequential and controversial government uses of algorithms take place in street-level bureaucracies such as courts, police departments, schools, welfare agencies, and social service agencies (Lipsky, 2010). Within street-level bureaucracies, street-level bureaucrats—such as judges, police officers, teachers, and social workers—make consequential decisions about how to instantiate public policy (e.g., allocating benefits or sanctions) with respect to specific individuals (Lipsky, 2010). Examples of algorithms in street-level bureaucracies include judges using risk assessments to inform pretrial and sentencing decisions (Angwin et al., 2016; Wisconsin Supreme Court, 2016), child services workers using predictive models to inform which families to investigate for child neglect and abuse (Eubanks, 2018), and welfare agencies using algorithms to determine eligibility for benefits (Calo & Citron, 2021; Charette, 2018; Henley & Booth, 2020). Policymakers call most strongly for human oversight of algorithms in high-stakes decisions such as these.

Given their role in translating between formal policies and contextual realities, street-level bureaucrats face a dilemma. Street-level bureaucracies represent a setting where the "central contradiction" of public services is particularly stark: the desire for universal and consistent application of rules exists alongside—and opposed to—the desire for discretion and responsiveness to individual contexts (Lipsky, 2010).[2] Street-level bureaucrats must strive to treat everyone equally by following pre-specified rules, yet also to be responsive to individual cases by exercising discretion; adherence to either of these goals necessarily conflicts with the other (Lipsky, 2010). Furthermore, there is "no 'objective' solution" regarding the appropriate balance between rules and discretion (Wilson, 2000).

The extent to which judges and other street-level bureaucrats should be granted discretion over individual cases is thus a source of strenuous and recurring debate (Christie, 1986; Lipsky, 2010). On the one hand, both publics and policymakers often express a strong desire to curb discretion in the interest of objectivity and consistency (Christie, 1986; Lipsky, 2010; Wilson, 2000). Allowing human decision-makers to exercise discretion means relying on the judgment and morality of those individuals, a prospect that causes significant unease (Zacka, 2017). Such discretion raises the specter of biased or arbitrary treatment by unelected agents of the state (Zacka, 2017). The United States is particularly prone to the response of using rules to limit discretion (Wilson, 2000).

[2] Lipsky's discussion of the dilemma facing street-level bureaucrats maps closely onto the common legal debate between rules and standards. While rules involve clear definitions and consequences, facilitating consistency and predictability, standards permit discretion, facilitating flexibility and sensitivity to the circumstances of specific cases (Solum, 2009).

On the other hand, there are many reasons to desire discretion in street-level bureaucracies. Discretion enables street-level bureaucrats to apply public policies laden with ambiguous and conflicting goals based on the particular context at hand (Zacka, 2017). Restricting discretion would entail imposing a rigid, formal logic on the unavoidably ambiguous, uncertain, and unpredictable situations that street-level bureaucrats encounter, preventing them from adapting bureaucratic decisions to complex or novel situations (Lipsky, 2010; Zacka, 2017). Reducing the decisions of judges, police officers, social workers, and other street-level bureaucrats to a set of pre-specified instructions and rules would violate deeply held social commitments to having decisions be responsive to individuals' conditions and needs (Lipsky, 2010). Discretion is therefore desirable—both normatively and practically—as it allows decision-makers to strike an appropriate balance between a variety of important values and objectives in light of each individual's particular circumstances (Lipsky, 2010; Zacka, 2017).

These competing demands on street-level bureaucracies lead to significant debate about the proper role and implementation of algorithms in government. The desire to reduce individual discretion and promote consistency present strong motivations for government agencies to use algorithms. Algorithms bring a promise of accuracy, objectivity, and consistency that is attractive to both policymakers and publics. Evidence suggests that algorithms make policy-relevant predictions more accurately, fairly, and consistently than public servants (Kleinberg et al., 2018; Kleinberg et al., 2015). Thus, in addition to goals such as cutting costs and enhancing efficiency, governments adopt algorithms hoping to attain greater accuracy when making predictive judgments, to replace biased human decisions with "objective" automated ones, and to promote more consistent decision-making (Calo & Citron, 2021; Engstrom et al., 2020; Green, 2019; New Jersey Courts, 2017).[3]

However, despite these desires for algorithmic accuracy, objectivity, and consistency, the specter of governments making high-stakes decisions with algorithms raises two significant concerns. First, evidence suggests that these algorithms are neither as accurate nor fair as hoped. Algorithms used in settings such as the criminal justice system (Angwin et al., 2016), education (Kolkman, 2020), policing (Fussey & Murray, 2020; Green, 2019), and welfare (Calo & Citron, 2021; Charette, 2018) have suffered from notably high error rates. Many of these and other algorithms used by governments have been shown to embed biases against women, minorities, and low-income individuals (Angwin et al., 2016; Buolamwini & Gebru, 2018; Kolkman, 2020; Richardson et al., 2019).

Second, making decisions via the rigid, rule-based logic of algorithms violates the strongly held desire for governments to be responsive to individual circumstances. Unlike street-level

[3] These efforts in government reflect a broader political economic trend toward replacing or undercutting workers with AI systems (Gray & Suri, 2019; Pasquale, 2020). Across domains, however, AI perennially relies on human assistance (Gray & Suri, 2019) and more just outcomes arise when AI is used to complement rather than replace professionals (Pasquale, 2020).

bureaucrats, algorithms adhere to predetermined decision rules and are unable to adapt reflexively to novel or marginal circumstances (Alkhatib & Bernstein, 2019). Scholars have thus raised concerns that automation and algorithms significantly reduce opportunities for discretion in street-level bureaucracies and administrative agencies (Bovens & Zouridis, 2002; Buffat, 2015; Calo & Citron, 2021). Automated decision-making also threatens due process, as such systems often levy judgments without providing notice to individuals or the ability to meaningfully inspect and challenge decisions (Citron, 2008). Although it may be inadvisable to provide individuals with a general right to have decisions about them be made by humans (Huq, 2020), individual justice requires human judgment for decisions that involve ethical and contextual analyses (Binns, 2020). This philosophically grounded notion is matched by human perceptions that decisions made by algorithms are less trustworthy and fair than those made by humans and that being evaluated by algorithms is dehumanizing (Binns et al., 2018; Lee, 2018). Government efforts to adopt algorithms are therefore often met with public skepticism and resistance, as communities reject the prospect of inhuman and inflexible automated systems shaping consequential decisions about their lives. For instance, protests against such tools in recent years have included signs saying "Families over Algorithms" (Scharfenberg, 2018) and chants of "Fuck the algorithm" (Kolkman, 2020).

Thus, although algorithms promise certain benefits, algorithmic decision-making raises significant concerns about the unreliability of algorithms and the lack of contextual human discretion in decision-making. These concerns motivate the turn to human oversight. By requiring that a human is kept in the decision-making loop, human oversight policies attempt to obtain the benefits of algorithms while ensuring that a person is in place to correct for any algorithmic errors and biases and to provide individualized judgment for each decision. The rest of this article describes and evaluates these policies.

3 Survey of Human Oversight Policies

This section summarizes how legislation and policy guidance describe the appropriate role for human oversight of algorithms used by governments. To conduct this survey, I collected policy documents related to government uses of algorithms. I considered documents that fall into one of the following three categories: 1) proposed or passed legislation; 2) policy guidance by government (or government-appointed) bodies; and 3) manuals, policies, and court cases related to two notably controversial and high-stakes risk assessment tools in the U.S. (risk assessments used in criminal justice settings and the Allegheny County Family Screening Tool). I discovered documents through a combination of searching for recently passed or proposed legislation related to AI and privacy, reviewing academic literature and news stories regarding AI regulation, and searching for documentation from the vendors and managers of criminal justice risk assessments and the Allegheny County Family Screening Tool. I reviewed each document discovered to determine whether it discusses human use or oversight of algorithms. This process yielded 40 policy documents that provide some form of official mandate or guidance regarding human oversight of public sector algorithms.[4] These 40 documents are listed in the Appendix (and are all cited at least once in this section).[5]

I analyzed these policy documents to determine how they describe the proper role for humans in government decision-making processes aided by algorithms. I used inductive coding, looking particularly for what each document emphasizes as the central principle guiding human oversight policy. I also looked for any specific mechanisms that policies provided as examples or suggestions for how to facilitate the desired form of human involvement and oversight in decision-making.

This coding process revealed that policies take three overlapping yet distinct approaches to human oversight. Each category rests on a central word or phrase that links the documents within that category to a common framework for human oversight.[6] I describe these categories in order from least to most stringent in terms of the requirements placed on human oversight. The first category involves restricting decisions that are made "solely" by algorithms. The second category provides a corollary to the first, affirmatively stating the importance of human discretion and oversight. Finally, the third category provides an extension of the first and second, emphasizing the need for "meaningful" human oversight. Although there is some overlap between the stated and implied goals of these categories, each reflects a distinct approach to regulating human oversight. Each approach is presented with quotes from the applicable policy documents in order to demonstrate the coherence of these themes.

3.1 Restricting "Solely" Automated Decisions

The first approach to human oversight is to directly prohibit or restrict decisions that are made through "solely" automated means. Nineteen of the 40 reviewed documents fall into this category of oversight. All of them are proposed or passed legislation.

The approach of prohibiting solely automated decisions is taken, most notably, by the European Union's General Data Protection Regulation (GDPR). Article 22 of the GDPR mandates (with limited exceptions) that "data subject[s] shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her" (European Parliament & Council of the

[4] Although most of these documents are specifically focused on government uses of algorithms, others provide broader guidance that encompasses—but does not exclusively consider—government settings. As such, much of the analysis that follows also applies to human oversight in non-government settings.

[5] Out of these 40 documents, 22 are proposed or passed legislation (category 1), eight provide policy guidance (category 2), and ten are related to the two controversial risk assessment tools (category 3).

[6] A significant number of key words and phrases appear across the policy documents in each category. This pattern is not particularly surprising given that recent privacy laws have developed in part by following the model of prior, high-profile privacy laws such as Europe's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) (Chander et al., 2021; Schwartz, 2019). A similar pattern has been observed in AI ethics, as recent statements of principles share many common elements (Fjeld et al., 2020; Jobin et al., 2019). This pattern simplified the coding process, as many of the policy documents contained identical (or very similar) words and phrases.

European Union, 2016). Nine of the EU member states specifically address solely automated decision-making in their national GDPR implementations.[7]

Since the passage of the GDPR, many countries outside the EU have passed laws that include GDPR-style restrictions on solely automated decision-making. In Argentina (Republic of Argentina, 2018), Mauritius (Parliament of Mauritius, 2017), Kenya (Republic of Kenya, 2019), and South Africa[8] (Republic of South Africa, 2013), these policies provide default prohibitions on solely automated decision-making, much like the GDPR. In Bahrain (Kingdom of Bahrain, 2018), Brazil (National Congress of Brazil, 2019), and Uganda (The Republic of Uganda, 2019), these policies do not prohibit any instances of solely automated decision-making, but require protections for the subjects of solely automated decisions. Similarly, Bill 64 in Québec and a proposed federal U.S. policy allow solely automated decisions but require certain rights for the subjects of those decisions (Brown, 2020; National Assembly of Québec, 2020).

All of the policies discussed in this section require that any solely automated decisions are accompanied by protections for the subjects of those decisions. One of the most emphasized and common safeguards is the right for subjects of solely automated decisions to obtain post hoc human intervention.In other words, after someone has been subject to a solely automated decision, they can request that a human inspect and, potentially, alter that decision. All but six of policies discussed here (Argentina, Austria, France, Mauritius, Slovenia, and South Africa) explicitly incorporate a right to human intervention.

3.2 Emphasizing Human Discretion

The second approach to human oversight reflects a corollary to the first approach. In these documents, policymakers, algorithm vendors, and courts emphasize human discretion as providing protection against the potential perils of automated decisions. Fourteen of the 40 policy documents fall into this category of oversight. One is passed legislation, three provide policy guidance, and ten relate to criminal justice and child welfare risk assessments.

Several policy documents point to human oversight and discretion as being essential for protecting values such as human rights. The Canadian Directive on Automated Decision-Making requires

[7] (Malgieri, 2019) provides an overview of how each member state implements Article 22, along with translations of the relevant sections from national laws that are not otherwise available in English. Austria (Austrian Parliament, 2018), Belgium (Belgian Federal Parliament, 2018), France (French Parliament, 2018), and Hungary (Hungarian Parliament, 2018) expand the scope of decisions that are subject to restrictions on solely automated decision-making. France (French Parliament, 2018), Germany (Bundestag, 2019), and the Netherlands (Dutch Parliament, 2018) distinguish the requirements for particular public and private sector settings. Ireland (Houses of the Oireachtas, 2018) and the United Kingdom (U.K. Parliament, 2018) describe detailed procedures that must accompany any permitted solely automated decisions (The UK's implementation of the GDPR went into effect in 2018, before the UK left the EU). Slovenia (Slovenian Parliament, 2020) calls for ex ante impact assessments.

[8] The Protection of Personal Information Act was passed 2013, following the release of the draft GDPR. Most of the provisions (including the section related to automated decision-making) went into effect in 2020.

that decisions likely to have “high” or “very high” social impacts “cannot be made without having specific human intervention points during the decision-making process; and the final decision must be made by a human” (Government of Canada, 2021). A discussion paper by the Australian Human Rights Commission proposes that algorithmic decisions “must be […] subject to appropriate human oversight and intervention” (Australian Human Rights Commission, 2019). Multiple reports by the New Zealand Government similarly emphasize the need to “retain” and “preserve” human oversight as core priorities for how governments should adopt and manage algorithms (Statistics New Zealand, 2018, 2020).

Discretion is emphasized in all eleven of the reviewed documents related specifically to controversial risk assessment algorithms in the United States. In the face of significant public scrutiny, these documents present human discretion as a safety valve that limits the influence of automated tools and mitigates the potential harms of mistaken or biased algorithmic predictions. The key mechanism meant to enable human oversight to perform this function is the discretion to override an algorithm’s judgments.

The Allegheny County Department of Human Services, which oversees the Allegheny Country Family Screening Tool (AFST), presents human oversight as an essential safeguard for decision-making. In response to concerns about the algorithm fully determining outcomes, the Department stresses that “[s]creening decisions are not in any way ‘dictated’ by the AFST,” as “supervisors have full discretion over call screening decisions” (Allegheny County Department of Human Services, 2019a). The Department further justifies the AFST due to the use of human discretion, claiming that the tool produces “few, if any, unintended adverse effects given workers’ willingness to use their own discretion in the screening decision” (Allegheny County Department of Human Services, 2019b).

Documents related to risk assessments used in criminal justice settings place a similar emphasis on discretion. Northpointe, which developed the COMPAS risk assessment, acknowledges that the algorithm can make mistakes and writes that “staff should be encouraged to use their professional judgment and override the computed risk as appropriate” (Northpointe, 2015). Arnold Ventures, which developed the Public Safety Assessment (PSA), strongly emphasizes that “[j]udges are *not required* to follow the PSA” (Arnold Ventures, 2019). The New Jersey Courts (which adopted the PSA in 2017) writes that the PSA “do[es] not replace judicial discretion” (New Jersey Courts, 2017). Other organizations that create and oversee the use of such tools also highlight the importance of professional judgment and the ability of staff to override a risk assessment’s recommendations (Andrews & Bonta, 2001; Steinhart, 2006; Wisconsin Department of Corrections, 2018).

Human discretion also played a central role in two court cases that justified the use of risk assessments in sentencing. In *Malenchik v. State of Indiana* and *State of Wisconsin v. Loomis*, defendants challenged their sentences due to the use of risk assessments to inform these sentences. Both courts hinged their allowance of risk assessments on the presence of human discretion. Despite stating that it would not be acceptable to base sentencing decisions solely or primarily on an algorithm, each court asserted that judicial discretion protects against the potential harms associated with automated sentencing. Noting that risk scores are "neither […] intended nor recommended" to replace individualized sentencing decisions, the Indiana Supreme Court stated that it "defer[s] to the sound discernment and discretion of trial judges to give the tools proper consideration and appropriate weight" (Indiana Supreme Court, 2010). Similarly, acknowledging that COMPAS is imperfect, the Wisconsin Supreme Court stated that "courts [should] exercise discretion when assessing a COMPAS risk score with respect to each individual defendant" (Wisconsin Supreme Court, 2016). These judgments suggest that as long as judges have discretion regarding how to incorporate algorithmic advice into sentences, then it is acceptable to use risk assessments to inform criminal sentencing.

3.3 Requiring "Meaningful" Human Input

The third approach to human oversight reflects an extension of the first and second categories. Rather than simply calling for human involvement in decision-making, documents taking this third approach recognize that some forms of human involvement can be superficial or inadequate. They therefore emphasize the need for human input and oversight to be "meaningful." Seven of the 40 reviewed documents fall into this category of oversight. Two are proposed or passed legislation and five provide policy guidance.

Efforts to promote meaningful human input are addressed, first and foremost, at avoiding the pitfalls of restrictions on solely automated decisions. The narrow scope of these restrictions makes it possible for superficial forms of human involvement to circumvent regulation without actually providing significant human judgment (Veale & Edwards, 2018; Wagner, 2019). Two influential European bodies have emphasized meaningful human input in direct reference to the GDPR. In its guidance related to the GDPR, the Article 29 Data Protection Working Party asserts that "[t]o qualify as human involvement, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture" (Article 29 Data Protection Working Party, 2018). The UK Information Commissioner's Office stresses that "human input needs to be **meaningful**," clarifying that "a decision does not fall outside the scope of [GDPR] Article 22 just because a human has 'rubber-stamped' it" (UK Information Commissioner's Office, 2020).

Other policies and policy guidance also emphasize meaningful human oversight of government algorithms. A Washington State law regulating government use of facial recognition requires that significant decisions concerning individuals "are subject to meaningful human review" (Washington State Legislature, 2020). The European Commission's High-Level Expert Group on AI lists "human agency and oversight" as the first of "seven key requirements for Trustworthy AI," describing the importance of "meaningful opportunity for human choice" (High-Level Expert

Group on AI, 2019). The European Commission's proposal for an Artificial Intelligence Act and a report commissioned by the Administrative Conference of the United States similarly stress the need for substantial forms of human oversight that avoid the pitfalls of simplistic forms of human oversight (Engstrom et al., 2020; European Commission, 2021).

Although none of the policy documents proposing this third approach provide precise or detailed definitions of what "meaningful" human involvement entails, three themes are most salient. First, human decision-makers must be able to disagree with the algorithm's recommendations.[9] Numerous of these documents emphasize that for human oversight to be meaningful, human reviewers must have the competence and authority to override algorithmic decisions (Article 29 Data Protection Working Party, 2018; European Commission, 2021; High-Level Expert Group on AI, 2019; UK Information Commissioner's Office, 2020; Washington State Legislature, 2020).

Second, human overseers must be able to understand how the algorithm operates and makes decisions. Several policy guidance documents suggest explanations of algorithmic decisions as a potential mechanism to facilitate human understanding (Engstrom et al., 2020; High-Level Expert Group on AI, 2019; UK Information Commissioner's Office, 2020). Numerous policies and reports also describe the need for transparency in algorithm design so that human decision-makers can interpret the output of algorithmic systems (Engstrom et al., 2020; European Commission, 2021; High-Level Expert Group on AI, 2019; UK Information Commissioner's Office, 2020; Washington State Legislature, 2020).

Third, human decision-makers must avoid relying on algorithms and instead thoroughly consider all of the information relevant to a given decision. Numerous documents raise concerns about people over-relying on algorithmic recommendations even when granted the ability to make final decisions (Engstrom et al., 2020; European Commission, 2021; High-Level Expert Group on AI, 2020; UK Information Commissioner's Office, 2020). The Article 29 Data Protection Working Party and the UK Information Commissioner's Office further stress that human reviewers must weigh algorithmic assessments alongside additional information and factors (Article 29 Data Protection Working Party, 2018; UK Information Commissioner's Office, 2020).

4 Two Flaws with Human Oversight Policies

In this section, I evaluate the efficacy of human oversight policies and find that they suffer from two flaws. First, drawing on recent empirical evidence about how people interact with algorithms in government and other settings, I consider whether people are capable of providing the types of oversight that human oversight policies call for. Despite the hopes of policymakers, the vast majority of evidence suggests that human oversight fails to provide the envisioned protections against algorithmic errors, biases, and inflexibility. Second, I consider the implications of these

[9] In this respect, documents calling for meaningful human oversight overlap with the documents described in Section 3.2, many of which explicitly call for humans being able to override an algorithm's judgments.

limits to human oversight. Ungrounded assumptions of effective human oversight promote a false sense of security in adopting algorithms and shift accountability for algorithmic harms from vendors and policymakers to frontline operators who typically have little agency or power.

4.1 Flaw 1: Human Oversight Policies Have a Meager Basis in Empirical Evidence

The first flaw of human oversight policies is that they have little basis in empirical evidence. The vast majority of research suggests that people are unable to provide reliable oversight of algorithms. The underlying problem here is a mismatch of skills and responsibilities: algorithms have been adopted for their superior prediction abilities relative to humans, yet then those same humans have been tasked with judging the quality of algorithmic predictions. Asking people to monitor automated systems that were adopted to improve upon human performance creates "an impossible task" for the human monitor (Bainbridge, 1983).

To show the empirical limits of human oversight policies, I will return to the three forms of human oversight introduced in Section 3 and describe why each is highly unlikely to provide the desired protections against algorithmic errors, biases, and inflexibility.

4.1.1 Restrictions on "Solely" Automated Decisions Provide Superficial Protection

Policies that restrict "solely" automated decisions have the clearest flaws: they provide protection for a limited number of cases and are susceptible to avoidance through superficial human involvement. It is uncommon for public sector algorithms—particularly in high-stakes settings such as criminal justice and child welfare—to operate without human involvement and without a human making the final decision. Policies prohibiting "solely" automated decisions are therefore unlikely to apply in the cases that have generated the most public scrutiny and outcry. Furthermore, the narrow scope of "solely" automated decisions creates flimsy and easily avoidable protections. At least by the letter of these laws, any nominal form of human involvement is sufficient to avoid the protections placed on solely automated decisions. Provisions like the GDPR's Article 22 thus may create an incentive to introduce superficial human oversight of automated decisions (i.e., "rubber stamping" automated decisions) as a way to bypass restrictions (Veale & Edwards, 2018; Wagner, 2019).[10]

In addition, the right to post hoc human intervention fails to provide robust protections against the harms of solely automated decisions. Procedurally, the right to human intervention puts the onus on individuals to request human review after they have already been harmed by a decision. Many people will have neither the means nor knowledge to take advantage of this right. Even when people do request human intervention, it can be slow and onerous to obtain remedies, allowing the harm of a flawed automated decision to manifest (Calo & Citron, 2021). Substantively, human intervention is unlikely to produce better decisions in most settings. This form of human

[10] This type of superficial human oversight would represent an example of "skeuomorphic humanity," providing the impression that a human is making decisions even when that is not the case (Brennan-Marquez et al., 2019).

intervention amounts to the ability of a human reviewer to override an automated decision. As the next section will describe, people are bad at evaluating the quality of algorithmic judgments, leading them to typically override algorithms in detrimental ways.

4.1.2 Human Discretion Does Not Improve Outcomes

Even when human oversight moves beyond "rubber stamp" approaches, such that people are granted agency to use discretion and make final decisions, human oversight is unlikely to provide protections against the harms of algorithmic decision-making.

Across a wide range of domains, automated decision-support systems tend to alter human decision-making in unexpected and detrimental ways. Numerous studies have demonstrated that humans (including experts) are susceptible to "automation bias"—i.e., deferring to automated systems without exercising the level of independent judgment that they would without an automated aid (Parasuraman & Manzey, 2010; Skitka et al., 1999). Automation bias can involve omission errors—failing to take action because the automated system did not provide an alert—and commission errors—following the advice an automated system even though it is incorrect and there is contradicting evidence (Parasuraman & Manzey, 2010; Skitka et al., 1999). Furthermore, automating certain parts of human tasks can make the remaining tasks more difficult and cause human skills to deteriorate (Bainbridge, 1983). As a result, automated systems may simply lead to different types of errors rather than reducing overall errors as intended (Skitka et al., 1999). Automation can also create a diminished sense of control, responsibility, and moral agency among human operators (Berberian et al., 2012; Cummings, 2006).

Similar issues arise when people collaborate with algorithms to make predictions. A significant body of research demonstrates that people are bad at judging the quality of algorithmic outputs and determining whether and how they should override those outputs. People struggle to judge the quality of algorithmic advice (Goodwin & Fildes, 1999; Green & Chen, 2019a, 2019b; Lai & Tan, 2019; Springer et al., 2017), leading them to discount accurate algorithmic recommendations and to rely on bad algorithmic recommendations (Dietvorst et al., 2015; Goodwin & Fildes, 1999; Lim & O'Connor, 1995; Springer et al., 2017; Yeomans et al., 2017). This means that even though algorithmic advice can improve the accuracy of human predictions, people's judgments about when and how to diverge from algorithmic recommendations are typically incorrect (Green & Chen, 2019a, 2019b; Grgić-Hlača et al., 2019; Lai & Tan, 2019). People have also been shown to exhibit racial biases when incorporating algorithmic advice into their predictions (Green & Chen, 2019a, 2019b). Furthermore, although an evaluation of the Allegheny Family Screening Tool found that staff were able to override many algorithmic errors (De-Arteaga et al., 2020), other evidence shows that algorithmic errors reduce the quality of expert judgments (Kiani et al., 2020).

The use of algorithms in policing and the criminal justice system highlights the limits of human oversight and discretion in practice. Police have been shown to follow incorrect advice from algorithms, even when tasked with overseeing an algorithm and under no mandate to follow its advice. For instance, police in London “overwhelmingly overestimated the credibility” of a live facial recognition system, judging computer-generated matches to be correct at three times the actual rate of accuracy (Fussey & Murray, 2020). Such behavior led to the first known case of arrest due to faulty facial recognition in the United States, when the Detroit Police Department arrested a man due solely to a facial recognition match that was clearly incorrect (Hill, 2020).

In contrast, judges across the United States regularly deviate from algorithmic advice, but typically in detrimental ways. Reports from several U.S. jurisdictions found that judges frequently override release recommendations in order to detain defendants, leading to inflated detention rates (Human Rights Watch, 2017; Sheriff’s Justice Institute, 2016; Steinhart, 2006; Stevenson, 2018; Stevenson & Doleac, 2021). Furthermore, risk assessments have exacerbated racial disparities in pretrial detention, in part because judges often make more punitive decisions regarding Black defendants than white defendants with the same risk score (Albright, 2019; Cowgill, 2018; Stevenson & Doleac, 2021). Thus, rather than enable decision-makers to identify and correct algorithmic biases, human discretion can enable decision-makers to inject new forms of inconsistency and bias into decisions.

4.1.3 Even “Meaningful” Human Oversight Does Not Improve Outcomes

Finally, calls for “meaningful” human oversight are also unlikely to deliver on their promises. Such calls face two major issues. First, although the policy documents reviewed suggest three core components of meaningful human oversight, none propose a straightforward definition of meaningful oversight. They agree that a human operator rubber stamping algorithmic decisions clearly does not constitute meaningful oversight but provide no standard for evaluating how meaningful any form of human oversight is.[11] It is not necessary for meaningful human oversight to have a singular precise definition, as many policies appropriately rely on standards whose meaning depends on contextual judgments (Lipsky, 2010; Solum, 2009; Zacka, 2017). However, for any definition of meaningful oversight to be desirable, there must be at least some components of meaningful oversight that are proven to improve outcomes. In the absence of such mechanisms, proposals for meaningful human oversight of public sector algorithms will sound reassuring without actually providing any benefits.

This definitional issue is compounded by a functional issue: the three components described as central to meaningful human oversight are either unlikely to improve decision-making or are very difficult to achieve. This means that none of the proposals for meaningful human oversight involve mechanisms that improve human oversight. The first requirement is that humans must be allowed to override algorithmic recommendations. This is already the case in most of the high-stakes

[11] In the related context of autonomous weapon systems, “meaningful human control” has gained widespread support as a governance principle, yet the “inherent imprecision” of this principle means that there is no consensus regarding what it actually entails (Crootof, 2016).

settings in which algorithms are used, such as the criminal justice system. Yet as described in Section 4.1.2, both laypeople and public servants cannot reliably evaluate the quality of algorithmic advice and tend to override algorithms in detrimental ways. Thus, although this tenet of meaningful human oversight is met in many instances, there is little evidence suggesting that it improves outcomes.

The second requirement is that human overseers must understand the algorithm's operations and outputs. This goal is meant to be facilitated by algorithmic explanations and transparency. As with the first requirement, despite the broad support for this idea, evidence suggests that algorithmic explanations and transparency do not actually improve human oversight. Studies have found that explanations do not improve people's ability to make use of algorithmic predictions (Bansal, Wu, et al., 2021; Green & Chen, 2019b). In fact, explanations can have the detrimental effect of prompting people to place greater trust in algorithmic recommendations even when the recommendations are incorrect (Bansal, Wu, et al., 2021; Jacobs et al., 2021) or when the explanations have no basis in the algorithm's actual functioning (Lai & Tan, 2019). Algorithmic transparency similarly reduces people's ability to detect and correct model errors (Poursabzi-Sangdeh et al., 2021). Based on this evidence, the proposed remedies of explanations and transparency appear to hinder—rather than improve—people's ability to identify algorithmic mistakes and make effective use of algorithmic recommendations.

The third requirement of meaningful human oversight is that decision-makers must avoid relying on algorithms and instead consider all of the relevant information. Although this goal is appealing, evidence suggests that it is difficult (if not impossible) to achieve in practice. Studies have found that automation bias persists even after training and explicit instructions to verify an automated system (Parasuraman & Manzey, 2010). Furthermore, even when people do not rely entirely on automation, an algorithm can still significantly alter how people make decisions. Experimental studies suggest that risk assessments increase the weight that judges, law students, and laypeople place on risk relative to other considerations when making simulated pretrial and sentencing decisions (Green & Chen, 2021; Skeem et al., 2019; Starr, 2014). These changes in decision-making processes represent an effective shift in policy and jurisprudence that can increase racial disparities (Green & Chen, 2021). In other words, it is not sufficient for an algorithm merely to improve the accuracy of human predictions, as an algorithm's effects on decision-making processes can counteract the potential benefits of enhanced predictions (Green & Chen, 2021). Thus, while it is desirable that decision-makers presented with algorithms balance the algorithm's advice with other information and factors, evidence suggests that people typically defer to automated tools and increase their attention to the factors emphasized by algorithms.

4.2 Flaw 2: Human Oversight Policies Legitimize Flawed and Unaccountable Algorithms in Government

The second flaw of human oversight policies follows from the first: because human oversight mechanisms are ineffective, human oversight policies legitimize flawed and unaccountable algorithms in government without mitigating the issues with these tools. This flaw has two dimensions. First, human oversight provisions paper over foundational concerns about algorithmic decision-making, providing policymakers and publics with a false sense of security that even flawed algorithms are safe to use in high-stakes arenas.[12] Second, by emphasizing the role of human operators, human oversight provisions obscure the responsibility of other actors in shaping government algorithms. All told, human oversight policies effectively create a loophole that justifies the adoption of flawed algorithms and enables vendors and agencies to shirk accountability for algorithmic harms.

4.2.1 The Assumption of Effective Human Oversight Provides a False Sense of Security in Adopting Algorithms

Human oversight policies cover up fundamental concerns about the use of algorithms in government decision-making. In response to community activism and media exposés, vendors and policymakers increasingly acknowledge that government algorithms can be error prone and biased and that relying on algorithms for high-stakes decisions can be unjust. In many cases, these concerns undercut the reasoning for adopting algorithms in government. However, rather than prohibit certain applications of algorithms, policymakers present human oversight as the salve that enables governments to obtain the benefits of algorithms without incurring the associated harms. Under this approach, human oversight plays a central role in enabling governments to employ high-stakes algorithms. Were any of the proposed forms of human oversight effective, then perhaps this remedy would be appropriate to mitigate the risks of government algorithms. But given the failures of existing human oversight mechanisms, such regulations serve only to divert attention from fundamental concerns about algorithms and justify the inappropriate integration of algorithms into government decision-making.

Two examples demonstrate how human oversight provides false comfort in the face of concerns that undermine the logic for algorithmic decision-making in government. The first example involves the assertion that humans should regularly override algorithmic decisions. Acknowledging that COMPAS is imperfect and that decisions must involve human judgment, Northpointe and the Wisconsin Department of Corrections describe an expected override rate of approximately 10% due to staff disagreeing with the assigned risk score (Northpointe, 2015; Wisconsin Department of Corrections, 2018). The Allegheny County Department of Human

[12] The “rubber stamp” loophole set up by prohibitions on “solely” automated decisions has proven relatively easy to identify (Veale & Edwards, 2018; Wagner, 2019), prompting some regulators developing legislation to drop calls for GDPR-style restrictions on solely automated decisions (Office of the Privacy Commissioner of Canada, 2020). However, the limits to more substantial forms of human oversight are more subtle and consequential, as these forms of oversight are increasingly common and have intuitive appeal.

Services takes this further, pointing directly to the observed rate of overrides (e.g., 37% for children assigned the highest risk category) as evidence that the algorithm is not replacing human judgment (Allegheny County Department of Human Services, 2019a). Similarly, multiple policy guidance documents calling for meaningful human oversight warn that, if humans agree with the algorithm too often, then decisions should be considered solely automated (Article 29 Data Protection Working Party, 2018; UK Information Commissioner's Office, 2020). In other words, for human oversight to be meaningful, decision-makers must routinely disagree with the automated system (Veale & Edwards, 2018; Wagner, 2019).

At first glance, these calls for overrides appear prudent. Policymakers are right to be concerned about the perils of solely automated decision-making. If a human decision-maker rarely overrides an algorithmic decision, then the decision-making process would indeed be nearly solely automated. Allowing humans to disagree with algorithms seems to provide an avenue for injecting discretion and error-correction into decisions. The problem, however, is that human overrides cannot actually remedy the concerns that motivate overrides. Policies calling for overrides therefore provide the appearance of quality control—legitimizing the use of controversial algorithms—but do not actually address the underlying issues.

Consider the two scenarios in which overrides of automated decisions appear particularly desirable. One reason to call for human overrides is because of a lack of trust in an algorithm to make accurate and fair decisions. Although human overrides seem reassuring in the face of doubts about the quality of algorithmic judgments, this remedy is unlikely to be effective: substantial evidence demonstrates that humans tend to override algorithms in detrimental rather than beneficial ways (Green & Chen, 2019a, 2019b; Grgić-Hlača et al., 2019; Lai & Tan, 2019). A second reason to call for human overrides is because an algorithm fails to account for considerations that are essential to a given decision. For instance, pretrial risk assessments do not consider the full range of factors that judges must balance (Green & Chen, 2021). In cases such as this, human overrides seem to improve decisions by incorporating considerations that the algorithm omits. However, although human overrides provide reassurances in the face of concerns about myopic algorithms, they are unable to address the underlying concerns. People cannot reliably balance an algorithm's advice with other important factors, as they often over-rely on automated advice (Parasuraman & Manzey, 2010; Skitka et al., 1999) and place greater weight on the factors that algorithms emphasize (Green & Chen, 2021; Skeem et al., 2019; Starr, 2014).

Thus, when proponents of criminal justice and child welfare risk assessments call for human overrides, they deflect criticism of these controversial tools but fail to mitigate the underlying concerns. Human oversight cannot address concerns about inaccurate, unfair, and myopic algorithms. More structural reforms are necessary. If an algorithm is so flawed that policymakers do not trust it to make decisions without a significant number of human overrides, then the appropriate remedy is to improve the algorithm. If the algorithm cannot be sufficiently improved, then the appropriate remedy is to stop using the algorithm. Similarly, if an algorithm ignores criteria that must be considered when making a given decision, then the appropriate remedy is to alter the algorithm so that it accounts for all relevant criteria. If this is not feasible, then the appropriate remedy is to stop using the algorithm.

The false sense of security provided by human oversight can also be seen in *State v. Loomis,* in which the Wisconsin Supreme Court relied on judicial discretion to justify allowing the COMPAS risk assessment in sentencing.[13] Due to concerns about biases, errors, and due process violations, the Court stated that risk assessments could not solely determine sentences. However, the Court also stated that risk assessments could be used as long there was human oversight. A concurrence from the Court's Chief Justice makes this particularly explicit, clarifying that "consideration of COMPAS is permissible; reliance on COMPAS for the sentence imposed is not permissible" (Wisconsin Supreme Court, 2016). In other words, human oversight makes it legitimate for risk assessments to inform sentences.

Although the Wisconsin Supreme Court was right to recognize the perils of relying on COMPAS to determine sentences, its reliance on human oversight to alleviate these concerns was misplaced. Consider the two central concerns that prompted the Wisconsin Supreme Court to call for human oversight. First, the Court responded to concerns about errors and biases in risk assessments by pointing to judges' discretion to ignore the tools. Most notably, the court mandated that COMPAS be accompanied by a list of concerns that have been raised about the tool (Wisconsin Supreme Court, 2016). Although it seems reassuring to prompt judges to use discretion when considering risk assessments, doing so leaves judges with conflicting guidance: on the one hand the court hailed risk assessments for their ability to provide reliable and accurate predictions, while on the other hand the court warned that risk assessments can be laden with errors and biases. Furthermore, discretion is unlikely to be an effective remedy, as evidence demonstrates that judges often use their discretion to ignore risk assessments in punitive and racially biased ways, (Albright, 2019; Cowgill, 2018; Human Rights Watch, 2017; Sheriff's Justice Institute, 2016; Steinhart, 2006; Stevenson, 2018; Stevenson & Doleac, 2021).

Second, the Wisconsin Supreme Court attempted to address concerns about risk assessments violating due process by limiting the extent to which judges could rely on them. The Court asserted that judges may use risk assessments only to inform—but not determine—particular aspects of sentences, defending the use of COMPAS because it had no impact on the final outcome. Even as it praised the ability of risk assessments to promote better outcomes, the court affirmed Loomis' sentence on the grounds that the circuit court "would have imposed the same sentence regardless of whether it considered the COMPAS risk scores" (Wisconsin Supreme Court, 2016). Although it seems reassuring to assert that judges cannot rely too heavily on risk assessments, this reasoning

[13] The Indiana Supreme Court followed similar reasoning when justifying risk assessment tools in *Malenchik v. State* (Indiana Supreme Court, 2010).

produces incompatible guidance: it is impossible for risk assessments to inform sentences without ever altering those sentences.[14] Furthermore, it is unlikely that judges will limit the role of risk assessments as the Court envisions, as evidence suggests that judges and other people often defer to automated advice and change their decision-making processes due to algorithms (Green & Chen, 2021; Parasuraman & Manzey, 2010; Skeem et al., 2019; Starr, 2014).

By calling for judicial discretion, the Wisconsin Supreme Court alleviated its own concerns about sentencing risk assessments without actually mitigating the harms associated with these tools. The foundational issue concerning the Court was the low quality of risk assessments and the conflict between risk assessments and due process. Judicial discretion cannot address these concerns. Rather than pointing to human oversight, the Court should have placed greater scrutiny on the algorithm's quality and on whether it is appropriate for an algorithm to alter a defendant's sentence. If the Court would not allow the use of COMPAS without human oversight, then there is scant evidence supporting its decision to allow the use of COMPAS with human oversight. In other words, if the Court was not comfortable with the accuracy and fairness of risk assessments, then it should not have allowed these tools to be presented to judges until their quality improves. Similarly, if the Court was not comfortable with algorithms altering sentences, then it should not have allowed algorithms to be incorporated into sentencing adjudications at all.

4.2.2 Relying on Human Oversight Diminishes Responsibility and Accountability for Institutional Decision-Makers

By appearing able to address foundational concerns about government algorithms, human oversight policies shift responsibility for algorithmic systems from agency leaders and technology vendors to human operators. Human oversight policies position frontline human operators as the scapegoats for government algorithms, even though the structure of these systems has been shaped by more powerful institutional actors. This approach allows vendors and governments to have it both ways: they can promote the use of an algorithm by proclaiming how its capabilities exceed those of humans, while simultaneously defending the algorithm and its creators from scrutiny by pointing to the protection (supposedly) provided by human oversight. When something goes well, governments and vendors can hail the benefits provided by an algorithm. When something goes wrong, governments and vendors can blame and punish the individuals operating these systems—even though algorithmic errors and injustices are typically due to factors over which frontline human overseers have minimal agency, such as the system design or the political goals motivating implementation (Elish, 2019; Perrow, 1999).

A notable instance of this convenient finger-pointing occurred in the aftermath of a Black man in Detroit being wrongfully arrested following an incorrect match by a Detroit Police Department

[14] If risk assessments can merely inform outcomes that would have been reached independently, then there is no functional reason to use such tools at all. However, if risk assessments are to inform outcomes that would not have been reached independently, then such tools must influence those sentences.

(DPD) facial recognition system (Hill, 2020). Representatives from each of the three technology companies that produced the system immediately blamed the mistaken arrest on human operators following an inappropriate investigation process (Hill, 2020). Appearing the following year on the national news program *60 Minutes*, the Detroit Police Chief similarly blamed "[s]loppy, sloppy investigative work" for the incident (CBS News, 2021). Noting that the detective and commanding officer have since been disciplined, the Detroit Police Chief added, "it wasn't facial recognition that failed. What failed was a horrible investigation" (CBS News, 2021).

Although the operators surely could have followed a more thorough investigative process in this particular case, placing blame on the operators obscures the role of other actors—particularly the technology vendors and police chief—responsible for the system-level decisions that led to this arrest. It is DPD leadership and the technology vendors who chose to implement an investigative technology known to have low accuracy when evaluating Black faces (Buolamwini & Gebru, 2018) in the U.S. city with the largest share of Black residents, against the opposition of many Black residents (Campbell, 2019). In fact, the Detroit Police Chief himself admitted that the DPD's facial recognition system is incorrect 96% percent of the time, and DPD data demonstrates that the system is used almost exclusively to investigate Black suspects (Koebler, 2020). No form of human oversight could make it appropriate for police to use a facial recognition system that violates civil liberties, is incorrect in the vast majority of cases, and is used to surveil Detroit's Black population. Instead, these harms can be remedied only by banning police facial recognition altogether (Hartzog & Selinger, 2018; Stark, 2019), as several jurisdictions across the United States have recently done (Hill, 2021) and many civil society organizations have called for (Amnesty International, 2021; European Digital Rights, 2021a; Fight For The Future, 2021).

5 An Alternative to Human Oversight Policies

This study has shown that policies mandating human oversight for government algorithms are flawed in two ways. First, human oversight is unable to provide the desired protections. In turn, human oversight policies legitimize flawed and unaccountable algorithms in government without remedying the issues with these tools. These findings demonstrate that policymakers must stop relying on human oversight as a central mechanism for protecting against the harms of government algorithms.

If legislators cannot depend on human oversight, then how should they regulate government algorithms? The answer is certainly not to allow algorithms to be used autonomously, fully replacing human judgment. Nor is it to entirely abandon any uses of algorithms, relying solely on human judgment for all decisions. Instead, drawing on the lessons from the two flaws of human oversight policies, it is necessary to develop an alternative strategy for determining whether (and in what form) to incorporate algorithms into government decision-making. In this section, I propose a two-stage approach to this task, centered on addressing both flaws of human oversight policies. Rather than crafting blanket rules that enable governments to use algorithms as long as a human provides oversight, policymakers must place far greater scrutiny on whether an algorithm is even appropriate to use in a given context and whether human oversight can actually function as desired.

5.1 Stage 1: Determining Whether an Algorithm is Suitable for a Decision

The first stage in determining the appropriate role for algorithms in governments is to evaluate whether a given algorithm is suitable for the proposed decision. This stage is intended to address the second flaw of human oversight policies, which is that they legitimize the use of algorithms despite flaws suggesting that these tools should not be used at all. If algorithms are going to be incorporated into government decision-making procedures, then we must be comfortable with those algorithms making (or at least altering) decisions there must be mechanisms for holding institutional decision-makers accountable.

Determining whether an algorithm is appropriate for a decision requires considering several components. Policymakers must first consider "red lines" that mark unacceptable uses of algorithms. Applications such as facial recognition and predictive policing violate fundamental notions of justice and human rights (Hartzog & Selinger, 2018; Richardson et al., 2019; Stark, 2019; Stop LAPD Spying Coalition, 2018). The issues with algorithms such as these cannot be remedied by more accurate algorithms or more reliable human oversight. Instead, it necessary to ban these applications of algorithms outright. In recognition of these dangers, numerous jurisdictions across the U.S. have recently placed bans or moratoria on police use of facial recognition and predictive policing algorithms (Hill, 2021; Ibarra, 2020; Stein, 2020).[15]

If an algorithm is not prohibited by a red line, policymakers should then consider whether it can be appropriately integrated into a given decision. Making this determination involves two dimensions of analysis: one focused on the decision and one focused on the algorithm. The first dimension of analysis is the extent to which the decision in question is amenable to algorithmic decision-making. As with decision-support systems in other domains (Cummings, 2006), the appropriate role for algorithms depends on the extent to which human discretion is essential to making the decision. Because algorithms make decisions according to predetermined rules, the more that a decision requires individualized human discretion, the less appropriate it is for algorithms to play a role in decision-making. While it may be appropriate to automate decisions guided by predetermined rules, decisions guided by standards require human discretion and cannot be adequately made by algorithms (Citron, 2008). Discretion is particularly desirable for decisions that require determining the appropriate application of ambiguous and conflicting goals in individual cases that are difficult to classify in advance (Binns, 2020; Zacka, 2017). This analysis suggests that machine learning algorithms may be appropriate for pure prediction problems, but not for decisions that involve balancing predictions with other factors.

[15] Policymakers should also consider additional factors at this stage, such as whether the algorithm in question actually represents a desirable approach to reform (Green, 2020).

The second dimension of analysis is the extent to which the algorithm in question is trustworthy, relative to the stakes of the decision at hand. The more that the algorithm is trustworthy, the more appropriate it is to have the tool influence decisions. Trust in an algorithm depends on several factors. An essential prerequisite is that the outcome of interest can be measured with reasonable accuracy and validity, as low-quality data and bad proxy variables present a significant limit on an algorithm's reliability. It is also essential that the algorithm has been rigorously evaluated for the task at hand and that those tests demonstrate that the algorithm makes predictions accurately and fairly. Transparency into the algorithm's source code, training data, and development process further enhances trust. Finally, trust is relative to the stakes of the decision: decisions that involve higher stakes associated with erroneous predictions require a higher standard for trusting algorithms. Many of these principles for evaluating the trustworthiness of an algorithm are already incorporated into the requirements for algorithmic impact assessments (Brown, 2020; California Legislature, 2021; Government of Canada, 2021).

Considering both discretion and trust in tandem can inform the appropriate role for an algorithm within a particular decision. Each dimension yields a general principle regarding algorithmic decision-making. The more that the decision requires discretion, the less appropriate it is to incorporate an algorithm into that decision. Similarly, the less that the algorithm is trustworthy, the less appropriate it is to incorporate that algorithm into a decision. These two principles combine, as summarized in Table 1. In quadrants 2 and 3, the appropriate mode of decision-making is relatively clear. If there is a high need for human discretion and low trust in the algorithm, then solely human decision-making is most appropriate.[16] Conversely, if there is a low need for human discretion and high trust in the algorithm, then we should rely primarily on algorithmic decision-making (with the precise form of human involvement subject to the results of empirical evaluation). In quadrants 1 and 4, the appropriate role for an algorithm is less clear. In these scenarios, we must turn to evidence regarding how people interact with the algorithm in the given setting. Although the default in both of these scenarios is to retain human decision-making, judgments should depend on case-by-case evaluations of whether (and in what forms) incorporating the algorithm into human decision-making can improve outcomes.[17] In sum, quadrants 1, 3, and 4 all involve a presumed or potential role for algorithms in decision-making. In these scenarios, the precise form of human involvement must be subject to the results of empirical evaluation. This is the focus of the next stage of analysis, as detailed in Section 5.2.

[16] Relying on solely human decision-making does not mean relying on unchecked discretion. There are many existing mechanisms that constrain—without eliminating—the discretion of street-level bureaucrats (Lipsky, 2010).

[17] Of course, many scenarios will not fit neatly into this 2x2 grid. Table 1 can therefore be seen as identifying poles of a two-dimensional spectrum. For instance, as we move from low-discretion to high-discretion decisions, the threshold for trust in an algorithm in order to consider using it increases accordingly.

Table 1
Summary of roles for human and algorithmic decision-making, based on the need for discretion within the given decision and the trust in the algorithm.

		Need for Discretion	
		Low	**High**
(Relative) Trust in Algorithm	**Low**	1) Primarily or solely human decision-making, with algorithms involved to the extent that rigorous research demonstrates benefits.	2) Solely human decision-making.
	High	3) Primarily or solely algorithmic decision-making.	4) Primarily or solely human decision-making, with algorithms involved to the extent that rigorous research demonstrates benefits.

Each of these three considerations—red lines, the need for discretion, and the trust in an algorithm—involve political judgments and must be evaluated democratically. The legislative developments of recent years demonstrate the importance of public input regarding these central faultlines of debate about algorithmic decision-making. The facial recognition and predictive policing bans in the United States were the result of concerted advocacy in each jurisdiction (Hill, 2021; Miller, 2020). In Europe, advocates and companies have jockeyed over the proper role for red lines. Industry representatives watered down the European Commission's High-Level Expert Group on AI ethics guidelines by removing calls for red lines (Metzinger, 2019). And even though a coalition of 62 civil society organizations preemptively called for the European Commission's AI Act proposal to include red lines (European Digital Rights, 2021a),[18] the proposed Act merely restricts most "high-risk AI systems" (European Commission, 2021). While civil society organizations and regulatory bodies responded to the proposal by calling for more stringent regulations (European Data Protection Board, 2021; European Digital Rights, 2021b), numerous technology companies complained that the Act was too stringent (Johnson, 2021), setting the stage for ongoing debate. Similarly, the extents to which a decision requires discretion and an algorithm is trustworthy cannot be determined without public input. For instance, it is common for vendors and government agencies to overstate the ability of algorithms to appropriately adjudicate decisions (Green, 2019) and to prevent thorough scrutiny of algorithms by hiding behind trade

[18] The coalition's open letter highlighted five categories of AI particularly deserving of red lines: biometric mass surveillance, border and migration control, social scoring systems, predictive policing, and risk assessments in the criminal justice system (European Digital Rights, 2021a).

secrecy protections and subpar record collection practices (Brauneis & Goodman, 2018; Wexler, 2018).

Centering political debate about whether an algorithm is suitable for a decision would shift responsibility and accountability toward the institutional actors that shape government algorithms. Whereas human oversight policies direct attention to human operators, starting with an analysis of an algorithm's suitability directs attention to actors such as government agencies and algorithm vendors. This proposed analysis would compel these actors to justify their choice to implement an algorithm in a given context, making it more difficult for them to direct blame at human operators when the algorithm produces harms.[19] The terms of debate would move upstream, from whether human operators exercised appropriate judgment to whether a given algorithm should be used in the first place.

5.2 Stage 2: Evaluating and Monitoring Human-Algorithm Collaborations

When the first stage of analysis suggests that there might be a role for algorithms in collaboration with humans (i.e., quadrants 1, 3, and 4 in Table 1), policymakers must then turn to the second stage of analysis. This stage involves evaluating whether incorporating algorithms into decision-making can actually improve outcomes. These evaluations must precede any decision to adopt algorithms in a manner that involves human-algorithm collaborations. This stage is intended to address the first flaw of human oversight policies, in which human oversight is presented as a safeguard against algorithmic harms despite minimal empirical evidence that it actually provides reliable protections. Unless an algorithm is intended to operate autonomously, it is not enough to show that the algorithm is reliable on its own. Instead, there must be evidence suggesting that people can oversee the algorithm's functioning and that incorporating the algorithm into decision-making will improve outcomes.

This second stage of analysis involves shifting the burden onto vendors and agencies to affirmatively demonstrate that people can effectively oversee an algorithm. Currently, policies calling for human oversight rarely point to empirical evidence or make empirical claims, reflecting the implicit assumption that human oversight is effective. Given the empirical evidence demonstrating the limits of human oversight, however, the default assumption should be that human oversight is likely to be ineffective, unless proven otherwise. The burden should therefore fall on those proposing human oversight of algorithms to provide affirmative evidence that this mechanism improves outcomes and acts a remedy for concerns about algorithmic decision-making.

[19] This should apply even in cases in which human operators may appear to be at fault for failing to provide proper oversight. If agency leaders choose to implement an algorithm without any evidence that staff can provide the desired form of oversight, then they should not be permitted to foist blame on staff for failing to do a task that never should have been expected of them.

This burden-shifting approach builds on the emerging trend in AI regulation toward requiring proactive assessments of an algorithm's performance and behavior. Several recent bills mandate that agencies or vendors must conduct algorithm impact assessments prior to any public sector implementation of an automated decision system (Brown, 2020; California Legislature, 2021; Government of Canada, 2021). Similarly, the EU AI Act requires providers of high-risk AI systems to conduct an ex ante assessment ensuring that their system conforms to the Act's rules (European Commission, 2021). These requirements represent a burden placed on vendors and agencies, reflecting knowledge of the limits and harms that government algorithms often generate. Although none of these pre-deployment assessments consider the quality of human oversight, legislators should extend these assessments to incorporate evaluations of human-algorithm collaborations.

The central method for assessing human oversight is to conduct experimental evaluations of human-algorithm collaborations. These evaluations should progress in two stages. First, in order to uncover breakdowns in human-algorithm collaboration and experiment with potential remedies, vendors should study how laypeople use the algorithm in a lab setting. Although these experiments would be with laypeople, they can shed light on some behaviors of experts in practice and can be conducted in a quick, low-stakes manner on platforms such as Amazon Mechanical Turk (Green & Chen, 2021). These initial experiments would provide a baseline of evidence regarding how people interact with the algorithm. Second, once preliminary evidence suggests that people can collaborate effectively with the algorithm, vendors and governments should study how practitioners use the algorithm in a lab setting. These experiments can test the mechanisms identified as most effective with laypeople and determine the likely effects of implementing the algorithm.[20] An algorithm should be incorporated into practice only when these proactive evaluations suggest that adopting the algorithm will improve human decision-making and that people are able to perform the desired oversight functions.

Even after an algorithm is adopted in practice, human interactions with it must be subject to ongoing monitoring. Once implemented, algorithms should complement rather than diminish the work of street-level bureaucrats (Pasquale, 2020). Persistent monitoring is particularly important in light of evidence that judicial uses of algorithms can shift over time (Stevenson, 2018) and that street-level bureaucrats' responses to algorithms depend on highly localized details of institutional implementation (Brayne & Christin, 2020). As with the proactive evaluations, this monitoring can be accomplished by extending recently proposed post hoc evaluation mechanisms (Brown, 2020; European Commission, 2021) to incorporate human interactions and oversight. Agencies should be required to collect information about human interactions with algorithms, for instance by tracking overrides of algorithmic decisions to check for racial disparities (Steinhart, 2006).

[20] Strictly speaking, it is possible to skip the first stage and start with the second stage. However, it will generally be beneficial begin with the first stage. Compared to studies with experts, studies with laypeople can be conducted more quickly and with more participants, enhancing the ability to study human-algorithm collaborations and to identify strategies for improving them (Green & Chen, 2021). As we gain deeper knowledge about human-algorithm collaborations, it may become more feasible to start with the second stage.

Furthermore, given that automation can reduce its users' sense of control, responsibility, and moral agency (Berberian et al., 2012; Cummings, 2006), it is also important to continuously monitor whether the algorithm distorts or erodes the moral agency of street-level bureaucrats.

These efforts to evaluate human-algorithm collaborations in specific government settings will be bolstered by research in computer science and related fields. There is growing academic interest regarding how humans make decisions with algorithmic support, leading to much of the evidence presented in Section 4. Further inquiry could provide general insights into mechanisms that improve human-algorithm collaborations and any fundamental limits of human oversight. Of particular relevance is research that considers the design and structure of human-algorithm collaborations rather than focusing only on optimizing algorithm performance.[21] Recent research demonstrates that the design of human-algorithm collaborations can significantly impact outcomes (Green & Chen, 2019b) and that algorithmic accuracy may not lead to the optimal outcomes (Bansal, Nushi, et al., 2021; Elmalech et al., 2015; Green & Chen, 2021; McCradden, 2021). One promising insight along these lines is that cognitive forcing functions (e.g., prompting people to make a preliminary decision before being shown the algorithm's suggestion) can improve human-algorithm collaborations (Buçinca et al., 2021; Green & Chen, 2019b). Future research should evaluate other mechanisms that have been implemented in practice or shown promise in other settings, such as providing training to human operators (Allegheny County Department of Human Services, 2019a), requiring written justification and supervisor approval for any overrides (Allegheny County Department of Human Services, 2019a; Steinhart, 2006), and providing decision support tools rather than specific recommendations (Parasuraman & Manzey, 2010).

6 Conclusion

This study evaluated the global policy trend toward requiring human oversight of algorithms used by governments. By considering human oversight policies in light of research on human-algorithm interactions, I found that these policies suffer from two significant flaws. First, the vast majority of evidence suggests that people are unable to adequately provide any of the envisioned forms of oversight. Second, the incorrect assumption of effective human oversight legitimizes the use of flawed and unaccountable algorithms in government. Thus, rather than enabling governments to attain the benefits of algorithms without incurring the associated risks, human oversight policies justify the inappropriate integration of algorithms into government decision-making and hinder accountability for institutional decision-makers such as agency leaders. In light of these findings, I proposed an alternative approach for determining whether and how to incorporate algorithms into government decision-making. First, governments must determine whether it is actually appropriate to employ an algorithm within a specific context. Second, in settings where they envision a potential role for algorithms alongside human judgment, governments or vendors must perform preliminary evaluations of whether people can collaborate with the algorithm as desired. This

[21] In prior work, I have described this as an "algorithm-in-the-loop" framework for evaluating how people make decisions with algorithms (Green & Chen, 2019a).

proposed process will promote more fine-grained decisions about how governments should use algorithms, helping to ensure that human oversight no longer operates as a superficial salve for fundamental concerns about algorithmic decision-making in government.

As governments incorporate algorithms into how they make consequential decisions, regulation is necessary to ensure that they do so in a manner that avoids producing injustices and violating fundamental legal principles. It is not enough merely to enact regulations, however: policymakers must ensure that their regulations actually provide the desired protections and benefits. Relying on intuitively appealing but ineffective regulation could lead to the worst of both worlds: the underlying problem persists, yet the presence of the regulation leads to the perception that the problem has been solved. Efforts to regulate government algorithms must therefore be particularly attentive to the social contexts in which algorithms are embedded and to empirical evidence about how algorithms influence human decision-making. Only within this type of sociotechnical and evidence-based frame can policymakers develop approaches to using algorithms that promote—rather than undermine—central values of public governance.

Acknowledgments

Thank you to Amit Haim, Amba Kak, Ben Lempert, and Andrew Schrock for insightful feedback on an earlier draft of this manuscript. I am also grateful to Alison Christiansen for providing excellent research assistance.

7 References

Albright, A. (2019). If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions. https://thelittledataset.com/about_files/albright_judge_score.pdf

Alkhatib, A., & Bernstein, M. (2019). Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. https://doi.org/10.1145/3290605.3300760

Allegheny County Department of Human Services. (2019a). Frequently-Asked Questions. https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/FAQs-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-8.pdf

Allegheny County Department of Human Services. (2019b). Impact Evaluation Summary of the Allegheny Family Screening Tool. https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Impact-Evaluation-Summary-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-5.pdf

Amnesty International. (2021). *Ban the Scan*. https://banthescan.amnesty.org

Andrews, D. A., & Bonta, J. L. (2001). The Level of Service Inventory-Revised Manual. *Multi-Health Systems Inc*. https://storefront.mhs.com/collections/lsi-r

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

Arnold Ventures. (2019). Public Safety Assessment FAQs ("PSA 101"). https://craftmediabucket.s3.amazonaws.com/uploads/Public-Safety-Assessment-101_190319_140124.pdf

Article 29 Data Protection Working Party. (2018). Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679. https://ec.europa.eu/newsroom/article29/redirection/document/49826

Australian Human Rights Commission. (2019). Human Rights and Technology: Discussion Paper. https://humanrights.gov.au/sites/default/files/document/publication/techrights_2019_discussionpaper_0.pdf

Austrian Parliament. (2018). Datenschutzgesetz (Data Protection Act). https://www.jusline.at/gesetz/dsg/paragraf/artikel2zu41

Bainbridge, L. (1983). Ironies of Automation. *Automatica, 19*(6), 775-779. https://doi.org/10.1016/0005-1098(83)90046-8

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., & Weld, D. S. (2021). Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(13), 11405-11414. https://ojs.aaai.org/index.php/AAAI/article/view/17359

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., et al. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1-81. https://doi.org/10.1145/3411764.3445717

Belgian Federal Parliament. (2018). Loi relative à la protection des personnes physiques à l'égard des traitements de données à caractère personnel (Law on the Protection of Natural Persons with regard to the Processing of Personal Data). https://www.ejustice.just.fgov.be/eli/loi/2018/07/30/2018040581/justel

Berberian, B., Sarrazin, J.-C., Le Blaye, P., & Haggard, P. (2012). Automation Technology and Sense of Control: A Window on Human Agency. *PLOS ONE, 7*(3), e34075. https://doi.org/10.1371/journal.pone.0034075

Binns, R. (2020). Human Judgment in algorithmic loops: Individual justice and automated decision-making. *Regulation & Governance*. https://doi.org/https://doi.org/10.1111/rego.12358

Binns, R., Kleek, M. V., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1-14. https://doi.org/10.1145/3173574.3173951

Bovens, M., & Zouridis, S. (2002). From Street-Level to System-Level Bureaucracies: How Information and Communication Technology is Transforming Administrative Discretion and Constitutional Control. *Public Administration Review, 62*(2), 174-184. https://doi.org/https://doi.org/10.1111/0033-3352.00168

Brauneis, R., & Goodman, E. P. (2018). Algorithmic Transparency for the Smart City. *The Yale Journal of Law & Technology, 20*, 103-176. https://yjolt.org/algorithmic-transparency-smart-city

Brayne, S., & Christin, A. (2020). Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts. *Social Problems*. https://doi.org/10.1093/socpro/spaa004

Brennan-Marquez, K., Levy, K., & Susser, D. (2019). Strange Loops: Apparent Versus Actual Human Involvement in Automated Decision Making. *Berkeley Technology Law Journal, 34*, 745-772. https://btlj.org/data/articles2019/34_3/02_Brennan-Marquez_Web.pdf

Brown, S. (2020). Data Accountability and Transparency Act of 2020. https://www.banking.senate.gov/imo/media/doc/Brown%20-%20DATA%202020%20Discussion%20Draft.pdf

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1-21. https://doi.org/10.1145/3449287

Buffat, A. (2015). Street-Level Bureaucracy and E-Government. *Public Management Review, 17*(1), 149-161. https://doi.org/10.1080/14719037.2013.771699

Bundestag. (2019). Federal Data Protection Act (BDSG). https://www.gesetze-im-internet.de/englisch_bdsg/englisch_bdsg.html

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 81*, 77-91. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

California Legislature. (2021). AB-13 Public contracts: automated decision systems. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220AB13

Calo, R., & Citron, D. K. (2021). The Automated Administrative State: A Crisis of Legitimacy. *Emory Law Journal, 70*(4), 797-845. https://scholarlycommons.law.emory.edu/elj/vol70/iss4/1/

Campbell, E. T. (Ed.). (2019). *Detroiters want to be seen, not watched*. Riverwise. https://detroitcommunitytech.org/system/tdf/librarypdfs/2019-2206_Riverwise-Surveillance.pdf?file=1.

CBS News. (2021). Police departments adopting facial recognition tech amid allegations of wrongful arrests. *60 Minutes*. https://www.cbsnews.com/news/facial-recognition-60-minutes-2021-05-16/

Chander, A., Kaminski, M. E., & McGeveran, W. (2021). Catalyzing Privacy Law. *Minnesota Law Review, 105*, 1733-1802. https://minnesotalawreview.org/article/catalyzing-privacy-law/

Charette, R. N. (2018). Michigan's MiDAS Unemployment System: Algorithm Alchemy Created Lead, Not Gold. *IEEE Spectrum*.

https://spectrum.ieee.org/riskfactor/computing/software/michigans-midas-unemployment-system-algorithm-alchemy-that-created-lead-not-gold

Christie, G. C. (1986). An Essay On Discretion. *Duke Law Journal, 1986*(5), 747-778. https://scholarship.law.duke.edu/faculty_scholarship/29/

Citron, D. K. (2008). Technological Due Process. *Washington University Law Review, 85*(6), 1249-1313. https://openscholarship.wustl.edu/law_lawreview/vol85/iss6/2/

Cowgill, B. (2018). The Impact of Algorithms on Judicial Discretion: Evidence from Regression Discontinuities. http://www.columbia.edu/~bc2656/papers/RecidAlgo.pdf

Crootof, R. (2016). A Meaningful Floor for "Meaningful Human Control". *Temple International & Comparative Law Journal, 30*, 53-62. https://sites.temple.edu/ticlj/files/2017/02/30.1.Crootof-TICLJ.pdf

Cummings, M. L. (2006). Automation and Accountability in Decision Support System Interface Design. *The Journal of Technology Studies, 32*(1), 23-31. https://scholar.lib.vt.edu/ejournals/JOTS/v32/v32n1/pdf/cummings.pdf

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3313831.3376638

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General, 144*(1), 114-126. https://psycnet.apa.org/doi/10.1037/xge0000033

Dutch Parliament. (2018). Uitvoeringswet Algemene verordening gegevensbescherming (Implementation Act General Data Protection Regulation). https://wetten.overheid.nl/BWBR0040940/2021-07-01

Elish, M. C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society, 5*, 40-60. https://doi.org/10.17351/ests2019.260

Elmalech, A., Sarne, D., Rosenfeld, A., & Erez, E. (2015). When Suboptimal Rules. *Proceedings of the AAAI Conference on Artificial Intelligence, 29*(1). https://ojs.aaai.org/index.php/AAAI/article/view/9335

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies. *Administrative Conference of the United States*. https://www.acus.gov/sites/default/files/documents/Government%20by%20Algorithm.pdf

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

European Commission. (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence

European Data Protection Board. (2021). EDPB & EDPS call for ban on use of AI for automated recognition of human features in publicly accessible spaces, and some other uses of AI that can lead to unfair discrimination. https://edpb.europa.eu/news/news/2021/edpb-edps-call-ban-use-ai-automated-recognition-human-features-publicly-accessible_en

European Digital Rights. (2021a). Civil society calls for AI red lines in the European Union's Artificial Intelligence proposal. https://edri.org/our-work/civil-society-call-for-ai-red-lines-in-the-european-unions-artificial-intelligence-proposal/

European Digital Rights. (2021b). EU's new artificial intelligence law risks enabling Orwellian surveillance states. https://edri.org/our-work/eus-new-artificial-intelligence-law-risks-enabling-orwellian-surveillance-states/

European Parliament, & Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Fight For The Future. (2021). *Ban Facial Recognition*. https://www.banfacialrecognition.com

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. C., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. *Berkman Klein Center Research Publication No. 2020-1*. https://cyber.harvard.edu/publication/2020/principled-ai

French Parliament. (2018). French Data Protection Act. https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000037090394/2018-05-25

Fussey, P., & Murray, D. (2020). Policing Uses of Live Facial Recognition in the United Kingdom. In A. Kak (Ed.), *Regulating Biometrics: Global Approaches and Urgent Questions* (pp. 78-85). AI Now Institute. https://ainowinstitute.org/regulatingbiometrics.pdf

Goodwin, P., & Fildes, R. (1999). Judgmental Forecasts of Time Series Affected by Special Events: Does Providing a Statistical Forecast Improve Accuracy? *Journal of Behavioral Decision Making*, *12*(1), 37-53. https://doi.org/10.1002/(SICI)1099-0771(199903)12:1%3C37::AID-BDM319%3E3.0.CO;2-8

Government of Canada. (2021). Directive on Automated Decision-Making. https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

Gray, M. L., & Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt.

Green, B. (2019). *The Smart Enough City: Putting Technology in Its Place to Reclaim Our Urban Future*. MIT Press.

Green, B. (2020). The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 594–606. https://doi.org/10.1145/3351095.3372869

Green, B., & Chen, Y. (2019a). Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 90–99. https://doi.org/10.1145/3287560.3287563

Green, B., & Chen, Y. (2019b). The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1-24. https://doi.org/10.1145/3359152

Green, B., & Chen, Y. (2021). Algorithmic Risk Assessments Can Alter Human Decision-Making Processes in High-Stakes Government Contexts. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2). https://doi.org/10.1145/3479562

Grgić-Hlača, N., Engel, C., & Gummadi, K. P. (2019). Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW). https://doi.org/10.1145/3359280

Hartzog, W., & Selinger, E. (2018). Facial Recognition Is the Perfect Tool for Oppression. *Medium*. https://medium.com/s/story/facial-recognition-is-the-perfect-tool-for-oppression-bc2a08f0fe66

Henley, J., & Booth, R. (2020). Welfare surveillance system violates human rights, Dutch court rules. *The Guardian*. https://www.theguardian.com/technology/2020/feb/05/welfare-surveillance-system-violates-human-rights-dutch-court-rules

High-Level Expert Group on AI. (2019). Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

High-Level Expert Group on AI. (2020). Assessment List for Trustworthy Artificial Intelligence. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342

Hill, K. (2020). Wrongfully Accused by an Algorithm. *The New York Times*. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Hill, K. (2021). How One State Managed to Actually Write Rules on Facial Recognition. *The New York Times*. https://www.nytimes.com/2021/02/27/technology/Massachusetts-facial-recognition-rules.html

Houses of the Oireachtas. (2018). Data Protection Act 2018. https://data.oireachtas.ie/ie/oireachtas/act/2018/7/eng/enacted/a0718.pdf

Human Rights Watch. (2017). “Not in it for Justice”: How California’s Pretrial Detention and Bail System Unfairly Punishes Poor People. https://www.hrw.org/report/2017/04/11/not-it-justice/how-californias-pretrial-detention-and-bail-system-unfairly

Hungarian Parliament. (2018). Data Protection Act. https://www.parlament.hu/irom41/00623/00623.pdf

Huq, A. Z. (2020). A Right to a Human Decision. *Virginia Law Review, 106*(3). https://www.virginialawreview.org/articles/right-human-decision/

Ibarra, N. (2020). Santa Cruz, Calif., Bans Predictive Policing Technology. *Government Technology*. https://www.govtech.com/public-safety/santa-cruz-calif-bans-predictive-policing-technology.html

Indiana Supreme Court. (2010). Malenchik v. State. *928 N.E.2d 564*.

Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry, 11*(1), 108. https://doi.org/10.1038/s41398-021-01224-x

Jobin, A., Ienca, M., & Vayena, E. (2019, 2019/09/01). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, K. (2021). The Fight to Define When AI Is 'High Risk'. *Wired*. https://www.wired.com/story/fight-to-define-when-ai-is-high-risk/

Jones, M. L. (2015). The Ironies of Automation Law: Tying Policy Knots with Fair Automation Practices Principles. *Vanderbilt Journal of Entertainment & Technology Law, 18*(1), 77-134. https://scholarship.law.vanderbilt.edu/jetlaw/vol18/iss1/3

Jones, M. L. (2017). The right to a human in the loop: Political constructions of computer automation and personhood. *Social Studies of Science, 47*(2), 216-239. https://doi.org/10.1177/0306312717699716

Kiani, A., Uyumazturk, B., Rajpurkar, P., Wang, A., Gao, R., Jones, E., et al. (2020). Impact of a deep learning assistant on the histopathologic classification of liver cancer. *npj Digital Medicine, 3*(1), 1-8. https://doi.org/10.1038/s41746-020-0232-8

Kingdom of Bahrain. (2018). Personal Data Protection Law. https://bahrainbusinesslaws.com/laws/Personal-Data-Protection-Law

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics, 133*(1), 237-293. https://doi.org/10.1093/qje/qjx032

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review, 105*(5), 491-495. https://doi.org/10.1257/aer.p20151023

Koebler, J. (2020). Detroit Police Chief: Facial Recognition Software Misidentifies 96% of the Time. *Vice*. https://www.vice.com/en/article/dyzykz/detroit-police-chief-facial-recognition-software-misidentifies-96-of-the-time

Kolkman, D. (2020). "F**k the algorithm"?: What the world can learn from the UK's A-level grading fiasco. *LSE Impact Blog*. https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/

Lai, V., & Tan, C. (2019). On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 29–38. https://doi.org/10.1145/3287560.3287590

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society, 5*(1). https://doi.org/10.1177/2053951718756684

Lim, J. S., & O'Connor, M. (1995). Judgemental Adjustment of Initial Forecasts: Its Effectiveness and Biases. *Journal of Behavioral Decision Making, 8*(3), 149-168. https://doi.org/10.1002/bdm.3960080302

Lipsky, M. (2010). *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*. Russel Sage Foundation. (1980)

Malgieri, G. (2019). Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations. *Computer Law & Security Review*, *35*(5), 1-26. https://doi.org/https://doi.org/10.1016/j.clsr.2019.05.002

McCradden, M. D. (2021). When is accuracy off-target? *Translational Psychiatry*, *11*(1). https://doi.org/10.1038/s41398-021-01479-4

McQuillan, D. (2018). People's councils for ethical machine learning. *Social Media+ Society*, *4*(2), 1-10. https://doi.org/10.1177/2056305118768303

Metzinger, T. (2019). Ethics washing made in Europe. *Der Tagesspiegel*. https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html

Miller, L. (2020). LAPD will end controversial program that aimed to predict where crimes would occur. *Los Angeles Times*. https://www.latimes.com/california/story/2020-04-21/lapd-ends-predictive-policing-program

National Assembly of Québec. (2020). Bill 64: An Act to modernize legislative provisions as regards the protection of personal information. http://m.assnat.qc.ca/en/travaux-parlementaires/projets-loi/projet-loi-64-42-1.html

National Congress of Brazil. (2019). General Data Protection Law. https://iapp.org/media/pdf/resource_center/Brazilian_General_Data_Protection_Law.pdf

New Jersey Courts. (2017). One Year Criminal Justice Reform Report to the Governor and the Legislature. https://www.njcourts.gov/courts/assets/criminal/2017cjrannual.pdf

Northpointe, Inc. (2015). Practitioner's Guide to COMPAS Core. http://www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf

Office of the Privacy Commissioner of Canada. (2020). A Regulatory Framework for AI: Recommendations for PIPEDA Reform. https://www.priv.gc.ca/en/about-the-opc/what-we-do/consultations/completed-consultations/consultation-ai/reg-fw_202011/

Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors, 52*(3), 381-410. https://doi.org/10.1177/0018720810376055

Parliament of Mauritius. (2017). The Data Protection Act 2017. https://rm.coe.int/dpa-2017-maurice/168077c5b8

Pasquale, F. (2020). *New Laws of Robotics: Defending Human Expertise in the Age of AI*. Belknap Press.

Perrow, C. (1999). *Normal Accidents: Living with High-Risk Technologies*. Princeton University Press. (1984)

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2021). Manipulating and Measuring Model Interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1-52. https://doi.org/10.1145/3411764.3445315

Republic of Argentina. (2018). Ley De Protección De Los Datos Personales (Personal Data Protection Law). https://www.argentina.gob.ar/sites/default/files/mensaje_ndeg_147-2018_datos_personales.pdf

Republic of Kenya. (2019). The Data Protection Act, 2019. http://kenyalaw.org/kl/fileadmin/pdfdownloads/Acts/2019/TheDataProtectionAct__No24of2019.pdf

Republic of South Africa. (2013). Protection of Personal Information Act. https://www.gov.za/sites/default/files/gcis_document/201409/3706726-11act4of2013popi.pdf

Richardson, R., Ed. (2019). *Confronting Black Boxes: A Shadow Report of the New York City Automated Decision System Task Force*. AI Now Institute. https://ainowinstitute.org/ads-shadowreport-2019.pdf

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. *New York University Law Review Online*, *94*, 15-55. https://www.nyulawreview.org/online-features/dirty-data-bad-predictions-how-civil-rights-violations-impact-police-data-predictive-policing-systems-and-justice/

Scharfenberg, D. (2018). Computers can solve your problem. You may not like the answer. *The Boston Globe*. https://apps.bostonglobe.com/ideas/graphics/2018/09/equity-machine/

Schwartz, P. M. (2019). Global Data Privacy: The EU Way. *New York University Law Review*, *94*(4), 771-818. https://www.nyulawreview.org/issues/volume-94-number-4/global-data-privacy-the-eu-way/

Sheriff's Justice Institute. (2016). Central Bond Court Report. https://www.chicagoreader.com/pdf/20161026/Sheriff_s-Justice-Institute-Central-Bond-Court-Study-070616.pdf

Skeem, J., Scurich, N., & Monahan, J. (2019). Impact of Risk Assessment on Judges' Fairness in Sentencing Relatively Poor Defendants. *Law & Human Behavior*, *44*(1), 51-59. https://psycnet.apa.org/doi/10.1037/lhb0000360

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, *51*(5), 991-1006. https://doi.org/https://doi.org/10.1006/ijhc.1999.0252

Slovenian Parliament. (2020). Zakon o varstvu osebnih podatkov na področju obravnavanja kaznivih dejanj (Personal Data Protection Act in the field of dealing with criminal offenses). http://www.pisrs.si/Pis.web/pregledPredpisa?id=ZAKO8157

Solum, L. (2009). Legal Theory Lexicon: Rules, Standards, and Principles. *Legal Theory Blog*. https://lsolum.typepad.com/legaltheory/2009/09/legal-theory-lexicon-rules-standards-and-principles.html

Springer, A., Hollis, V., & Whittaker, S. (2017). Dice in the Black Box: User Experiences with an Inscrutable Algorithm. *The AAAI 2017 Spring Symposium on Designing the User Experience of Machine Learning Systems*, 427-430. https://aaai.org/ocs/index.php/SSS/SSS17/paper/view/15372

Stark, L. (2019). Facial Recognition is the Plutonium of AI. *XRDS: Crossroads, The ACM Magazine for Students*, *25*(3), 50-55. https://doi.org/10.1145/3313129

Starr, S. B. (2014). Evidence-Based Sentencing and the Scientific Rationalization of Discrimination. *Stanford Law Review*, *66*(4), 803-872. https://www.stanfordlawreview.org/print/article/evidence-based-sentencing-and-the-scientific-rationalization-of-discrimination/

Statistics New Zealand. (2018). Algorithm Assessment Report. https://www.data.govt.nz/assets/Uploads/Algorithm-Assessment-Report-Oct-2018.pdf

Statistics New Zealand. (2020). Algorithm Charter for Aotearoa New Zealand. https://data.govt.nz/assets/data-ethics/algorithm/Algorithm-Charter-2020_Final-English-1.pdf

Stein, M. I. (2020). New Orleans City Council bans facial recognition, predictive policing and other surveillance tech. *The Lens*. https://thelensnola.org/2020/12/18/new-orleans-city-council-approves-ban-on-facial-recognition-predictive-policing-and-other-surveillance-tech/

Steinhart, D. (2006). Juvenile Detention Risk Assessment: A Practice Guide for Juvenile Detention Reform *The Annie E. Casey Foundation*. https://www.aecf.org/m/resourceimg/aecf-juveniledetentionriskassessment1-2006.pdf

Stevenson, M. T. (2018). Assessing Risk Assessment in Action. *Minnesota Law Review*, *103*, 303-384. https://scholarship.law.umn.edu/mlr/58/

Stevenson, M. T., & Doleac, J. L. (2021). Algorithmic Risk Assessment in the Hands of Humans. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3489440

Stop LAPD Spying Coalition. (2018). Before the Bullet Hits the Body: Dismantling Predictive Policing in Los Angeles. https://stoplapdspying.org/wp-content/uploads/2018/05/Before-the-Bullet-Hits-the-Body-May-8-2018.pdf

The Republic of Uganda. (2019). The Data Protection and Privacy Act, 2019. https://ict.go.ug/wp-content/uploads/2019/03/Data-Protection-and-Privacy-Act-2019.pdf

U.K. Parliament. (2018). Data Protection Act 2018. https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

UK Information Commissioner’s Office. (2020). Guidance on the AI Auditing Framework. https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

Veale, M., & Edwards, L. (2018). Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling. *Computer Law & Security Review*, *34*(2), 398-404. https://doi.org/https://doi.org/10.1016/j.clsr.2017.12.002

Wagner, B. (2019). Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. *Policy & Internet*, *11*(1), 104-122. https://doi.org/https://doi.org/10.1002/poi3.198

Washington State Legislature. (2020). SB 6280 - 2019-20: Concerning the use of facial recognition services. http://lawfilesext.leg.wa.gov/biennium/2019-20/Pdf/Bills/Senate%20Passed%20Legislature/6280-S.PL.pdf?q=20210513071229

Wexler, R. (2018). Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System. *Stanford Law Review*, *70*(5), 1343–1429. https://www.stanfordlawreview.org/print/article/life-liberty-and-trade-secrets/

Wilson, J. Q. (2000). *Bureaucracy: What Government Agencies Do and Why They Do It*. Basic Books. (1989)

Wisconsin Department of Corrections. (2018). Electronic Case Reference Manual. https://doc.wi.gov/GuidanceDocumentsV2/Reentry/OOS_Electronic%20Case%20Reference%20Manual.pdf

Wisconsin Supreme Court. (2016). Wisconsin v. Loomis. *2016 WI 68*.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2017). Making Sense of Recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403-414. https://doi.org/https://doi.org/10.1002/bdm.2118

Zacka, B. (2017). *When the State Meets the Street: Public Service and Moral Agency*. Harvard University Press.

8 Appendix: Summary of Human Oversight Policies

This table summarizes the 40 policy documents that I reviewed as part of this study. Document Classification refers to the type of policy document (1 = proposed or passed legislation; 2 = policy guidance by government or government-appointed bodies; 3 = manuals, policies, and court cases related to risk assessments used in the US criminal justice settings and the Allegheny County Family Screening Tool). Approach to Human Oversight refers to how each policy presents the appropriate role for human oversight (1 = restricting solely automated decisions; 2 = emphasizing human discretion; 3 = requiring meaningful human input).

Author/Publisher	Title	Year	Document Classification	Approach to Human Oversight
Austrian Parliament	Datenschutzgesetz (Data Protection Act)	2018	1	1
Belgian Federal Parliament	Loi relative à la protection des personnes physiques à l'égard des traitements de données à caractère personnel (Law on the Protection of Natural Persons with regard to the Processing of Personal Data)	2018	1	1
Bundestag (German Parliament)	Federal Data Protection Act (BDSG)	2019	1	1
Dutch Parliament	Uitvoeringswet Algemene verordening gegevensbescherming (Implementation Act General Data Protection Regulation)	2018	1	1
European Parliament and the Council of the European Union	General Data Protection Regulation (GDPR)	2016	1	1
French Parliament	French Data Protection Act	2018	1	1
Houses of the Oireachtas (Irish Parliament)	Data Protection Act 2018	2018	1	1
Hungarian Parliament	Data Protection Act	2018	1	1
Kingdom of Bahrain	Personal Data Protection Law	2018	1	1
National Assembly of Québec	Bill 64: An Act to modernize legislative provisions as regards the protection of personal information	2020	1	1
National Congress of Brazil	General Data Protection Law	2019	1	1
Parliament of Mauritius	The Data Protection Act 2017	2017	1	1
Republic of Argentina	Ley De Protección De Los Datos Personales (Personal Data Protection Law)	2018	1	1

Republic of Kenya	The Data Protection Act, 2019	2019	1	1
Republic of South Africa	Protection of Personal Information Act	2013	1	1
Republic of Uganda	The Data Protection and Privacy Act, 2019	2019	1	1
Senator Sherrod Brown	Data Accountability and Transparency Act of 2020	2020	1	1
Slovenian Parliament	Zakon o varstvu osebnih podatkov na področju obravnavanja kaznivih dejanj (Personal Data Protection Act in the field of dealing with criminal offenses)	2020	1	1
UK Parliament	Data Protection Act 2018	2018	1	1
Allegheny County Department of Human Services	Frequently-Asked Questions	2019	3	2
Allegheny County Department of Human Services	Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County	2019	3	2
Annie E. Casey Foundation	Juvenile Detention Risk Assessment: A Practice Guide for Juvenile Detention Reform	2006	3	2
Arnold Ventures	Public Safety Assessment FAQs (“PSA 101”)	2019	3	2
Australian Human Rights Commission	Human Rights and Technology: Discussion Paper	2019	2	2
Government of Canada	Directive on Automated Decision-Making	2021	1	2
Indiana Supreme Court	Malenchik v. State	2010	3	2
Multi-Health Systems Inc.	The Level of Service Inventory-Revised Manual	2001	3	2
New Jersey Courts	One Year Criminal Justice Reform Report to the Governor and the Legislature	2017	3	2
Northpointe	Practitioner’s Guide to COMPAS Core	2015	3	2
Statistics New Zealand	Algorithm Assessment Report	2018	2	2
Statistics New Zealand	Algorithm Charter for Aotearoa New Zealand	2020	2	2
Wisconsin Dept of Corrections	Electronic Case Reference Manual	2018	3	2
Wisconsin Supreme Court	Wisconsin v. Loomis	2016	3	2
Administrative Conference of the United States	Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies	2020	2	3

Article 29 Data Protection Working Party	Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679	2018	2	3
European Commission	Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)	2021	1	3
High-Level Expert Group on AI	Ethics Guidelines for Trustworthy AI	2019	2	3
High-Level Expert Group on AI	Assessment List for Trustworthy Artificial Intelligence	2020	2	3
UK Information Commissioner's Office	Guidance on the AI Auditing Framework	2020	2	3
Washington State Legislature	SB 6280 - 2019-20: Concerning the use of facial recognition services	2020	1	3