The Right to Explanation*

*This article has benefitted from many critical interlocutors. For comments and discussion on earlier drafts of this article, I'm grateful to David Estlund, Renée Jorgensen Bolinger, Joshua Cohen, Ned Hall, Gabbrielle M Johnson, Geoff Keeling, Niko Kolodny, Seth Lazar, Rune Nyrup, Marco Meyer, Jakob Reckhenrich, Ronni Gura Sadovsky, Kieran Setiya, Lucas Stanczyk, two anonymous reviewers for the *Journal of Political Philosophy*, and audiences at ANU's Philsoc Seminar, the Centre for the Future of Intelligence at Cambridge University, Harvard University's Moral and Political Philosophy Workshop, the MIT philosophy department, Northeastern Univerity's department of philosophy and religion, Stanford University's Institute for Human-Centered AI, and Stanford University's Political Theory Workshop.

ABSTRACT

This paper argues for a right to explanation. The argument is structured according to an interest-based account of rights, where rights are constraints on the discretion of decision-makers to act that are necessary to protect a weighty, widespread interest against standing threats, and come at a tolerable cost. The right to explanation is grounded in the interest in so-called informed self-advocacy, or the ability to represent one's interests and values to decision-makers and to conform one's behavior to a set of rules. Institutional opacity, both in the form of algorithmic decision-making and complex institutional rules, is argued to threaten the interest in informed self-advocacy. Explanations are necessary means to protect this interest, and their provision comes at a tolerable cost. Finally, a new content of the right to explanation is proposed, in the form of rule-based explanations and population-level causal explanations provided by free experts.

1. Introduction

On your thirtieth birthday, you awake to find two police officers standing over you. They inform you that you're under arrest. When you ask what for, they tell you that they don't know, but their boss, the inspector, will be along soon, and he'll be able to tell you. A few hours later, you're ushered into a neighboring tenant's bedroom. The inspector tells you that he doesn't know what you've been charged with, but you're free to go about your life, except that you must report to a hearing on Sunday. And so you move through the criminal justice system, never knowing what crime you've allegedly committed, nor understanding the rules of the process. On the eve of your thirty-first birthday, you're taken away and executed.

Fortunately, this story is not about you, but is instead about Josef K., the fictional protagonist of Kafka's *The Trial*. Unfortunately, a broadly similar story is probably true about you, thanks to the use of algorithms to aid decision-making in a variety of institutions. One such example comes from Washington, D.C. In 2009, the city set out to improve its school system, where only eight percent of eighth graders were performing at grade level in math. It introduced IMPACT, a model to evaluate teacher performance developed by Mathematica

Policy Research. The aim of the model was to identify poorly performing teachers, as input to firing decisions. The model that Mathematica Policy Research developed was complex and proprietary, and teachers couldn't find anyone to explain to them why they'd been fired. In response to the firing of her colleagues on the basis of the model's outputs, Sarah Bax, a math teacher, asked the following: "How do you justify evaluating people by a measure for which you are unable to provide an explanation?".[1]

How, indeed. And answering this question becomes even more pressing in light of the myriad ways in which opaque algorithmic decision-making seriously impacts individuals' life prospects. Job applicants are found, screened, and interviewed by algorithms, with no human input until the final stage of the application process; the success of loan applications is determined by one's credit score; matching algorithms influence the people we date, the news stories we read, and the products we buy. Many of these algorithms are so complex that their outputs cannot be explained to the affected parties, the wider public, or even the decision-makers themselves. Furthermore, the details of the algorithm are often protected by trade secrecy law, adding another layer of inscrutability.

Opaque algorithms threaten to undermine the legitimacy and fairness of the institutions in which they are used. Because Josef K. neither understands why he's been arrested nor the rules governing the criminal justice system, for example, he's unable to do many of the things that are necessary for the proceedings to be legitimate and fair. He cannot appeal the verdict, nor, does it seem, could he have avoided punishment by better conforming his behavior to his society's rules.

In response to the many cases like Sarah Bax's, this paper argues for a right to explanation, on the basis of its necessity to protect the interest in what I call *informed self-advocacy* from the serious threat of opacity.[2] The argument for the right to explanation proceeds along the lines set out by an interest-based accounts of rights (§2). §3 presents and motivates the moral importance of informed self-advocacy in hierarchical, non-voluntary institutions. §4 argues for a right to so-called *rule-based normative and causal explanations*, on the basis of their necessity to protect that interest. §5 argues that this protection comes at a tolerable cost.

[1] O'Neill 2016, p. 8.

[2] This paper's moral case is bolstered by the European Union's 2016 General Data Protection Regulation's (GDPR) arguable establishment of a legal right to explanation (Goodman and Flaxman 2016).

Civil society and regulatory demands for transparency have focused on technology companies and the secret, complex algorithms they develop. While this paper draws heavily on examples of algorithmic opacity, the argument for the right to explanation does not depend on moral complaints unique to algorithmic decision-making. As *The Trial* evinces, bureaucracies can be opaque in much the same way that algorithms can. The argument for the right to explanation applies to any institution that relies on informed self-advocacy for its fairness and legitimacy. Algorithms just make the stakes more vivid.

2. Rights and explanation

A successful argument for a right to explanation must show, firstly, that explanation is a *right*, and secondly, that the right is *to explanation*. I argue that explanation is a right by arguing that the proposed right to explanation satisfies the criteria a particular account of rights, an interest-based account. To show that the right is a right to explanation, the argument must show that the interest is best protected by requiring explanations from decision-makers. This section lays the groundwork for this argument by distinguishing between three types of explanation that might be the content of such a right: intentional, causal, and normative explanation.

I follow Scanlon (2003: 3) in taking rights to be "constraints on the discretion of individuals or institutions to act, which are justified on the grounds that they are necessary and feasible means to prevent unacceptable results that would flow from unlimited discretion." Rights are important sources of control for individuals, as well as protection from interference.[3] A right to freedom of association, for example, offers me control over my associates, and protects me from powerful decision-makers interfering in my gathering of a group of like-minded people to advance an ideal or shared political agenda.

Such constraints, however, are burdensome on decision-makers, and may have adverse effects on third parties and rights-holders themselves. Accordingly, these constraints must satisfy three conditions in order to amount to a right. First, the constraints must protect a morally weighty and widespread interest that is under threat, to justify their imposition. Second, the constraints must be necessary, i.e., they must better protect the interest than alternatives. Third, the constraints must be tolerably costly. I assume a contractualist account

[3] Scanlon 2003, p. 4 and p. 28.

of tolerable cost, where a policy or institution comes at a tolerable cost if it is justifiable to each person, i.e., could not be reasonably rejected.[4] Thus, like any contractualist argument for a right, the argument for the right to explanation faces the following *justificatory burden*: the proposed explanation-related protections for informed self-advocacy must be justifiable to each.

The argumentative strategy for a right *to explanation,* by contrast, does not start by committing to a particular account of explanation. Instead, I will use a functional strategy to identify the explanatory content, by asking what kind of information would enable informed self-advocacy. Still, this content must be a type of explanation.

There are three different types of explanation to which individuals might have a right. I will illustrate the difference between them using the example of explanations of individual action. First, individuals might have a right to an *intentional explanation*. Intentional explanations are psychological explanations of agents' actions in terms of their motivating reasons, or reasons on which they act that they take to count in favor of their action. Second, individuals might have a right to a *causal explanation*. Causal explanations may explain an action in terms of the agent's motivating reasons, but they need not – Aruna's failure to collect her friend Myisha from the airport may be explained by Aruna's illness, which is a physical cause but not a motivating reason.[5] Finally, individuals might have a right to a *normative explanation*, in terms of the normative reasons that count in favor of an agent's action.

With an account of rights and distinction between three types of explanation in hand, we can move on to the argument for the right to explanation. The argument will proceed using the methodology for rights theorizing that flows from Scanlon's account. First, identify the reasonable complaints that people would have in a world without the rights protections (§3), and what is necessary to address their complaints (§4). Then, consider the reasonable complaints that people would have a world where the rights protections are instituted (§5). Finally, make sure the rights protections pass the justificatory burden (§5).[6]

[4] Scanlon 1998, especially Chapter 5.

[5] This paper is neutral on whether intentional explanations of action are causal explanations (Davidson 1963).

[6] Scanlon 2003, p. 4.

3. Informed self-advocacy

The right to explanation is grounded in the interest in informed self-advocacy. To start to get a grip on the nature and moral importance of informed self-advocacy, it will help to dig deeper into the source of the peculiar horror engendered by *The Trial*. The outcome of *The Trial* feels both arbitrary and inevitable, in a way that is deeply connected with the opacity of the criminal justice system in Josef K.'s world. On the one hand, Josef K.'s arrest and punishment strike both him and the reader as unfairly arbitrary. Because the citizens of this autocratic society do not understand the rules of their institutions, they cannot intentionally adjust their behavior in order to comply with the law, a fact which undermines the legitimacy of punishment. On the other hand, the outcome of *The Trial* seems inevitable: given Josef K.'s inability to contest decisions, he is unable to demonstrate his innocence and so is punished.

One of the sources of the horror of *The Trial*, then, is that Josef K. is unable to *do* certain things that are necessary for his institutions to be legitimate and fair. He does not have a say in what the rules governing his society are. He's unable to conform his behavior to the rules, or to contest mistaken or unfair decisions. In other words, he's unable to engage in informed self-advocacy.

Informed self-advocacy is a cluster of abilities to represent one's interests and values to decision-makers and to further those interests and values within an institution. §3.1 taxonomizes those abilities, in terms of representation, accountability, and agency. §3.2 argues that individuals can reasonably reject institutional set-ups in which they cannot engage in informed self-advocacy, because the ability to engage in informed self-advocacy is necessary for hierarchical, non-voluntary institutions to be legitimate and fair.

3.1 Informed self-advocacy

This section explains informed self-advocacy in terms of three central abilities. I will begin with representation. It is, I take it, uncontroversial that individuals have an interest in their interests being taken into account. In the public sphere, individuals have an interest in decisions being made democratically, or by a process in which individuals have equal opportunity to influence decision-making.[7] In the private sphere, individuals generally do not have a claim to equal opportunity for influence, but still have an interest in their interests being

[7] Kolodny 2014, p. 197.

taken into account. This interest is a basic one, exercises of which allow individuals control over the rules that shape their lives. An additional ground of this interest is epistemic: individuals are often the best judges of their interests and values, and thus have good reason to want to be able to represent those interests accurately to decision-makers. One important way that individuals represent their interests and values is by providing input to the content of rules, an example I will return to throughout.

Individuals also have an interest in engaging in informed self-advocacy in the face of existing sets of rules. Such self-advocacy comes in two types: forward-looking exercises of *agency* to navigate systems of rules to achieve one's goals, and backwards-looking exercises of *accountability* to remedy mistakes or unfairness.[8] Individuals have an interest in being able to conform their behavior to a set of rules, which requires forward-looking and temporally extended agency.[9] In order to qualify for a position, for example, the agent needs to take a series of steps over time to gain the required qualifications. Individuals also have an interest in being able to hold decision-makers to account for mistakes or unfairness. This interest is grounded in an interest in living under systems of rules that are predictably and fairly applied. Being able to hold decision-makers to account for mistakes is necessary to engage in robust forwards-looking exercises of agency: it is rational for agents to engage in temporally extended planning only if they are reasonably confident that they can reliably correct mistakes and there is not systemic unfairness that would curtail their plans.

3.2 A morally weighty interest

For there to be a right to explanation, the interest in informed self-advocacy must generate morally weighty complaints if it is not protected. Below, I will argue that the inability to engage in informed self-advocacy generates weighty complaints in hierarchical and non-voluntary institutions, as it undermines their fairness and legitimacy.

Why does the argument for the moral importance of informed self-advocacy focus on institutions? There are some rights, such as the right to life, that are grounded in stable features of human nature, and thus protect interests that people have strong reason to want to be

[8] Two justifications for a requiring administrative decision-makers to give reasons discussed in UK common law are (1) "reasons can provide guidance to others on the body's likely future decisions," and (2) "a reasoned decision is necessary to enable the person prejudicially affected by the decision to know whether he has a ground of appeal" in cases where individuals have a right of appeal on questions of law (De Smith 2020: Part II, Chapter 7, 7-093—7-095).

[9] Venkatasubramanian and Alfano 2020.

protected across various ways of arranging society. Other rights, however, protect interests that people have strong reason to want because they find themselves living in a certain kind of society, in which they are subject to certain kinds of institutions. In societies where private property is required to realize one's rational life plan, for example, individuals have a strong interest in access to credit on fair terms.[10] The interest in informed self-advocacy is the latter kind of interest, one that is generated by the structure of rule-governed, involuntary hierarchies of the institutions that are characteristic of complex modern societies.

In rule-governed hierarchies, rules create "a stable distribution of strictly limited authority in which activities are 'assigned as official duties'."[11] The unequal distribution of authority is the first property of hierarchies that gives rise to a weighty interest in informed self-advocacy. Individuals higher up in the hierarchy have the authority to generate obligations for others to complete certain tasks, backed by sanctions. They also have the power to set goals and standards, and discretion in applying rules to distribute important benefits and burdens. The second property is distributed knowledge.[12] Hierarchies coordinate divided labor to produce goods or offer services at scale, but a by-product of this coordination is that knowledge of institutional rules and relevant facts are distributed throughout the hierarchy.

Because of the asymmetric relations of authority and distributed knowledge within hierarchies, individuals have an interest in being able to engage in informed self-advocacy. Decision-makers may make rules or particular decisions on the basis of the wrong reasons, in ways that are unfair or constitute an abuse of power. And so, individuals have a weighty interest in being able to protect themselves from such arbitrary uses of power by having a say in what the rules are and holding decision makers accountable. Powerful decision-makers also face barriers to fair rule creation due to distributed knowledge, as they have imperfect knowledge of affected parties' interests and values.[13]

However, rule-governed hierarchies can be enormously beneficial, through, for example, increasing efficient production at scale.[14] Arguably, not any difference in power and authority due to hierarchy warrants the costly interventions required to enable informed self-advocacy. In the case of the right to explanation, such costs are warranted when and because

[10] Meyer 2018.
[11] Herzog 2018, p.62, quoting Weber 1968, p. 956.
[12] Herzog 2018, Chapter 6.
[13] Herzog 2018, Chapter 6.
[14] Herzog 2018.

the hierarchical institutions are non-voluntary. Here I include both public and private institutions that unavoidably exercise coercive or manipulative power[15] or distribute justice-relevant goods.[16] The high benefit of participation make the latter non-voluntary.[17]

Such institutions invest decision-makers with power over decisions that seriously and unavoidably impact individuals' life prospects. In order for such institutions to be legitimate and fair, individuals must be able to engage in informed self-advocacy. When individuals cannot avoid being subject to an institution, they must be able to conform their behavior to those rules and correct mistakes for exercises of power to be legitimate. For example, a commonly accepted necessary condition on the legitimacy of punishment is that individuals had the opportunity to conform their behavior to the rules, thereby having it within their control to avoid punishment.[18] Or, informed self-advocacy is also required for fair competition, as individuals need to be on a roughly equal footing to put themselves in a position to compete for scarce benefits, i.e., can equally well conform their behavior to the rules.[19]

The ability to engage in informed self-advocacy, however, is under threat. Kafka, alongside other early twentieth century critics of bureaucracy, was alive to the ways in which bureaucracies could be opaque to those subject to them. This threat is no accident. The same properties of hierarchical, non-voluntary institutions – power and complexity – that give rise to the moral importance of informed self-advocacy also ground the threat to that interest, absent protections. And, with the use of opaque algorithms to inform consequential decisions in a variety of institutions, this threat has taken a new form. For example, opaque AI has introduced novel managerial control mechanisms in the workplace.[20] The automation of processes like firing enables managers to be more opaque about workplace rules and decisions. Platforms can automatically discipline workers, by, say, automatically removing them from the

[15] Anderson 2017.

[16] Firms, for example, distribute an opportunity for saving and a chance to make a social contribution (Gheaus and Herzog 2016).

[17] Scanlon 2003, p. 46 and p. 54.

[18] See Fuller (1965), Hayek (2011: Chapters 9 and 10), and Scanlon 2003, p. 230.

[19] Criticisms of formal equality of opportunity, such as Williams 1962, often focus on the necessity of material resources for substantive equality of opportunity. By contrast, this paper focuses on informational requirements to realize substantive equality of opportunity.

[20] Kellogg, Valentine, and Christin 2020.

platform if their ratings – calculated algorithmically based on real time data – fall below a certain level.[21]

Protections for the ability to engage in informed self-advocacy are thus required to mitigate serious complaints from those subject to opaque decision-making in hierarchical, non-voluntary institutions. The next two sections argue for one such protection: a right to explanation.

4. The right to explanation

A right to explanation is necessary to protect the interest in informed self-advocacy. I first argue for the content of the right, in the form of a claim right to rule-based causal and normative explanations (§4.1). §4.2 argues that without explanations, individuals are not in a good epistemic position to engage in informed self-advocacy. Thus, explanations are typically necessary means to protect informed self-advocacy.

4.1 A claim right to rule-based explanations

The right to explanation is a claim right of individuals against decision-makers in hierarchical and non-voluntary institutions. Following Hohfeld, theorists distinguish between rights as *liberties*, where individuals are free take a certain action, and *claim rights*, where individuals can command others to take certain actions and those others are obligated to comply.[22] The right to explanation as a mere liberty would not protect the underlying interest because individuals would only have access to explanations at the discretion of decision-makers, thus ceding too much control to decision-makers and failing to protect individuals from unfairness or the arbitrary use of power.

Decision-makers are required to provide individuals with *rule-based normative explanations* and *rule-based causal explanations*. As I argue in this and the next section, such explanations are necessary to enable individuals to engage in informed self-advocacy and are tolerably costly.

The term "rule-based explanations" is a term of art. It captures the thought that enabling informed self-advocacy requires an explanation in terms of the relevant rules. Without knowing the relevant rules, individuals are in the epistemic and practical situation of Josef K. and Sarah Bax – they cannot successfully engage in informed self-advocacy, because

[21] Rosenblat and Stark 2016.
[22] Hohfeld 1919.

they can only guess at the rules behind the decision. In order to have meaningful input into the content of rules, for example, one must know what they are. Or, if individuals don't know which of a set of permissible criteria are used to distribute a position, they won't know what to do in order to qualify. Finally, since accountability aims to detect when the rules have been misapplied, explanations that enable accountability must cite those rules.

Different types of explanation, however, better serve different types of informed self-advocacy. Representation and accountability paradigmatically require normative explanations in terms of the relevant rule and supporting normative reasons. Agency, by contrast, paradigmatically requires causal explanations in terms of the relevant rules and causal generalizations that agents can use to satisfy those rules.[23]

To make the case for the necessity of normative explanations, I will use the example of notice and comment rulemaking. Administrative law in the United States and the United Kingdom recognizes that administrative agencies do not merely implement policy, but also make rules that profoundly impact the lives of citizens. Since administrative decision-makers are not formal representatives, administrative law establishes a number of procedures to allow citizens to have input into administrative rules. Notice and comment rulemaking, for example, requires government agencies to notify the public and to seek public comments when they create new administrative rules.

One of the first steps in notice and comment rule-making is the issuance of a notice of proposed rulemaking, which "explains the need, source of authority, and reasons for the proposed rule changes."[24] Why do administrative agencies issue notices of proposed rulemaking when seeking public comments on a new rule? Imagine that you are a United States citizen in late 2017, when the Federal Communications Commission (FCC) sought public comments on its Net Neutrality repeal proceedings. Now imagine if the FCC had asked you,

[23] One might worry that I idealize the extent to which decision-making in institutions is rule-governed (e.g., Haslanger 2018, Zacka 2017). I do take representation and accountability to be responding to the normative properties of rules and their application, paradigmatically. However, the account could be extended to social norms, practices, and other ways of organizing social life that produce stable patterns of behavior for which there are normative reasons for and against. In the case of agency, decision-makers may have more discretion, in which case any rules may be relatively uninformative about patterns individual outcomes. In that case, the causal explanations required to enable agency can just be causal explanations of social outcomes. Thanks to Sally Haslanger for raising this objection.

[24] Federal Communication Commission, from https://www.fcc.gov/about-fcc/rulemaking-process.

a member of the public, the following: "Should the FCC require that all internet service providers treat all transmissions of data over the Internet equally?"

You may not have known what the question meant, or had time to research and deliberate about the matter. And, even if you did have time, you likely wouldn't have had the expertise to know what kinds of factors to consider in deliberating. In other word, you would not be in a good epistemic position to consider the rule and to provide meaningful input on the basis of your deliberation. This, however, undercuts your ability to engage in representation. Individuals have an interest in *informed* influence – they do not merely want to be able to vote on a policy, for example, but want to be able to vote in a way that reflects their preferences.[25]

Normative explanations of why rules are desirable are needed for individuals to provide meaningful input from a good epistemic position. Such explanations should state (1) what the relevant rules are, and (2) the primary normative reasons in favor of the rules.[26] Once provided with such explanations, individuals have a starting point for deliberation, allowing them to think through how the stated reasons relate to their interests and values, the weight of the purported reasons in favor of the new rule, and so on. Normative explanations also enable accountability, by explaining decisions in terms of the relevant rules and the primary normative reasons why those rules were applied appropriately. This claim is defended in §4.2 and §5.

Agency, by contrast, is better enabled by causal explanations. They explain what an agent would have to do to get a desired outcome, in terms of the relevant rules and robust population-level causal generalizations. The latter are an important source of information for individuals to know how to change their behavior in order to get a desired outcome. Such causal generalizations represent the causes of social outcomes, such individuals repaying a loan or not, *not* the causal history of a particular decision.[27] Relevant rules are an important addition

[25] Kolodny 2014.

[26] Rule-based normative explanations are thus different than public justifications (e.g. Gaus 1990, Rawls 1993, and Quong 2011). Public justifications give sufficient reasons in favor of a rule or policy that all citizens should find reasonable (or adequate, or whatever one's preferred account requires). By contrast, the provision of a rule-based explanation is less demanding, as it requires providing the relevant rule and some information about why the rule or its application is appropriate.

[27] The proposal of this paper is thus distinct from proposals to flesh out GDPR's purported right to explanation in terms of information about the decision-making model, often in the form of counterfactuals linking model inputs and outputs (Wachter, Mittelstadt, and Russell 2018).

to such causal explanations because the decision process may privilege one causal pathway over another: a credit scoring system may favor individuals with a lower loan to debt ratio, rather than individuals with a lower debt to income ratio. By understanding what behavior tends to produce social outcomes judged as desirable by the rules, individuals thereby understand what they can change about themselves to better engage in agency with respect to the rules.

By reflecting on what kind of information is required to enable individuals to engage in different types of informed self-advocacy, we see that individuals need to be provided with explanations in terms of the relevant rules and the relevant causal or normative reasons.

4.2 Explanation or evidence?

This section argues for superiority of explanations over other kinds of information in enabling informed self-advocacy. The case for the superiority of causal explanations is straightforward: causal explanations allow agents to intervene on the world, unlike information about correlations.[28] But, one might be skeptical that normative explanations are necessary to enable representation and accountability, as compared to mere evidence of wrongdoing.

Consider the case of Kyle Behm. In 2012 and 2013, Kyle applied for a number of low-skill service jobs for which he was qualified, but was rejected from every job.[29] All of the job applications involved personality tests. A friend told Kyle that he'd scored "red" on the personality test, the lowest score. Kyle had been diagnosed with bipolar disorder about eighteen months prior, but he had not received negative behavioral feedback on previous jobs after his diagnosis. In response, Kyle's father, a retired lawyer, filed a complaint with the Equal Employment Opportunity Commission (EEOC) claiming that the companies had committed hiring discrimination by using personality tests.

Kyle was not given an explanation of his rejections. Instead, he was given some evidence – the information that he'd scored "red" on the personality test – of potential discrimination, and his father used that evidence to file a complaint with the EEOC. That evidence, which came from a friend's testimony, was easier for Kyle to acquire than an explanation of the decision, which he was never given. More generally, one might take existing protections, such as auditing mechanism, as sufficient, as they produce evidence of unfairness

[28] Woodward 2003.
[29] Weber and Dwoskin 2014.

that is communicated to the public, enabling them to engage in accountability. Since explanations are more costly for decision-makers to produce than evidence and are not necessary, the objection goes, there is no right to explanation.

However, consider other versions of Kyle, whose fathers are not lawyers. Would these other Kyles have filed a complaint with the EEOC? More often than not, no, they would not have. Individuals subject to algorithmic decision-making are often not in a good epistemic position to know that there has been a mistake or that they've been treated unfairly, even if they are provided with evidence of a mistake or unfairness. And that is because it is often difficult for non-experts to see that the use of a particular feature in decision-making is objectionable. Consider the use of consumer credit scores in hiring, where they are used as proxies for trustworthiness.[30] It is doubtful that an individual's credit score is a good proxy for trustworthiness, or for desired employee traits and behavior more generally.[31] However, an individual with evidence that she was rejected from a job because of her credit score is unlikely to know what the credit score is a proxy for, nor that credit scores are not good proxies for trustworthiness.

Without the reasons why particular features are taken to be relevant to the outcome, individuals are not in a good epistemic position to contest a decision. What is needed are the reasons why a decision or a decision procedure is appropriate – in other words, a normative explanation.[32] This point also holds for representation, as the discussion in §4.1 of notice and comment rulemaking illustrates: without a normative explanation of a proposed set of rules, individuals are not in a good epistemic position to provide meaningful input. Explanations are thus necessary in part because individuals are not experts regarding the institutions they navigate, and so are not in a good epistemic position to engage in informed self-advocacy. Hierarchical decision-making and legally sanctioned secrecy impose additional epistemic

[30] Traub 2013.

[31] Kiviat 2019, p. 288 discusses the existing literature on correlations between personal financial behavior and employee data.

[32] The Kyle Behm case motivates a general conclusion – that the interests in accountability and representation require normative explanation – based on the interest in being able to hold decision-makers to account who make discriminatory decisions. However, the anti-discrimination protections may be morally unusual in the labor market, in that they require explanations from decision-makers (see footnote 45). Of course, there are legal differences between countries in the stringency of employment protections: UK employment law, for example, requires that employers give particular reasons for a dismissal if requested to do so (Employment Rights Act c. 18, s. 92). Still, the Kyle Behm case does not show that explanations are required of all or most labor market decisions, as this paper is neutral on the exact scope of requirements of fairness and legitimacy in economic domain.

barriers to using evidence to engage in informed self-advocacy. Successful informed self-advocacy requires explanation.

5. The justificatory burden

Thus far, I have argued that explanations are necessary to protect the widespread and morally weighty interest in informed self-advocacy. For there to be a right to explanation, the proposed rights protections must come at a tolerable cost. The final step of the argument addresses this justificatory burden. Would imposing a right to explanation create similar or greater complaints, such that it could be reasonably rejected?

Below, I argue that the right to explanation is justifiable to those subject to the decision process and to decision-makers. To do so, I outline and discharge the most serious justificatory challenge to the right to explanation. This challenge takes the form of a purportedly irresolvable dilemma claiming that a right to explanation will be too costly on any plausible way of spelling out its content.[33]

The first horn of the dilemma considers costs to rights holders. It starts from the observation that explanations need to be both true and intelligible to effectively enable informed self-advocacy. False explanations are instrumentally useless or harmful: individuals will change their behavior but do no better in achieving their goals, misrepresent their values, or fail to hold decision-makers to account. And, if they cannot grasp the relevant explanations, they cannot use them to effectively engage in informed self-advocacy. Unfortunately, there seems to be a tradeoff between intelligibility and effectiveness. On the one hand, more intelligible explanations may be so simple as to fail to capture the relevant factors in enough detail; on the other hand, if informed self-advocacy requires more complex explanations, they will be too costly for all but the most expert – and those who can afford experts – to use. According to this first horn of the dilemma, more complex and accurate explanations are too costly for rights holders to use, but simpler, more intelligible explanations are ineffective. A requirement of either more detailed or more intelligible explanations undermines the right to explanation.

[33] Thanks to anonymous reviewer 2 for pressing this dilemma.

Personalization is a tempting answer to this seeming dilemma: require decision-makers to explain only what will put a particular individual in a position to engage in informed self-advocacy in that context. However, such a requirement would be too costly for decision-makers, requiring them to collect extensive data about individuals, rigorously document the minutia of decisions, and devote resources to delivering intelligible, personalized explanations that are also effective. Thus, it seems as if the only way of escaping the first horn of the dilemma also undermines the right to explanation, by creating intolerable costs for decision-makers.

In sum, the right to explanation seems to run into the following dilemma: either the explanations are abstract enough to be tolerably costly to decision-makers, but the right is undermined because it becomes ineffective for rights bearers; or, explanations ought to be personalized to the rights bearer, to be effective, but the right is undermined because such personalization is intolerably costly for decision-makers.

Rather than undermine the case for the right to explanation, these potentially weighty complaints are an opportunity to further refine the content of the right. The dilemma shows us that a requirement to provide explanations imposes a significant burden on duty-holders, and this burden sets a limit on what the content of the right to explanation can plausibly be. To respond to this dilemma, I will argue that individuals are owed explanations in terms of high-level descriptions of the relevant rules, provided by decision-makers or by free experts.

What do I mean by "high-level descriptions of the rules"? I will first discuss algorithms, and then draw a parallel with organizations. High-level descriptions of an algorithm's functioning provide "functional transparency," or knowledge of the high-level rules of how inputs relate to outputs.[34] Creel (2020) contrasts this type of transparency with two other types: so-called structural transparency, which allows one to know how results are generated by the particular code, and so-called run transparency, which allows one to know how a particular output was produced by running a program on a particular instance, on the basis of the input data.

Some types of algorithms, such as decision trees, are usually judged by experts to be functionally transparent. Other algorithms, such as neural nets, are so complex that it is difficult to understand the rules by which inputs are related to outputs. For such algorithms,

[34] Creel 2020.

computer scientists have developed explainability techniques to increase functional transparency, such as creating a simpler version of the algorithm that has explicit rules for how inputs relate to outputs, and testing whether the simpler algorithm's outputs match the opaque algorithm's outputs, when both are given new inputs.[35] Such explainability techniques create functional transparency without structural or run transparency, as one can know, for example, what algorithm a program instantiates without being able to understand how it is realized in code.[36]

An analogous type of transparency can be used to explain organizational decision-making. In centralized hierarchies, shared sets of rules are needed to coordinate behavior and achieve standardization. Those rules do not fully determine decisions in the organizations, as organizations are only partly centralized decision-making units,[37] and organizational activity is not entirely structured by explicit rules.[38] For example, in the face of cases that fall into grey areas, neither permitted nor forbidden by the rules, decision-makers on the ground utilize heuristics, which may be idiosyncratic or shared amongst one's team or office.[39] But, the rules can still explain and justify individuals' decisions, even if they are approximately true of a complex process, and incomplete normative guides.

The requirement to provide explanations in terms of abstract descriptions of the rules addresses costs to decision-makers: decision-makers need only know what the relevant rules are that apply to the decision, and to provide an explanation in terms of those rules. The complex algorithm or institutional rules can be used for decision-making, reaping the predictive and other benefits of complexity, while the simpler rules can be used to explain decisions, enabling informed self-advocacy.

Further evidence for the tolerable costliness of such explanations comes from existing legal systems. Such requirements are found across different domains and different legal systems, indicating that they are tolerably costly for decision-makers. For example, UK government guidance on discipline and grievance at work requires clear rules and standards of conduct, as well as record keeping about the reasons for any actions taken.[40] Furthermore,

[35] e.g., Bastani, Kim, and Bastani 2017.
[36] Creel 2020.
[37] March and Simon 1993.
[38] Haslanger 2018.
[39] Zacka 2017.
[40] Acas 2015.

explanations are less costly for decision-makers to provide in cases of algorithmic decision-making. As discussed above, computer scientists have developed explainability techniques for complex algorithms, from the tools to create functional transparency discussed above to tools to identify the particular criteria that determine a decision, and to state the weight each criterion has. Given that such explanations are both possible and cheap, the objections of demandingness don't apply. Requirements for algorithms should be at least as stringent – if not more stringent – than the requirements for human decision-makers.

Rule-based normative explanations therefore are tolerably costly for decision-makers. Even for complex algorithms or institutions, it is possible to provide the relevant high-level rules required for rule-based normative and causal explanations. In the case of normative explanations, decision-makers only need add the reasons why the rules are appropriate, or the decision is appropriate in light of the rules. Generating causal explanations requires doing social science. Neither require extensive details of the actual decision-making process, nor intolerably costly personalization.

However, in shaping the right to explanation around costs to decision-makers, we seem to fall onto the other horn of the dilemma. Will such explanations be intelligible and effective?

Causal explanations are effective in virtue of the type of explanation that they are: they inform agents what people with certain properties generally need to do in order to bring about the intended effect.[41] The effectiveness of rule-based normative explanations requires more defense. How can individuals use explanations in terms of abstract rules to correct a mistake or an instance of unfair decision-making? The proposal seems like a non-starter: individuals seem to require detailed information about the actual decision-making process to correct mistakes or unfairness, especially the motivating reasons of decision-makers.

However, explanations in terms of the motivating reasons behind a particular decision are not usually required to enable accountability. That is because, first of all, explanations of how decisions tend to be made, in terms of the rules, are often sufficient to enable accountability. Consider the Australian government's use of a controversial algorithm to detect

[41] The choice of causal explanations of social outcomes also addresses costs to third parties from certain forms of undesirable gaming. Sometimes, rules are stated in terms of proxies for the desired outcome and its causes: a ranking of universities, for example, may rank on the basis of proxies for student welfare, such as access to mental health services and recreational facilities. However, if universities only focused on acquiring those proxy traits to rise in the rankings, these actions may not increase student welfare. Causal explanations address such gaming by encouraging individuals to develop causally efficacious properties rather than the proxies.

and auto-generate notices of welfare fraud, by predicting whether individuals owe the government money because their income was in some fortnightly period was too high, given the level of benefits they received. To calculate fortnightly income, the algorithm takes an individuals' yearly reported income and divides it by twenty-four, the number of fortnightly periods in the year. The algorithm's method of detecting fraud based on the assumption that an individual's income is spread evenly over the year, however, leads it to issue many mistaken debt notices. Many of those who claim welfare benefits have seasonal or precarious work, and thus have fortnightly periods of low to no income and fortnightly periods of much higher income. But the algorithm mistakenly assigns them a higher income in those low to no income periods. Releasing the general method by which predicted debts are calculated, as public awareness campaigns have done, has allowed individuals to correct mistakes without detailed information of why a debt notice was issued in their particular case.

The second line of defense argues that triggering explanatory requirements on the basis of evidence of unfairness is more effective to enable accountability than a process that requires up front explanations in terms of the motivating reasons of agents. Consider, for example, US discrimination law, which applies at all stages of the employment relationship, from hiring to firing. It recognizes that unfairness is difficult to detect in a single case, given that the motivations of decision-makers are often opaque and that discriminatory actions often seem permissible. It therefore relies on the identification of statistical patterns to detect discrimination.[42] If a troubling statistical pattern is detected, then decision-makers are required to justify the hiring practice. Thus, a normative explanation of the decision process is required when there is evidence of unfairness.

There is still, however, the problem of intelligibility. It is neither realistic nor desirable to expect individuals to use explanations to engage in informed self-advocacy within all the institutions that affect them, absent support. Protections for informed self-advocacy need to accommodate individuals' interest in pursuing other plans and interests outside of the relevant institution, i.e., considerations of personal autonomy.[43] Individuals require epistemic support in order to engage in tolerably costly informed self-advocacy.

[42] For example, the EEOC (1979) uses a heuristic to identify adverse selection, which is evidence of discrimination: no hiring procedure should result in members of protected groups being hired at a rate of 80% or less than the group with the highest selection rate.

[43] This argument is based on an argument from Shiffrin 2018, who provides an autonomy-based justification liability for deceptive advertising set up by US law.

Rights holders should be provided with access to free experts and advocates, to aid them in informed self-advocacy. Here, Eubanks' (2018) discussion of a pilot program in Indiana to automate welfare eligibility processes is enlightening. Eubanks quotes Indiana State Senator Vi Simpson as commenting the following:

> People don't know what it means when they get 'failure to cooperate' on a denial notice. In the old days, they used to be able to *call their caseworker* and find out what piece of paper they were missing, or what signature line they forgot to sign, or whatever the problem was. Now they don't have anyone to call.[44]

Bureaucracies are opaque, which is why individuals navigating welfare systems, for example, are assigned caseworkers. Caseworkers function in part as explanation aides: A caseworker could explain to an individual why she received a "failure to cooperate" notice, and what she could do to contest a mistaken receipt of such a notice or to alter her behavior going forward. Caseworkers also sometimes function as advocates, helping individuals to access benefits and entitlements.[45] Since it is costly for individuals to advocate for themselves, they should be provided with free advocates as well as experts.

This additional rights protection is necessary in both just and unjust societies. Any society with a division of labor will have differentially distributed expertise that raises concerns about a fair distribution of the capacity to engage in informed self-advocacy.[46] Free experts and advocates are even more important in unjust societies, because knowledge hoarding is a mechanism by which dominant group members maintain a monopoly on positions of power.[47] And, in both, free expertise and advocacy are important to ensure that everyone is well-positioned to balance their interest in informed self-advocacy with other values, such as autonomy.

The provision of free experts is tolerably costly to decision-makers as well. Much of the information required to be a good expert and advocate is information about how the institution works, rather than about particular decisions. For example, because experts

[44] Eubanks 2018, p. 70.

[45] Zacka 2017. A union representative is another example of this dual role. They explain managers' justifications of policies or rules, and engage in advocacy on behalf of workers, through collective bargaining or by acting as an aid in a grievance process.

[46] Downs 1957. Downs' point can be illustrated by comparing two differently classed professions. Lawyers are paid to develop valuable expertise on how to navigate legal rules. Plumbers, by contrast, are not paid to develop institutional expertise they can use to their personal advantage.

[47] This mechanism is well-documented in the "social closure" literature (Parkin 1979, following Weber 1978).

understand how an institution works and have information about many individuals who interact with an institution, they can more easily identify mistakes or unfairness from patterns of decisions. This lowers costs from data collection and personalization. To detect discrimination, for example, the many counterparts of Kyle Behm could go to a free advocate to register concern if they are systematically shut out of the labor market. If troubling patterns are found, those patterns can be explained to individuals, alongside the relevant rules that seem to be violated. Having a free expert as a "collecting point," e.g., someone to put together information from disparate sources to formulate an explanation, lowers costs to decision-makers and those affected.[48]

One might worry, however, that the above arguments have glossed over a crucial fact about the irreducible opacity of many algorithmic and institutional decision-making processes, one that re-raises the dilemma.[49] Simplified models of very complex algorithms, while intelligible, will not be accurate enough to effectively enable informed self-advocacy. A similar point holds for institutions: they are not games like chess or basketball, with clear and simple rules that participants use to negotiate various moves. And, even worse, such simple models may open up even more scope for unfair decision-making, by enabling decision-makers to engender an illusion of understanding in individuals who grasp a misleading pseudo-explanation.[50] Ruling out such complex decision-processes across the board is unacceptably costly: complex but predictively accurate algorithms, for example, can aid the fair distribution of benefits and burdens.[51] But, allowing them seems to undermine the right to explanation.

I am skeptical that many of these seemingly troubling cases cannot be handled by the provision of free experts, for reasons discussed above. But, let us accept the claim that there will be some complex decision-processes that can only be explained in terms of very abstract and somewhat inaccurate rules. Can the right accommodate this point?

In some cases, the right to explanation ought to rule out opaque decision-making; in other cases, it ought to allow for differing levels of explanatory abstraction in different contexts, without weakening the explanatory requirement so much as to undermine the right. To close this section, I will outline a methodology to guide reasoning about when to take

[48] See Simon 2019, p. 41 on "collecting points."
[49] Thanks to Kieran Setiya for raising this dilemma.
[50] Dimanov et. al. 2020.
[51] Gates, Perry, and Zorn 2002.

which option. I will also argue that neither option undermines the case for the right to explanation.

The first step asks about the stakes of the decision process, and whether it distributes harms or entitlements on the one hand, or benefits on the other. If the stakes are low, the right to explanation does not apply, per the discussion of §2. If the stakes are high and harms or entitlements are distributed, the right to explanation rules out opaque decision procedures where individuals do not have the opportunity to understand how to conform their behavior to the rules. And, there is a stronger case to provide free human experts in decision processes that distribute harms, as the stakes for particular decisions are higher: consider the provision of legal aid in criminal cases, or caseworkers in welfare systems.

If the stakes are high and benefits are distributed, there is more room for abstract explanations. For example, a norm of equality of opportunity does not require that individuals reach an absolute level of understanding of how to conform their behavior to the decision criteria; instead, individuals must be in an equally well-positioned relative to each other. And, since the stakes for any particular decision are lower, free experts could be AI systems designed to generate explanations in response to user queries, with human labor saved for individuals who need to navigate an appeals or other institutional process.

We can see how the distinction between entitlements and benefits grounds a difference in explanatory requirements by comparing US credit and discrimination law. Credit, considered as an entitlement, is subject to explanatory requirements for all adverse decisions.[52] Discrimination law, by contrast, requires explanations when evidence of discrimination is found. This difference in process can be explained by the methodology outlined here: a particular credit decision is higher stakes than a single employer's decision about a job application, say. The rights protection applies in both domains, but the specific duties of decision-makers differ in light of morally relevant facts such as the nature of the good that is distributed.

The second step asks how important informed self-advocacy is for fairness and legitimacy. While the ability to engage in informed self-advocacy is necessary for institutions to be legitimate and fair in any society, it may be more or less weighty under different

[52] For example, US credit law requires creditors to provide consumers with the "principal reasons" for adverse decisions (Barocas, Selbst, and Raghavan 2020).

institutional arrangements. In societies with higher income dispersion, for example, credit may be more important for individuals in lower income brackets to access educational opportunities. In such a society, correcting mistakes in loan applications is more important than in a society where credit is less important for realizing one's life plan, and so requires more personalized explanations. Or, consider a society where the government requires banks to use multiple credit-scoring models. A diversity of decision criteria makes it less important to have personalized explanations regarding any particular set of criteria. While there is a right to explanation across societies with different institutional arrangements, the obligations of decision-makers will vary in terms of the amount of personalization demanded (within reasonable limits).

This section has argued that the right to explanation is tolerably costly for both rights and duties holders. Considerations of cost led to a refinement of the content of the right to explanation: decision-makers are required to provide explanations in terms of the relevant rules, as well as access to free experts.

6. Conclusion

This paper calls for a re-thinking of the informational duties of public and private decision-makers. Once we see that decision-makers have a duty to provide explanations, for example, we have reason to revisit legal protections for opaque decision-making, such as intellectual property law.[53] Or, states and companies may need to invest more resources in providing individuals with free experts – human or algorithmic – to help them navigate hierarchical, non-voluntary institutions.

The right to explanation can seem intolerably costly, requiring human or algorithmic expertise, paperwork, and a reduction in complexity. The variety of existing legal protections that require explanations from decision-makers belie this costliness. Due process is another prominent example, alongside administrative procedure, employment law, and anti-discrimination law, all discussed above. It requires decision-makers to justify their actions by

[53] Burrell 2016.

appeal to suitable reasons in public, with protections backed by a system of appeal to judicial decision-makers.[54] This paper gives a moral grounding for such existing legal protections.

I want to end, though, on a cautionary note. Sometimes, requiring explanations of decision-makers is intolerably costly. One reason is that other legal protections for informed self-advocacy are weak. The right to explanation is one right in a mutually supporting rights package to promote informed self-advocacy. An explanation of why one's state benefits were denied, for example, cannot be used to overturn that decision without an appeals process. So, too, is an appeals process inert if individuals do not know whether they should appeal a decision.

A second reason is that sometimes, it is better to promote fairness and legitimacy by reducing the need for informed self-advocacy, and thus the requirement for explanations. If there are few pathways to a desirable good, for example, advice on how to play by the rules becomes more pressing. In such cases, it may seem as if a right to explanation implies that highly personalized, intolerably costly advice is owed to individuals. Sometimes, however, the best response to an individual's complaint recognizes a serious moral fault with the scarcity of pathways to positions of advantage, and reduces the institutional importance of informed self-advocacy. Explanations are not a silver bullet, and the moral importance of explanations should not distract us from the other fundamental changes that are needed to make our institutions more legitimate and fair.

[54] Scanlon 2003, Chapter 3. The importance of informed self-advocacy partly springs from one of the same underlying concerns as due process protections, namely, a concern about arbitrary decisions.

References

Acas. (2015). Code of Practice on disciplinary and grievance proceedings. Retrieved from https://www.acas.org.uk/acas-code-of-practice-for-disciplinary-and-grievance-procedures/html.

Anderson, E. (2017). *Private Government: How Employers Rule Our Lives (and Why We Don't Talk about It)*. Princeton: Princeton University Press.

Barocas, S., A. Selbst, and M. Raghavan. (2020). The hidden assumptions behind counterfactual explanations and principal reasons, *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*: 80-89. doi: https://doi.org/10.1145/3351095.3372830

Bastani, O., C. Kim, and H. Bastani. (2017). Interpretability via model extraction, eprint arXiv:1706.09773.

Burrell, J. (2016). How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data & Society* 3: 1–12.

Creel, K. (2020). Transparency in complex computational systems, *Philosophy of Science* 87(4): 568-589. Doi: https://doi.org/10.1086/709729.

Davidson, D. (1963). Actions, reasons, and causes, *The Journal of Philosophy, 60,* 685-700.

De Smith. (2020). *Judicial Review,* 8th edition. UK: Sweet & Maxwell Ltd.

Dimanov, B., U. Bhatt, M. Jamnik, and A. Weller. (2020). You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods, *European Conference on Artificial Intelligence (ECAI)*.

Downs, A. (1957). *An Economic Theory of Democracy*. New York: Harper and Row.

Employment Rights Act 1996 c. 18, s. 92 (UK).

Eubanks, V. (2018). *Automating Inequality*. St. Martin's Press: New York.

Federal Communication Commission. "Rulemaking Process." Retrieved from https://www.fcc.gov/about-fcc/rulemaking-process.

Fuller, L. (1965). *The Morality of Law*, revised edition. New Haven: Yale University Press.

Gates, S.W., V.G. Perry, and P.M. Zorn. (2002). Automated loan underwriting in mortgage lending: Good news for the underserved?, *Housing Policy Debate, 13*, 369-391. doi: 10.1080/10511482.2002.9521447

Gaus, G. (1990). *Value and Justification: The Foundations of Liberal Theory*. Cambridge: Cambridge University Press.

Gheaus, A. and L. Herzog. (2016). The goods of work other than money, *Journal of Social Philosophy 47*, 70-89. doi: 10.1111/josp.12140

Goodman, B and S. Flaxman. (2016). European Union regulations on algorithmic decision-making and a 'right to explanation', arXiv:1606.08813.

Haslanger, S. (2018.) What is a social practice? *Royal Institute of Philosophy Supplement* 82: 231-247. doi:10.1017/S1358246118000085.

Hayek, F. A. (1996). *Individualism and the Economic Order*. Chicago: University of Chicago Press.

Hayek, F. A. (2011). *The Constitution of Liberty*. ed. Hamowy, R. Chicago: University of Chicago Press.

Herzog, L. (2018). *Reclaiming the System: Moral Responsibility, Divided Labor, and the Role of Organizations in Society*. Oxford, Oxford University Press.

Hohfeld, W. (1919). *Fundamental Legal Conceptions*. New Haven: Yale University Press.

Kellogg, K., M. Valentine, and A. Christin. (2020). Algorithms at work: the new contested terrain of control, *Academy of Management Annals, 14*, 366-410. doi:

https://doi.org/10.5465/annals.2018.0174.
Kiviat, B. (2019). The art of deciding with data: evidence from how employers translate credit reports into hiring decisions, *Socio-Economic Review 17*, 283-309. doi: 10.1093/ser/mwx030
Kolodny, N. (2014a). Rule over none I: what justifies democracy?, *Philosophy and Public Affairs*, *42*, 195-229. doi: 10.1111/papa.12035
Kolodny, N. (2014b). Rule over none II: social equality and the justification of democracy, *Philosophy and Public Affairs*, 42.
Lewis, D. (1987). Causal explanation. *Philosophical Papers: Volume II*. Oxford: Oxford University Press.
March, J. and H. Simon. (1993). *Organizations*. Hoboken: Wiley-Blackwell.
Meyer, M. (2018). The right to credit, *The Journal of Political Philosophy, 26*, 304-326. doi: 10.1111/jopp.12138
O'Neill, C. (2016). *Weapons of Math Destruction*. New York: Crown Publishers.
Parkin, F. (1979). *Marxism and Class Theory: A Bourgeois Critique*. London: Tavistock Publisher.
Quong, J. (2011). *Liberalism without Perfection*. New York: Oxford University Press.
Rawls, J. (1993). *Political Liberalism*. New York: Columbia University Press.
Regulation (EU) 2016/679 2018 Article 15, 2f (EU).
Rosenblat, A. and L. Stark. (2016). Algorithmic labor and information asymmetries: A case study of Uber's drivers, *International Journal of Communication*, 10, 3758-3784.
Scanlon, T.M. (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.
Scanlon, T.M. (2003). *The Difficulty of Tolerance: Essays in Political Philosophy*. Cambridge: Cambridge University Press.
Simon, H. 2019. *The Sciences of the Artificial*. Cambridge, MA: MIT University Press.
Traub, A. (2013). Credit reports and employment: findings from the 2012 national survey on credit card debt of low- and middle-income households, *Suffolk University Law Review, 46,* 983–995.
The US Equal Employment Opportunity Commission. (March 2nd, 1979). Adoption of questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures, *Federal Register*, *44*. Retrieved from https://www.eeoc.gov/policy/docs/qanda_clarify_procedures.html
Venkatasubramanian, S. and M. Alfano. (2020). The philosophical basis of algorithmic recourse, *FAT* 20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 284–293.
Wachter, S., B. Mittelstadt, and C. Russell. (2018). Counterfactual explanations without opening the black box: automated decisions and the GDPR, arXiv:1711.00399.
Weber, L. and E. Dwoskin. (2014, September 29th). Are workplace personality tests fair, *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/are-workplace-personality-tests-fair-1412044257.
Weber, M. (1947). *The Theory of Social and Economic Organizations*. Trans. Parsons AM, T. Parsons. New York: Free Press.
Weber, M. (1978). *Economy and Society: An Outline of Interpretative Sociology*. Berkeley: University of California Press.
Williams, B. (1962). The idea of equality. In P. Laslett and W.G. Runciman (eds.), *Philosophy, Politics, and Society, Series II*. Oxford: Basil Blackwell.
Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

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