EPISTEMIC INJUSTICE AND DATA SCIENCE TECHNOLOGIES

John Symons and Ramón Alvarado

Forthcoming in *Synthese*
2/19/2022

Abstract

Technologies that deploy data science methods are liable to result in epistemic harms involving the diminution of individuals with respect to their standing as knowers or their credibility as sources of testimony. Not all harms of this kind are unjust but when they are we ought to try to prevent or correct them. Epistemically unjust harms will typically intersect with other more familiar and well-studied kinds of harm that result from the design, development, and use of data science technologies. However, we argue that epistemic injustices can be distinguished conceptually from more familiar kinds of harm. We argue that epistemic harms are morally relevant even in cases where those who suffer them are unharmed in other ways. Via a series of examples from the criminal justice system, workplace hierarchies, and educational contexts we explain the kinds of epistemic injustice that can result from common uses of data science technologies.

1. Introducing Data Ethics and Epistemic Harms

Data ethics is the lively interdisciplinary enterprise that engages critically with normative aspects of software-intensive, data-driven technologies. In this paper, when we refer to 'data science technologies' we use the terms to encompass machine learning, statistical modelling, artificial intelligence, and similar techniques. Data science technologies have become pervasive features of contemporary life and they have obvious moral and political significance insofar as they often

justify important governmental and commercial decisions.[1] On a more personal level, data science provides the ability to monitor and sometimes intervene into the most intimate and important aspects of our lives. Education, sexual relationships, family life, medical care, mobility, commercial activity, and political agency are all vulnerable to interventions from technologies that draw on data science techniques.[2]

Our focus in this paper is on a less familiar kind of harm that these technologies can cause, namely the harm to individual human persons as knowers, interpreters, and sources of testimony. We argue that this is a neglected, but morally significant kind of harm that can result from the design, development and deployment of complex and opaque data-driven technologies such as machine learning, deep neural networks, and big data analysis. We argue that some of the characteristics of these technologies make them harmful to epistemic aspects of our personhood in morally and politically relevant ways. Some of these epistemic harms are unjust and when they are we are obliged to correct or prevent them.

The furtive nature of data collection in contemporary life when combined with our inability to understand or review the operations of the systems that use these data, leave many of us with a sense that these technologies are morally suspect. These technologies exacerbate inequality in potentially worrying ways given that data and the techniques for processing them are not accessible to everyone (Crawford and boyd, 2010; O'Neil, 2016). Corporations and governments collect vast troves of information and have access to proprietary techniques for processing those data. Worries range from effects on privacy and autonomy to questions of distributive justice and social bias. Additional consequences of these technologies include the added difficulty they impose when one attempts to respond to governmental or corporate actions. In many contexts where data science technologies are deployed, we find ourselves with limited recourse or are silenced in ways that strike us as morally wrong and politically oppressive. All of these are legitimate concerns. Another unsettling aspect of these technologies is the inchoate feeling that we are diminished as knowers, interpreters, and sources of testimony. As we will

[1] In some cases, decision makers have no choice but to rely on computational models and simulations as means of determining the best course of action. For a discussion of why this is the case see Boschetti et al. (2012).
[2] Some of these issues predate the computational context. For an account of the development of data-tracking and large-scale record keeping that connects the pre-computational era to present concerns see Colin Koopman (2019). For a conceptual analysis of the development of statistical methods in general, see also Desrosières (1998). In their edited volume *Life by Algorithms*, Catherine Besteman and Hugh Guterson provide an overview of the morally significant effects of what they call 'roboprocesses' on contemporary life (Besteman and Gusterson, 2019).

show, understanding how uses of these technologies can count as instances of epistemic injustice can help us to articulate the vague impression that we are sometimes wrongly diminished by these technologies even in cases where they otherwise seem to be working for the good.

One of the distinctive features of data science techniques is that they typically involve levels and kinds of epistemic opacity that make it difficult to question their results or to understand their operation (Alvarado 2020; 2021a; 2021b). For reasons we will explain below, this characteristic means that data science is liable to result in epistemically unjust harms. When used in important decision-making contexts, this technology can introduce significant epistemic disadvantages. This changes the character of the relationships between commercial and governmental decision makers and those who are subject to their decisions. For example, epistemic opacity can make it difficult for a prisoner to appeal an automatically generated decision on their parole, it can leave a borrower unable to understand a decision with respect to their credit-worthiness, and it can make it hard for an employee to understand the ways their work is evaluated. When uses of data science technology result in unjust epistemic disadvantages or illegitimate diminutions of epistemic status they are morally blameworthy.

Data science technologies influence how people perceive and judge others and in particular they can shape how individuals are ranked with respect to their epistemic capacities. We will show how these technologies can systematically empower administrative and corporate decision makers in unjust ways and can intrude on the personal lives of individuals in ways that alter our relationships and make it difficult for us to find, redress and correct errors. The ways in which we judge the cognitive and hermeneutical worth of one another involves morally and politically significant questions (Bratu and Haenel, 2021). These judgments are inextricably tangled up with power and social position. It matters to our standing as persons *that* we are capable of knowing, *what* we are capable of knowing, and *what others think we know*.[3] How we navigate (or are allowed to navigate) the social and political worlds we inhabit is partly dependent on judgments with respect to our epistemic capacities. These judgments are constitutive of what we call one's *epistemic standing*. In this paper we assume that one's epistemic standing has implications for how one relates to others in communities, within

[3] There are formal features of judgments with respect to the collective aspects of knowledge, for example, common knowledge, that epistemic logicians have shown are foundational for inclusion in certain kinds of norm-governed social behavior. Participation in some norms seems to require that one is judged capable of sharing in common knowledge (see Rendsvig 2021).

institutional hierarchies, in commercial transactions, in legal proceedings, sexual relationships, friendships, or family life. Our approach to these topics is shaped by the assumption that one's dignity as a person is constituted to a significant extent by one's capacity to understand, to reason, to interpret, and to experience. Thus, on our view the introduction of obstacles to understanding, to interpreting, or to experiencing can be harmful to us and in some cases these harms will be morally blameworthy.

Using real examples and imagined scenarios, we will explain how data science technologies can serve to create and amplify conditions in which people suffer epistemic harms and epistemic injustice. We will explain and defend the concept of epistemic injustice in more detail below. For now, we simply note that an epistemic injustice takes place when one's standing as a knower or one's credibility as a source of testimony is incorrectly reduced in virtue of one's marginal or subordinate social position. One of our tasks in this paper is to explain why the illegitimate diminution of one's testimony and perspective that can result from the effects of data science technologies sometimes counts as a form of injustice regardless of the distribution of other material or social goods.

By highlighting some of the ways that data science is liable to epistemically unjust consequences we hope to add some new considerations to the ongoing work of data ethicists. Data ethicists have engaged with data science technologies in roughly four ways. They have critically examined practices around the acquisition or curation of data (Benjamin, 2019; Jo and Gebru, 2020), they have exposed and criticized ways that these technologies are abused or wrongly deployed (O'Neil, 2016; Noble, 2017), they have revealed their susceptibility to various kinds of error in the "data analysis pipeline" (Suresh and Guttag, 2019; Horner and Symons 2020), or they have zoomed out from the technology itself to explore important political considerations such as the social power dynamics surrounding the design, development, and use of data science technologies (Amoore, 2020; Kalluri, 2020). Special attention has been devoted to the harmful consequences of socially biased data in machine learning contexts (Mehrabi et al., 2021). For the most part, recent discussions in data ethics have focused on practices that might ensure that the technology works correctly and on ensuring that it is not misused in ways that cause harm to individuals (Buolamwini and Gebru, 2018). Data ethicists have been guided by

the goal of ameliorating the impact of mismanaged uses of data and avoiding the socially biased uses of the technology (O'Neil, 2016; Butterworth, 2018; Noble, 2017; Benjamin, 2019).[4]

There is no doubt that data science technologies can deliver significant enhancement of our epistemic capacities and as such are an undeniable benefit to their users. And yet, as data ethics scholars have noted, in practice, these enhancements are being distributed and deployed unequally. There are a variety of kinds of injustice that can result from inequitable access to these technologies. However, in the pages below we conceptually distinguishing epistemically unjust actions from those that result in some harm with respect to the distribution of material goods or social status. It is certainly the case that epistemic injustice typically occurs in social contexts where illegitimate social biases are prevalent, where individual autonomy is not respected, or where other harms co-occur. Nevertheless, it is important to distinguish these other harms from epistemic harms. [5]

2. Epistemic injustice

In this section we provide an overview of the concept of epistemic injustice as originally conceived by Miranda Fricker (Fricker 2007; Kidd et al., 2017). In the first part of this section, we defend her conceptual framework as capturing a distinct kind of harm done to persons in their capacity as knowers.[6] Our approach takes epistemic injustice to be a discriminatory injustice,

[4] Notable exceptions are philosophers of technology who do not draw a conceptual line between social contexts and the technical artifacts that emerge from or are deployed within it (See Latour 1988, written under the pseudonym of J. Johnson; Simondon, 2017; Slater, 1980). For these philosophers, the ethical aspect of technological development is not addressed by exploring the relationship between societal harm and technological use, but rather by exploring the values and interests of society that bring these technologies into being. Hence, scholars that follow this approach, such as Amoore (2020), claim that "the algorithm already presents itself as an ethicopolitical arrangement of values, assumptions, and propositions about the world and cannot/should not be analyzed on its own". Similarly, Green states that "Data scientists must recognize themselves as political actors engaged in normative constructions of society and, as befits political work, evaluate their work according to its downstream material impacts on people's lives" (2020). While we agree that treating technology in isolation from its societal context is limited, in this paper we also defend conceptual distinctions that allow us to identify harms particular to a specific technology independently of the settings in which they were developed or deployed. See our discussion of Jeroen van den Hoven's (2000) taxonomy of the kinds of moral wrong-doing associated with different kinds of technologies in Section Five of this paper.

[5] See Keyes et al., 2019 for a thorough assessment of how an algorithm that is Fair, Accountable, Transparent, *and* used for good can nevertheless have ethically worrisome implications. Keyes (2020) also offers valuable insight as to how the data sciences can influence discourse about 'knowers.'

[6] It is important here to note the originality of Miranda Fricker's contribution while contextualizing it with other works that also deploy or function within an epistemic framework and that are an important contribution to the understanding of epistemic harms. Fricker's original contribution lies in the fact that her account of epistemic

which is sometimes, but not always, accompanied by distributive harms.[7] We begin by focusing on how and when such harms occur to individuals in communities that were already vulnerable to other kinds of harm and injustice in virtue of their social position. However, we also argue that, under this discriminatory (rather than distributive) framework, the category of epistemic harm has a broader application and can be applied to individuals at many different social levels, from prison inmates to university faculty. We provide a set of examples showing how data science methods can cause both hermeneutical and testimonial epistemic injustices. The primary reason that these methods are liable to being harmful, we argue, is due to their opacity and their inability to permit corrective recourse.

We follow the example of the early discussions of epistemic injustice by considering people who are most vulnerable to governmental, institutional, or corporate authorities that deploy data science. Our first examples focus on prison inmates and members of oppressed racial minorities. As our discussion proceeds, we consider situations in which relatively privileged persons can be subject to epistemic injustice. Thus, while harms to persons belonging to minority and marginal communities serve as our starting point, we will argue that epistemically unjust harms can potentially befall anyone who is in a subordinate relationship to the kinds of entities or persons capable of effectively deploying data science technologies.

Although there is a substantial literature on the concept of epistemic injustice, it exhibits disagreement about the nature of epistemic harm itself. In many influential views, for example, the term is deployed to capture unjust social obstacles to epistemic goods such as education or information (Coady, 2017). Elsewhere the term is used as synonymous with harms that *result from* epistemic prejudices such as the denial of a job, or unequal pay for equal work (see Green,

injustice sought to identify the possible harms directly related to the unjust diminution of an agent's epistemic status, due in part to irrelevant social factors (we thank an anonymous reviewer for helping us emphasize this point). As we will see below, similar concepts such as *epistemic violence* (Dotson, 2011) sought to capture and account for a different set of phenomena such as *physical or social harms* done to agents in virtue of epistemic reasons or elements such as ignorance. Similarly, as we will see in detail below, distributive accounts of the exact same term 'epistemic injustice' seek to identify social harms and obstacles such as poverty or segregation that *result in an unjust distribution of an epistemic good* such as education. A social harm that has an epistemic source and an unfair distribution of epistemic goods that was caused by a social harm, though often contiguous or related, are not conceptually the same phenomenon as a discriminatory diminution of an agent's epistemic status. It is this latter phenomenon that Fricker's framework brings to the fore (2017) and it is this discriminatory account of epistemic injustice that best fits the kind of phenomenon we seek to account in the context of our interactions with data science technologies and methods.

[7] As we will see, this is an important distinction made by Fricker herself (2017) that addresses a prevalent conflation in the literature between discriminatory harms of an epistemic nature and distributive asymmetries of epistemic goods such as education or other observable harms stemming from epistemic motives such as ignorance or so.

2020). In the following subsection, we defend a view that emphasizes discriminatory rather than distributive injustice as essential to the concept of epistemic injustice. Discriminatory injustices can happen even when a distributive injustice does not, and because of this we argue that an excessive focus on distributive injustice risks ignoring this other important and distinctive kind of harm associated with data science technologies.

2.1 Epistemic Injustice as a Discriminatory Rather than a Distributive Harm

Fricker argued that there exist distinctively epistemic kinds of harms that, under some circumstances, count as unjust (2007). She asks us to consider the indignation felt by those who are unjustly minimized as worthy or credible contributors to a debate or conversation for epistemically irrelevant reasons - because of sexism, racism, or classism for example. Indignation, can often be the first signal that there is a harm associated with an epistemic injustice that is not reducible to the unjust distribution of social status or material goods. Even if there are no other material or social consequences at stake, an indignant reaction to having one's perspective or epistemic capacities ignored or diminished is understandable. The reaction is warranted insofar as having one's testimony or interpretation ignored or diminished for no reason, for no good reason, or incorrectly is straightforwardly harmful.[8] An instance where one's standing as an epistemic agent is diminished primarily or solely in virtue of one's social position, Fricker argues, would counts as an unjust harm.

Notice that if one's contribution to a conversation is ignored because of the chauvinism of one's interlocutor, it is likely that one will be upset or angered not simply because one's interlocutors have chauvinist beliefs. Instead, one's indignation is likely to be due primarily to one's being *devalued* in a particularly personal way – qua epistemic agent. That harm is unjust, Fricker argues, insofar as it is motivated by bias towards the group to which one belongs. Thus, already in Fricker's analysis we find the distinction between the harm to an important aspect of

[8] It is important to note that some epistemic injustices are so extreme that scholars, such as Medina (2017), use the term 'hermeneutical death' to refer to them. These are instances in which there is a total erasure of an agent's voice and capacity to engage in meaning-making cultural activities. In hermeneutical injustices, the agent has difficulty articulating the harm they are being subjected to and in some cases may not even recognize their diminished condition as a harm. We thank one of our anonymous reviewers for suggesting that in these extreme cases, individuals will not be indignant.

one's personhood and the judgment that this is an unjust harm in virtue of the biased motivation of the source of the harm.

While Fricker provides the first explicit defense of the concept of epistemic injustice, other thinkers who explained the epistemic marginalization and silencing of minority communities informed her work. There is a rich tradition of discussion around the political and social role of knowledge, testimony, and interpretation. Kristie Dotson (2011) connects this previous work with questions of epistemic justice and explains how the testimonial standing of members of marginalized communities has sometimes been illegitimately diminished through silencing.[9]

It is important to recognize that one's status as a knower, or as a witness, can be harmed even when one makes gains or increases one's status in some other social or material dimension. One can, for example, be helped by another person because the helper prejudicially deems the recipient of assistance incapable of helping themselves. This would be an instance of a kind of benevolent condescension. Or consider how one might provide an explanation in some circumstances to another person because one regards them as incapable of providing their own equally credible explanation. In cases such as these, those who intervene on another's behalf may indeed be helping, but they are doing so because of the incorrectly diminished epistemic status of the person they are seeking to help. Thus, an illegitimate diminution of their epistemic standing may still obtain.

In some circumstances, the recipient of the benefit might decide that the reduced epistemic standing is not significant, or that it is somehow offset by other benefits. However, in order to be in a position to make such judgments in a case-by-case basis or to be able to reason competently about such tradeoffs, one must first recognize harms to one's epistemic standing. The framework of epistemic injustice helps us to identify, evaluate, and reason about such diminutions of epistemic standing.

In coining the term *epistemic injustice* Fricker indicated that her aim was to "delineate a distinctive class of wrongs, namely those in which someone is ingenuously downgraded and/or disadvantaged in respect of their status as an epistemic subject." (Fricker, 2017, p. 53). According to Fricker epistemic injustice is primarily a kind of *discriminatory injustice*. This is in

[9] See also Gayatri Spivak's discussion of what she calls 'epistemic violence' against subaltern groups (2003). These ideas have been developed further by Patricia Hill Collins (2017).

contrast to other views in which epistemic injustice is sometimes treated as the presence of social and financial obstacles to the acquisition of epistemic goods such as education or information (Origgi and Ciranna, 2017, see also Coady, 2010; 2017). Epistemic injustice, on this interpretation, is a form of distributive rather than discriminatory injustice. When, for example, members of oppressed groups are denied proper schooling or access to libraries in virtue of their social status they are unfairly deprived of epistemic goods. In this case the epistemic good of schooling or libraries is distributed unfairly. However, a focus on distributive aspects of epistemic injustice risks neglecting the kind of discriminatory harm that Fricker identified. While it is typically the case that discriminatory and distributive injustices occur together this is not always or necessarily the case. Fricker's account of epistemic injustice as a *discriminatory* injustice and not solely a distributive one illuminates a special kind of harm or degradation that we can suffer qua epistemic agents.[10] One can suffer a discriminatory injustice, in other words, even when the distribution of goods one has access to is equal to or more than others receive.

In order to illustrate this, we can imagine cases in which the allocation of goods is more favorable for a person, *because* they are subject to an epistemic injustice. For example, in *Everybody Hates Chris*— the semi-autobiographical TV series based on the life of comedian Chris Rock— the main character, Chris, consistently receives special attention from his teacher because she believes that as a black teenager in a predominantly white middle school, he has diminished epistemic capacities. The show makes clear that the teacher's stereotypical assumptions are mistaken yet she continuously provides Chris extra instructional support because she assumed, for example, that Chris did not know his father, that his father was homeless, or that he had never had the chance to travel outside of his neighborhood.

In this example, we can see that the distribution of goods is distinguishable from the discriminatory actions of his teacher. Chris is receiving *more* assistance, attention and goods *because* he is prejudicially deemed less capable of a knower than others. Chris was undervalued

[10] In this sense, the concept of epistemic injustice differs substantially from terms such as 'epistemic violence' used by Kristie Dotson, for example. In particular, while the term 'epistemic injustice' entails an unjust diminution of someone's epistemic status— whether by a discriminatory diminution of their testimony or by a systemic neglect of or the imposition of obstacles to epistemic participation — the term 'epistemic violence' picks out the wrongdoers pernicious ignorance. By contrast, 'epistemic' as an adjective in the term 'epistemic injustice' picks out the aspect of the agent whose status is being diminished. See Dotson, 2011 p.240, especially her example of the pyromaniac toddler. This distinction seems to also apply to other neighboring concepts such as "discursive harms" (Congdon, 2017; Keyes, 2020).

in virtue of his social position despite consistently showing capacities equal or superior to his peers in class. However, this undervaluing resulted in the teacher providing *extra* resources on his behalf. Thus, discriminatory harms can happen independently of distributive benefits and harms. Hence, someone may be discriminated against even when the overall distribution of goods ends up benefiting them, as is the case with biased assessments of competence related to gender that are categorized either as *benevolent sexism* (Glick and Fiske, 1997) or *white knighting* (Ruiz, 2017).

Consider another occasion that more clearly illustrates the point above. Unbeknownst to Chris, his teacher applied on his behalf for a scholarship reserved for children belonging to single-parent, renting households. When Chris informs his teacher that in fact he does have a two-parent household and that they are homeowners, the incredulous response from his teacher is "Sure you do." Here, Chris is being incorrectly judged an unreliable source of testimony concerning his family life and finances in virtue of his social position. This is an instance of what Fricker called *testimonial epistemic injustice*. Testimonial injustice will figure centrally when we discuss cases where people are illegitimately taken to be inferior arbiters of truth due to their perceived inferiority to an automated data analytics system.

Another form of epistemic injustice identified by Fricker is what she calls *hermeneutical injustice*. Hermeneutical injustices are those in which a socially unjust arrangement is embedded in the operation of a social system in such a way that its victims are incapable of understanding that a harm has been inflicted on them. This kind of injustice can result from the design of institutions, bureaucratic processes, and policies, but it can also result from biases embodied in cultural practices or traditions (Medina, 2017; Bratu and Haenel, 2021). As we will show here, hermeneutically unjust systems can result from data science methods, practices and technologies.

It is important to note that from Fricker's perspective these harms should be distinguished from instances of deliberate, manipulative malice. While instances of malicious actions could certainly diminish the epistemic status of an agent, it is not a necessary condition for epistemic injustice that it results from malicious intent. To illustrate this, she offers the example of three people, one of whom wants to deliberately diminish the credibility of the second in the eyes of the third. In her example the harm obtains in the interaction between the second and the third person. Specifically, the harm is inflicted by the third person on the second. Because of

deception, the third person in this example might be blameless with respect to the harm against the second. Note also that the first person, the malicious deceiver wishing to reduce the epistemic status of the second person, does not believe that the third person in fact has a diminished epistemic status. Fricker insists we ought to keep the term epistemic injustice 'strict' enough that it can capture the kinds of harm that the third person does to the second. Furthermore, we should be able to distinguish such harms from the kind that the deceiver in this example is engaged in (Fricker, 2017 p.54). In the eyes of a third person, the second has a diminished epistemic status. In the eyes of the first person, the epistemic status of the second person remains unchanged.[11]

As we have seen, Fricker explains the various forms of epistemic injustice as they arise in the individual lives of socially marginalized or oppressed people. Elizabeth Anderson urged philosophers to extend Fricker's account beyond individual interactions so as to explore principles for the cultivation of epistemically just social and political institutions (2012). The rest of our paper responds to Anderson by showing how technologies, as well as social and political institutions can be more or less epistemically just.

3. Opacity in data science technologies can lead to hermeneutical epistemic injustice

In this section we explain the opacity inherent to many data-science methods, particularly methods such as machine learning, deep neural networks and big data analytics. The opacity at play in data-driven technologies often results from procedures that are so complicated or complex that understanding the nature and sources of associated harms becomes challenging. As we will see, when knowers lack a conceptual framework to identify, name or understand harms they experience, they may be subject to a hermeneutical epistemic injustice. In this section we offer two examples of knowers in very distinct social circumstances but facing very similar data-driven technologies.

Opacity in machine learning systems and other computational methods has been well documented in an extensive body of literature (Humphreys, 2009; Crawford and boyd, 2010; Symons and Horner, 2014; Symons and Alvarado, 2016; Kaminski et al., 2017; Alvarado and

[11] Injustices can also happen in the event of wrongful ascription of an *inflated* epistemic status. We thank our anonymous reviewer for encouraging us to mention this point.

Humphreys, 2016; Burrell, 2016; Pascale, 2015; Alvarado, 2020; 2021a; 2021b). In this section we explain how such opacity can contribute to both hermeneutical and testimonial forms of epistemic injustice as articulated by Fricker.

The term 'epistemic opacity' was introduced by Paul Humphreys (2004) to characterize the epistemic inaccessibility of the underlying processes and properties of some computational systems. Humphreys focused on computational models and simulations, but the concept applies broadly to computational methods in general and to data-intensive computational methods in particular (Symons and Alvarado, 2016; Alvarado, 2020). In addition to the formal characteristics of computational processes, social phenomena, such as division of labor, proprietary limitations on access, intentional secrecy, etc. can also contribute to making something epistemically opaque. In these instances, what is at play is a kind of social epistemic opacity (Kaminski et al., 2017; Burrell, 2016). Philosophers (Kaminski et al., 2018; Alvarado, 2020) and sociologists of science and computation (Saam, 2017) have provided accounts of the various ways in which particular computational processes, such as computer simulations and machine learning algorithms can be opaque.

A given device, method or process can be opaque simply on the basis that its inner processes are beyond what any given human agent can reasonably track and understand (Humphreys, 2004; 2009). This is particularly the case when big data and associated methodology is involved (Alvarado, 2020; Burrell, 2016; Symons and Alvarado, 2016). A system, device, or set of processes that is opaque in this manner—beyond the epistemic resources of any one individual — has what Humphreys called essential opacity (2009). Barberousse and Vorms (2014), for example, mention that in circumstances in which complex software is used for scientific inquiry, opacity arises when those using it do not have access to the underlying processes by which the software or the software-intensive instrument works or arrives at its results. This potentially undermines confidence in the results of such inquiry since in their view "the scientist's partial blindness to the details of the computational process seem to result in a serious lack of epistemic control upon the empirical validity of its outputs" (Barberousse et al., 2014).[12] Data-intensive computational processes are often epistemically

[12] The sense of trust being discussed here is one close to the epistemology of science and in particular to the epistemology of computational methods in the sciences. Hence, we can call it 'scientific trust'. In such settings, important debates are taking place regarding the relationship the term trust has to reliability, explanatory understanding and transparency. For a thorough though deflationary account of these efforts, see Durán and

opaque in this stronger sense. This is due to a number of factors including the complexity and complication of the software required to run the kinds of analytics that yield meaningful insights from vast amounts of data.[13] In addition, opacity is almost intractably difficult in many machine learning and artificial intelligence contexts that depend on neural networks approaches.[14] Finally, as Boschetti and Symons argue, computational systems are typically marked by irreversibility that often makes the history of the processes involved opaque in principle.[15]

The different kinds of opacity mentioned so far are often in place simultaneously in these systems. A given predictive algorithm, such as the ones used to assess risk in, for example, legal or financial settings, may be opaque in any or all of the ways discussed above. As we will explain in more detail below, when it comes to risk assessment algorithms used in United States courts to predict recidivism, for example, researchers have identified opacity resulting from social, engineering, or mathematical characteristics of these systems (Amoore, 2011; 2014; boyd and Crawford, 2012; Kitchin, 2014). Individuals who are subject to the action of these technologies are typically unable to assess their operation. However, the ways by which such systems produce their results may be inaccessible not only to those of us who are affected by it but, sometimes, even to employees of the companies that produce them (O'Neil, 2016; Burrell, 2016; Alvarado and Humphreys, 2017).

Epistemic opacity can generate instances of both testimonial and hermeneutical epistemic injustice as we will see below. Very roughly, because such computational methods are often opaque in a way that prevents us from simply looking 'under the hood' they leave us without epistemic recourse and therefore vulnerable to being diminished or excluded from seeking justification or participating in decisions.

Formanek (2018) and Durán and Jongsma (2021). Hence, some theoretical frameworks assume that trust within scientific inquiry is necessarily linked to transparency and should not be otherwise. See Symons and Alvarado (2016) and Alvarado (2021a) to see how and why this applies to issues of data science in particular and Alvarado and Symons (2019) to see why this applies to other software-intensive technology in scientific inquiry and policy-making. We thank our anonymous reviewers for encouraging this clarification.

[13] Path complexity in the execution of code is an obstacle to software verification and also poses a challenge to the surveyability of systems that rely on software intensive processes. For details see Symons and Horner (2014; 2019)

[14] For a detailed account of opacity in these kinds of systems, see Burrell (2016), Alvarado and Humphreys (2017), and Alvarado (2021a).

[15] By irreversibility Boschetti and Symons mean the fact that computational models can generally arrive at the same state via many possible sequences of previous states. "Thus, while in the natural world, it is generally assumed that physical states have a unique history, representations of those states in a computational model will usually be compatible with more than one possible history in the model" (2013, 809).

The opacity of these systems is directly connected to their political and moral status. (McKinlay, 2020). Kitchin for example, observes that such methods are "largely directed by black box algorithms working on data of unknown provenance, and […] generally closed to recourse" (Kitchin, 2014). What Kitchin means by being closed to recourse is that there is nothing for individuals to do in relation to the technology: no transparent assessment, no possibility to challenge, and, importantly, limited ability to correct the outputs of these systems. Because they are embedded in a larger system with cascading consequences—often negative—the implications of these systems are sometimes irreversible. This is different from other complex processes, such as regular court proceedings or other bureaucratic procedures. To understand how data science technologies of the kind we discuss below differ from traditional legal proceedings consider how the transcribed testimony of a prejudiced witness could later be explicitly read out as part of an appeal process or challenged during a trial. By contrast "the workings of a [computerized] recidivism model are tucked away in algorithms, intelligible only to a tiny elite." (O'Neil, 2016 p.25). As boyd and Crawford put it "when computational skills are positioned as the most valuable, questions emerge over who is advantaged and who is disadvantaged in such a context. This, in its own way, sets up new hierarchies around 'who can read the numbers'" (2012). However, the central issue is that the lack of access to the technical details of these technologies means that it is generally difficult to determine when and how the system has made a mistake and how to respond to the relevant decision makers appropriately with counterarguments and evidence.

Burrell (2016) as well as Alvarado and Humphreys (2017) discuss examples of algorithms that are opaque in the sense that as a recipient of their output "rarely does one have any concrete sense of how or why a particular classification has been arrived at from inputs" (Burrell, 2016). Mittelstadt et al. note that the uncertainty and opacity of some of these methods "inhibit the identification and redress of ethical challenges in the design and operation of algorithms". Furthermore, they argue that software artefacts used in data analysis bring with them the kinds of harms caused by what *algorithmic activity*, which "is hard to debug (i.e. to detect the harm and find its cause)" (Mittlestadt et al., 2016 p.5; Floridi 2014). Mittelstadt et al. cite Rubel and Jones (2014) as suggesting that "the failure to render the processing logic [of some ML algorithms] comprehensible to data subjects disrespects their agency." (Mittelstadt et al., 2016 p.7). Critics are clearly converging on the view that in these contexts, it is challenging

to determine whether or when a harm is being done to us. If one is in a subordinate position with respect to decision makers who use these tools then, as we will argue below, one suffers a hermeneutical epistemic injustice; as we will also argue, when one is automatically considered a lesser knower vis-à-vis the tools themselves, one suffers a testimonial epistemic harm.

As we have seen, a hermeneutical epistemic injustice is one in which unjust social conditions prevent the subject of the harm from having access to the relevant terms, criteria, words, ideas, etc. necessary to understand or articulate that a harm has been done to them. Fricker uses the example of 'sexual harassment' to elucidate this kind of harm. In societies where a term does not exist to name this specific harm, those who suffer that harm will have greater difficulty finding a remedy.[16]

The kind of opacity described here plays out in a socially consequential manner in contexts where corporate or governmental authorities use data-driven systems for decision-making purposes. For Glenn Rodriguez, for example, opacity was key to his now famous struggle with automated data-driven decision-making processes (Wexler, 2017; 2018; Rudin and Ustin, 2018; Welz, 2019; Rudin, 2019). Rodriguez was denied his petition for parole in the New York prison system because of the results of the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) risk assessment algorithm (Wexler 2017; 2018). Systems like COMPAS are used by many court systems in the United States to aid or compliment decision-making processes such as parole hearings, sentencing, bail calculations, and other bureaucratic procedures. The treatment of prisoners like Rodriguez constitutes a clear instance of epistemic injustice. Few members of American society are as low on the social hierarchy as prison inmates and in Rodriguez's case, the injustice is striking. His case exemplifies the distinct layers of opacity that can be at play when challenging systems that use data science to provide risk assessments and other consequential decisions. It is instructive and rare insofar as Rodriguez was able to navigate serious obstacles in order to finally understand and attempt to challenge the technological system that the New York State prisons were using. Rodriguez' story shows why opacity—whether intentional or not— plays such an important part in the epistemic injustices related to data science methods and products.

[16] Likewise, the norms governing marriage in some cultures make it difficult for victims of marital rape to explicitly articulate what has happened to them in an institutional context such as a court.

While facing a parole hearing, Rodriguez's record indicated that while he had been in prison, he had been a model of rehabilitation. However, the board denied his petition for parole citing his COMPAS risk score as the basis for their decision. COMPAS is one of many risk assessment algorithms used by court systems in the United States to aid or compliment decision-making processes such as parole hearings, length of sentences, bail calculations, and others.[17] Details of the inner workings of COMPAS are proprietary trade secrets and are therefore protected to a very large degree from inspection.[18]

By consulting with fellow inmates, Rodriguez ingeniously managed to identify the relevant value in the qualitative survey part of the assessment that was likely responsible for his derogatory score.[19] He was able to compare his score and the answers to the qualitative survey part of the assessment by an evaluator with those of other inmates in order to determine that one question likely had an inappropriately weighty influence on the output of the algorithm. When compared to those with otherwise similar traits and records it seemed that the algorithm was producing an unfair result. However, knowing this was not enough. Even after providing evidence of the disproportionate weight this particular question had on the operation of the algorithm, no public access to the process by which the algorithm weighed the different inputs was granted. Rodriguez' story began as an instance of hermeneutical injustice. He was unfairly prevented from understanding what had been done to him.[20] Rodriguez did not initially even

[17] Note here that while COMPAS has been widely cited in research involving bias and fairness metrics, here we are talking about a different problematic aspect of this kind of technology: their opacity. While the bias accusations towards COMPAS have been widely deemed as problematic and taken to be a construct of the metric by which the results of the COMPAS software were measured (Corbett-Davies and Goel, 2018), the issue of whether COMPAS is or is not biased may be more complicated than simply looking at the COMPAS results or the metrics used by those who judge it as biased (to see an insightful discussion of how this may be the case see Hübner, 2021).

[18] Trade secrecy - which according to Burrell (2016) belongs to the category of social epistemic opacity - is at the root of the epistemic injustice in this case. As Wexler notes, this highlights another worrying aspect of these arrangements since "private companies increasingly purport to own the means by which the government decides what neighborhoods to police, whom to incarcerate, and for how long. And they refuse to reveal how these decisions are made—even to those whose life or liberty depends on them." (Wexler, 2017)

[19] Wexler notes that "We do know certain things about how COMPAS works. It relies in part on a standardized survey where some answers are self-reported and others are filled in by an evaluator. Those responses are fed into a computer system that produces a numerical score." The important part is that the developers and owners of the software "consider the weight of each input, and the predictive model used to calculate the risk score, to be trade secrets. That makes it hard to challenge a COMPAS result" (Wexler 2017).

[20] Since we have no access to the algorithm itself or the way its weights and inputs are entered and computed it is simply not possible for us to judge whether it was indeed this one question that was skewing results against Rodriguez. COMPAS, as a data science tool takes into consideration many other attributes and features of those subjected to its processes. The qualitative survey referred to by Rodriguez is but one of the more tangible aspects of the system.

understand that epistemic recourse to the relevant elements of the process which produced his score were also beyond his reach.

Now, it is important to recognize that everyone involved—the defendant, his lawyer, and even the members of the parole board — confront the opacity of the process by which the results were arrived at. None of them were in a position to challenge, question, or even understand the technology that supports the process in which they are participating. They are all epistemically limited by the many levels of opacity surrounding the technology. In this sense, the technology diminishes the epistemic status of the judge, the lawyer, and the members of the parole board. Arguably they are all epistemically harmed in the process. However, since they are not *subject* to the decision of the system in question, or more precisely since they are voluntarily engaging with the process, they are not subordinate in the same way that Rodriguez is. Hence, we can distinguish the (modest and voluntarily assumed) epistemic harm suffered by the bureaucrats in the prison system from the profound epistemic injustice suffered by the prisoner himself.

The opacity of these systems put Rodriguez in a very difficult legal position. As Wexler explains, "generally, a defendant who wants to see evidence in someone else's possession has to show that it is likely to be relevant to his case. When the evidence is considered "privileged," the bar rises: he often has to convince the judge that the evidence could be *necessary* to his case—something that's hard to do when, by definition, it's evidence the defense hasn't yet seen." (Wexler, 2017) Rodriguez was *subject* to the consequences of the opaque system and therefore the opacity of the system was not only an epistemic harm, but an unjust epistemic harm. His lawyer, for example, was diminished epistemically, but not unjustly so. Likewise for Rodriguez's supervisor, the person charged with providing reasons for or against his parole. When Rodriguez thought he found the question seemingly skewing his COMPAS score, he provided the evidence to his supervisor who shared his assessment. The supervisor wrote a letter to the prison system and to those in charge of the use of the COMPAS system at the prison where Rodriguez was detained urging them to reassess the score. This had no effect on either the process, the score, or on Rodriguez's overall situation. Thus, the hermeneutical injustice, which was to a certain extent overcome by Rodriguez's own efforts, was accompanied by a harm to the testimonial standing of his supervisor. The supervisor in direct contact with the inmate was deemed less credible than the algorithmic assessment.

The COMPAS recidivism software that Rodriguez faced is used to assist authorities in many jurisdictions across the United States in parole decisions. The software takes a wide range of data in order to model the likelihood of recidivism. This probability is converted to a risk score. Beyond a certain threshold, parole is denied and one continues to serve one's sentence in prison (Larson et al., 2016). A group of independent journalists attempted to analyze the workings of COMPAS and claimed that the algorithm was biased against black defendants (Larson et al., 2016). Initially they determined that the software produced more errors when producing risk scores for black defendants than when evaluating white defendants. Specifically, the distribution of false negatives across subgroups in the data set did not achieve group parity, a measure of cultural fairness that seeks to equalize accuracy metrics across groups considering features such as race (Corbett-Davies, 2018; Hitchison and Mitchell, 2019; Saxena et al., 2019). This was taken to indicate that the software was unfairly producing different results for people in socially marginalized groups. Ultimately, the method by which the software was deemed to be racially biased was itself found to be faulty.[21] Nevertheless, notice that in order to determine whether a racially biased method was being used requires extensive expertise in computational statistics and data curation. Barriers to recognizing, let alone rectifying, harms in these cases mean that the systematic deployment of COMPAS in the criminal justice system fits Fricker's characterization of hermeneutical epistemic injustice.

As mentioned above, many of the processes and techniques that underlie automated decision-support tools like COMPAS will be accessible and understood, solely by a highly specialized segment subset of the population. Not only are these systems typically closed due to their proprietary nature and difficult to decompile, it is also the case that some systems can be epistemically opaque due to their formal characteristics (Horner and Symons 2019; Alvarado, 2021a; 2021b). While this can sometimes be true of other aspects of modern science and engineering, its effects in applications of data science are particularly consequential insofar as data science directly influences our capacity to make decisions in ways that are often directly relevant to the well-being of others. A significant practical harm in these contexts is that some of these data science processes—in and of themselves— are closed to epistemic recourse while generating results whose uses have morally significant consequences (Kitchin, 2014). This is

[21] See Flores et al. (2016), Dieterich et al. (2016), Feller et al. (2016), Chouldechova (2017), and Corbett-Davies and Goel (2018) for a thorough examination of the details of the question of racial bias in COMPAS.

especially harmful in the case of individuals who have been economically and otherwise marginalized and are subject to administrative or corporate control in institutional settings like prisons, schools, and hospitals (Young, 2013). A serious moral challenge is posed by applications of data science that are likely to exacerbate the unequal status of individuals by empowering members of administrative or managerial elites while simultaneously sheltering their decisions from scrutiny.

Louise Amoore notes that technologies like risk-assessment algorithms are "political because they precisely involve combinatorial possibilities whose arrangement has effects in the world" (Amoore, 2014). That is, the effects they have in the world is what makes these technologies politically significant. There is a sense in which this is uncontroversial. However, an exclusive focus on politics, understood in terms of power relations, can lead commentators like Pratyusha Kalluri (2020) to urge researchers to forgo ethical questions and to focus instead on investigating the ways in which the technology changes power relations. One obvious problem with this approach is that a technology may bring about new moral problems that are orthogonal to the existing political or social dynamics. The tendency in recent data ethics has been to accept Latour's slogan that "technology is society made robust" meaning that technology is shaped largely by social values and existing power relations (Latour and Vent, 2002). Notice that this neglects the possibility that social values can themselves be shaped and altered by technological developments (Winner, 1980). In the case of the kinds of technologies we are considering, we have been arguing that they can harm us qua epistemic agents in a manner that other technologies typically do not (Burrell, 2016; Alvarado and Humphreys, 2017).

4. Testimonial Injustice and Data Science Technologies

Let's consider cases where the testimony and judgment of individual persons is devalued inappropriately by data science technologies in a bureaucratic or institutional setting. For example, in North American higher education, administrators commonly use expensive commercial data analytics services to determine the relative contribution of individual faculty members to the overall ranking of their universities. These tools consider factors like the number of articles published, the impact of the journals where they publish, the number of citations,

prizes, grants, and the like. Companies like Academic Analytics promise to provide administrators insight into the career progression of individual faculty relative to their peers, allowing them not only to rank but also to predict the trajectory of faculty careers.[22] Administrators are willing to use institutional resources for this purpose because they have been persuaded that using such systems offer them objective insight into factors affecting their institution's ranking and their ability to garner extramural research funding. In practice, these systems are used in decisions concerning the allocation of financial resources, hiring and retention decisions, and in some cases tenure and promotion decisions (Basken 2018; Else, 2021).

Consider a dean needing to decide whether a faculty member in the art history department who is offered a position at a competing university should be given a competitive retention offer. The software reports that the art historian's departure will have no impact on the reputation of the department or of the university and the dean uses this output to justify the decision not to expend resources on a retention offer. Now imagine that the head of the art history department insists to the dean that the art historian's departure would be a significant loss to the institution.

In disagreements of this kind, data analytics companies promise to empower administrators to make decisions based on unbiased data. In their sales pitch, this is frequently contrasted with the personal, intuitive (and thereby supposedly unreliable) judgments of department heads and other subordinates. Companies that develop these instruments provide upper-level administrators with the advantage of many more data points, aggregated from the discipline as a whole, from other institutional resources, and from places that would be unavailable to their subordinates. The instrument itself is generated according to proprietary methods and data that are not accessible to either customers or those who are subject to decisions justified by this instrument. These tools are marketed to administrators as authoritative decision support tools built with proprietary methods using data sets that would be difficult to acquire otherwise.

Is the department head in our hypothetical dispute concerning the art historian treated in an epistemically unjust manner? The dean might defend their decision by arguing that the system

[22] For a demonstration of the user interface of the system marketed by Academic Analytics see https://youtu.be/U_Li7ZEp3e0 (Last accessed Oct. 3rd, 2021).

is better able to capture and analyze a broader and more complex data set than an unaided human. If so then the department head *is* in an epistemically inferior position and should simply defer to the judgment of the system. If this is the case, the department head was not epistemically harmed, let alone harmed unjustly.

The morally problematic epistemic issue arises with the *assumption* that the system is superior to the judgment of the department head with respect to the faculty member's work. Let's add some details to the case in order to show why such judgments are frequently mistaken: imagine that the art historian in question has published four essays in edited volumes or in exhibition catalogs. In addition, she has three highly-cited articles published four years prior, along with two other more recent articles in relevant journals of her subdiscipline. Finally, she is also involved with important community-building contributions within her field. From the perspective of her department head and by the standards of her discipline this counts as a good record.

It is commonly the case that analytical tools of the kind deployed by administrators are not sensitive to different disciplinary norms with respect to judgments of quality. Most obviously, for example, unlike engineering or the natural sciences, judgments of quality in disciplines like art history are not substantially influenced by citation measures. It is also the case that contributions to edited volumes and exhibition catalogs are sometimes highly regarded by art historians but are frequently neglected by indexing services of the kind that commercial systems scrape for data. It is also the case that the time horizon for judgments of quality vary across the disciplines. Papers four years or older might not count towards the reputation of faculty in some disciplines whereas other slower disciplines concern themselves with work from decades earlier.

Given that the methods underlying the instrument are protected business secrets, there is no way to determine whether inappropriate standards or quantitative measures are being applied to her case. Faculty in disciplines like art history are right to suspect that the analytical approaches that are marketed to university administrators are governed by norms derived from high status disciplines in the STEM fields. However, it is difficult to make this case in response to the output of the system given the proprietary nature of these services. Faculty in non-STEM disciplines note that the incentive structures created by these instruments will have a distorting

effect on research and scholarship, harmfully encouraging rapid publication on high-profile topics in order to maximize short-term citation scores (Basken 2018).

In the case as described, a testimonial epistemic injustice (albeit a relatively mild one) has occurred. Even though a department chair at a university has a far higher social and economic status than a prisoner like Glenn Rodriguez, the department chair is unfairly diminished epistemically in virtue of their position in the institutional hierarchy. Let's begin with the harm. The epistemic standing of the department head, as an expert and well-informed insider in the world of art historical scholarship was undermined by the dean's use of the data analytics instrument. The assumption that the system is superior to the department head in the evaluation of the faculty member is unjustified given the scenario above. As we have explained above, being incorrectly diminished in this way is a harm, even if the chair's material conditions are unchanged. In order to see why it counts not only as a harm but as an injustice it is useful to reframe the scenario by replacing the software with people. Consider for example a consultant who is not acquainted with the academic field in question, being brought in with a checklist, a calculator and a set of complicated metrics to assess the value of an individual faculty member's work. Consider also that this consultant's judgment is taken to serve as the justification of the dean's decision. If the dean were to take the final score given by the consultant as more important than the input of the department head then the mistake, in the scenario described above, is obvious. In the human case, the department head would be entitled to object to the specific criteria that the consultant used and would be in a position to offer more appropriate alternatives. For someone in a position of institutional power to override the testimony of a subordinate who is an expert in a field in favor of a non-expert for non-epistemic reasons is an epistemic injustice. It is simply an illegitimate and unjust diminution of their standing as a knower and testimony. [23]

At this point we can see that the use of these technologies in institutional contexts of the kind described here is particularly pernicious insofar as it makes it difficult for subordinates to understand that a harm is being done to them with respect to their epistemic status. It diminishes subordinates as sources of testimony in unfair ways and it allows authorities to defer responsibility for decision making to opaque processes. In this example we have seen that the

[23] To see why *novel* technical artifacts, such as computational methods, should not be granted the same levels of trust as human experts see Symons and Alvarado (2019).

epistemic injustice occurs in an unexpected place, namely in the interaction between the chair, the software, and the dean. Unlike the Rodriguez case, the people involved in this example are relatively privileged academics in North America whose material condition and social status are minimally affected by the events in the scenario. Nevertheless, as we argued above, questions of distributive justice and social status are orthogonal to this instance of epistemic injustice. Simply because those involved may not have been socially or financially harmed does not mean they were not unjustly harmed in the kinds of ways described by Fricker, namely by having their epistemic status unjustifiably diminished. Similarly, simply because they are not members of recognized marginalized groups, does not mean that an epistemic injustice did not take place. This is particularly the case when one considers the subordinate status of many in institutional hierarchies. While someone that has a job may be less susceptible to harms than someone who cannot even get a job, we would not say that the person with a job cannot suffer harms within the hierarchies inherent to their place of work. While the people involved in this second example may not suffer consequences as severe as those faced by Rodriguez, being incorrectly deemed a less reliable source of testimony than one objectively is counts an epistemic harm. Within the context of an institutional or workplace hierarchy it can become an injustice. Hence, the example above qualifies as a testimonial epistemic injustice.

5. Understanding the Nature and Sources of Harms in Data Science Technology

Technologists will sometimes react to criticism from data ethicists by claiming that technology, by itself, is innocent.[24] Indeed, one of the most significant factors highlighted in the examples of epistemic injustice that we have considered so far has been the relative social status of the parties involved. Data science is certainly not responsible for the social power relations in which these technologies appear. Data science is built on straightforward and innocuous mathematical concepts and tools such as regression, classification, and clustering techniques that are applied to data sets using familiar computational processes (Saltz and Stanton 2017). Since mathematical

[24] The problems with such a position have been extensively addressed by historians of technology such as Lewis Mumford and Langdon Winner. Regarding computational processes in particular, similar positions regarding the neutrality of information technologies and methodologies was noted by Richard De George (2008, 5) in his discussion of "the myth of amoral computing".

techniques and computational processes themselves are morally and politically neutral, technologists are sometimes tempted to judge criticisms from the data ethics community as misdirected or confused. This response is not completely wrong and in order to respond, data ethicists must work to distinguish the background social critiques from careful assessment of the technologies themselves. Understanding the nature and sources of the harms that are associated with these technologies is crucial if we hope to identify possible remedies.

While a loaded gun by itself is innocent, if one introduces it into an ongoing bar fight, the quantity and quality of the harm that result are different from the harms and sources of harms at play prior to its introduction. By analogy, introducing data science technologies into social contexts with existing social injustices can change matters in ways that are due to the specific characteristics of the technologies. In this sense, the moral implications of technology cannot be explained solely by pointing to preexisting social conditions or political dynamics. The differences between effects of introducing a loaded gun, a knife, or a hand grenade into a bar fight result from the differences between the technologies. [25] At the same time, context certainly matters in the assessment of technologies. The same algorithm can have wildly different moral and social consequences when deployed in astronomy when compared with its use in the criminal justice system. Nevertheless, the harms that result from the introduction of a particular technology can be distinguished from the harms that stem from characteristics of the social setting in which the technology figures.

Thus, it is be important for the data ethics community to understand the morally relevant aspects of the technology itself, rather than pointing back to the background social and political injustices that frame the uses of these technologies. For example, in response to charges of *algorithmic bias*, technologists sometimes note that while there is a formal or learning theoretic sense in which all machine learning algorithms are biased, this bias is a politically or morally neutral source of error. It should not be confused with the colloquial uses of 'bias' in the context of racial or gender prejudices. Indeed, they are correct to note that it is a mistake to confuse the formal concept of bias in, for example the bias-variance tradeoff in machine learning models with the notion of bias that is involved in consideration of unjust social outcomes.[26] In the latter

[25] See Alvarado (2020) for an argument for this distinction.

[26] See Neal (2019) and Vapnik (2000) for additional details on the bias-variance tradeoff.

case, the technologist will claim, the fault usually lies with how data was gathered or with historical biases inherent in that data rather than with the algorithm. Data samples that are derogatory or prejudiced in some socially unjust ways are the responsibility of those who collected or curated the data rather than the fault of the technologist who created the algorithms. In scholarly data science and data ethics literature, 'bias' in the popular sense of unjust social prejudice is usually distinguished correctly from the technical sense of bias as it figures in statistics. However, popular presentations of data ethics issues often confuse the two. In statistics and machine learning many modeling projects face a bias-variance tradeoff as they attempt to usefully generalize beyond its training set. Bias in this context leads to underfitting a training set, variance leads to overfitting the data set. It is a fundamental feature of most learning models that they trade-off between some level of *bias* and some level of *variance*. Bias in the technical sense is when an algorithm has assumptions that lead it to fail to detect relationships between features in the data set and the target output - *the algorithm doesn't take the training set seriously enough*. While variance is the tendency of a model to *take the data set too seriously*, taking too many features as salient, and as a result not being able to generalize beyond it. Both are sources of error to be minimized. As Suresh and Guttag (2019) have noted, besides this technical notion of bias-variance tradeoff, there are many other kinds of bias that fall within and transcend the scope of algorithmic technology. While bias in the formal sense is simply a feature of these algorithms, social and other human biases that are more pernicious can influence the machine learning developing process at other stages of the production of the model.

Sometimes, even if both the data gathering method and the algorithm is free of social bias, historical inequities can influence the data. On other occasions, the bias is related to curatorial choices in the process of gathering data. Occasionally, however, the morally relevant bias is due to formal features of the algorithmic process. For example, depending on how one incentivizes success during the training phase of a model an algorithm may get very good at the aspects of the task it performs well, while never improving at the tasks where it performs badly. This can have significant consequences for systems that classify people via facial recognition or those deployed in judicial proceedings as we discussed previously. Hence, technologists cannot completely excuse themselves from blame by offloading moral responsibility to derogatory data collection practices or morally problematic uses of their technology. A blameworthy data set by itself will not always be source of the kinds of harm that can result when data is manipulated

algorithmically for decision-making purposes. At this point there are a wide range of well-documented harms related to the design, development, and deployment of data science (Amoore, 2011; 2020; O'Neil, 2016; Leonelli, 2016; Noble, 2018; Buolamwini and Gebru, 2018, Benjamin, 2019).

Some of the points discussed in this paper may be taken as capturing harms that while relevant to data science, are nevertheless not unique to its methods. In other words, one may agree that the harms addressed above, epistemic or otherwise, are harms that can arise with many other algorithmic technologies that are not data-driven *per se*. Nevertheless, this does not undermine our argument concerning the connection between epistemic injustice and opacity. In fact, we recognize that some of these concerns extend to computational methods other than the data-driven ones we discuss here. At this point, it is worth considering how many of the issues we discuss here are peculiar to data science methods and products.

In highlighting the specific ways in which these technologies change the moral and political landscape it is helpful to follow Jeroen van den Hoven's taxonomy (2000) of the kinds of moral wrong-doing associated with different kinds of technologies. Sometimes, for example, a harm can arise from a technology itself, other times the harm can come from the context in which the technology is deployed. As we have seen above, in data science, context matters. Using opaque technologies in legal proceedings may not be fair, while using it for sorting handwritten addresses is unproblematic. Still, other concerns may arise from the social dynamics that guide its development and still other harms arise from features of the technology itself. These should not be conflated. With respect to epistemic injustice we have argued that it is the features of data science technology itself that makes it peculiarly susceptible to epistemically unjust outcomes: its automated and opaque nature, for example, mean that is generally not amenable to investigative or corrective recourse.

In order to better understand the distinction between the harms *associated with the technology* and the harms *associated with the context* in which a technology is deployed, consider van den Hoven's (2000) classification of moral issues related to the internet. He distinguishes moral issues that are internet-related, internet-dependent, internet-determined and internet-specific. Here, we can generalize from van den Hoven's focus on the internet as a specific technology to technology in general. A technology-related issue is for van den Hoven an issue in which the technically specific aspects of a technology in question are incidental. There

are moral issues related to pornography on the internet, for example. But the internet is only playing a role as a communication device and, he argues, nothing about the infrastructure of the internet plays a substantial role in determining the moral issue at hand.[27] He believes that similar moral questions were at stake when VHS distribution networks arose in the early 80's.[28] We can apply van de Hoven's taxonomy to the specifics related to data-driven methods. Hence, a technology-dependent issue is one in which the technology is necessary for the harm to arise but just because the technology is in use doesn't mean that the harm necessarily follows. In other words, the technology in this case is necessary but not sufficient for the moral issue to arise. In order to hack a computer, for example, one usually needs a computer. But just because one can use a computer does not mean that it will be used for hacking. It is important to differentiate, van den Hoven notes (2000 p.134), between a technology-dependent issue and a technology-determined one. A technology-determined issue is one in which the mere introduction of a technology is sufficient to bring about a moral issue. The novel introduction of internet technology, for example, automatically generates the moral and political question of who ought to have access to it (van den Hoven, 2000, 134).

Using van den Hoven's categorization we can say that debates in the data ethics to date have typically been focused on either data science-related or data science-determined issues. However, this approach neglects the distinctive contributions of the technology itself to the ethical landscape. For example, when we are speaking about technology-determined issues, "although sufficient, the [technology] is not necessary for this type of moral question to arise, since we encounter the same moral problems of equal access and responsibility […] in card catalogs and books" (2000, 130). Similarly, technology-related issues are such that the technology involved is neither sufficient nor necessary to elicit the moral issue by itself. There is a fourth category in this taxonomy: technology specific issues, in which the technology is both necessary and sufficient for the issues to arise. It seems to be the case, particularly taking into consideration the kind of opacity discussed above that the epistemic harms and the injustices

[27] Sofya Noble (2016) highlights the ways in which Google's search algorithms exacerbate racial bias and the sexualization of young girls via its search results. Since racism and sexualization of young girls existed independently of and prior to the existence of algorithmic technology, the moral problem related to Google's algorithms is rather that of either exacerbating, enabling, or perpetuating an existing moral problem rather than creating the moral problem itself.

[28] Arguably, the example of internet pornography is not a good choice for him in the formulation of this distinction given novel characteristics of the pornography industry that have arisen in conjunction with the emergence of internet technologies.

related to them are in fact specific to data-driven, automated technologies such as the ones discussed in this paper. Returning to Rodriguez's case above for example, while some judicial proceedings may be beyond the reach of some individuals, for example, those of us without a legal education, the data-driven processes discussed in this paper seem to be opaque in a more insurmountable manner such that no amount of education can overcome them. If this is the case, then epistemic injustices like those that befell Rodriguez's are directly related to the introduction of these novel data-driven technologies themselves in a technology specific way.

In machine learning, determining which of the above-mentioned taxonomic categories apply depends on which of the many kinds of biases inherent in the machine learning production pipeline are under consideration. Suresh and Guttag (2018), for example point to six kinds of biases, each of which happens at different stages of the technology's development.[29] At some points, harms related to data gathering processes will be only data science-related and not data science-determined. By contrast, those that are related to automated clustering and classification technique may be data-science dependent issues and hence a characteristic of the methods and products of data science practice. In the study of epistemic harms, the same framework can serve to help identify the sources and kinds of blameworthy applications of data science technologies. The main lesson here is that it is a relatively straightforward matter to distinguish between the harms previously endemic to contexts in which a technology is deployed and the harms brought about because of the characteristics of a particular technology.

6. Conclusion

In this paper we have tried to illuminate the technology-specific sources of epistemic harms. We have also attempted to persuade our readers that epistemic harms are distinctive types of harm that deserve serious consideration. If we are correct, then the data ethics community ought to take account of the epistemic standing of people who are subject to data science technologies in their deliberations. We have explained how Fricker's concept of epistemic injustice can illuminate some blameworthy features of technology based on data science techniques and methods. We have also emphasized that epistemic harms can happen even when one is unharmed

[29] Similarly, as Hübner (2021) points out, some instances of algorithmic bias may have their sources in existing historical inequities while others may be the product of an analytic process.

(or even benefited) in other ways. In our view, these technologies can pose a range of morally relevant threats to our standing as knowers and decision makers.

Over the course of this paper we were careful to distinguish epistemic injustice from the structural injustices of the varying social contexts in which data science is deployed. Similarly, we were not concerned here with the practical consequences of those using the technology in misguided or malicious ways. These are important and serious issues to consider and they have been extensively documented by other researchers (Amoore, 2011; 2014; Benjamin, 2019; O'Neil, 2016; Noble, 2017). Rather, in this paper we focused on what we take to be one particularly morally significant feature of the technology itself; its opacity. Our goal has been to explain how these features of the technology account for its tendency to unjustly diminish the epistemic status of persons or groups of persons in virtue of their socially subordinate status.

The recognition of epistemic injustice as a meaningful moral category has emerged over the past two decades thanks in large part to the contributions of thinkers who were especially concerned with injustices that affect people in marginal or oppressed social groups (Fricker, 2007; Dotson, 2011; Harding, 2016; Code, 2017; Grasswick, 2018). As a general methodological commitment, we agree that ethical and political reflection is often most illuminating when one pays attention to the situations of those among us who are subject to the most serious harms and injustices. Injustice is often clearest when seen from the margins. However, as we have shown in this paper, epistemic injustices can also occur in conjunction with, or in some cases even independently of, other kinds of injustice and marginalization. Some technologies can threaten the dignity of human persons as interpreters, knowers, and sources of testimony. Knowing how to properly respond to these threats requires attention to how one's epistemic standing can be unjustly raised or degraded. This paper offers an example of how to do this in the context of data science technologies.

Acknowledgements

This paper has benefited enormously from the critical feedback of three excellent referees for this journal, we are greatly indebted to them for their careful and rigorous work. We have presented earlier versions of this paper to a wide range of venues over the past three years and

are deeply grateful to audiences for both their criticism and encouragement. Critical feedback from Markus Ahlers, Brooke Burns, Martin Cunneen, Nico Formanek, Luciano Floridi, Miranda Fricker, Stephanie Harvard, Jack Horner, Dietmar Hübner, Denisa Kera, Colin Koopman, Nicolae Morar, Kasper Luppert-Rasmussen, Camisha Russell, Laura Schelenz, Jake Searcy, Paul Showler, Irina Symons, Eran Tal, Michał Wieczorek, and Eric Winsberg has been especially helpful to us as this manuscript developed. John Symons's work is partly supported by NSA Science of Security initiative contract #H98230- 18-D-0009.

Bibliography

Alvarado, R. (2021a) "Should we replace radiologists with deep learning? Pigeons, error and trust in medical AI" *Bioethics* (Forthcoming).

Alvarado, R. (2021b) "Explaining Epistemic Opacity." (Preprint).

Alvarado, R. "Epistemic Opacity, Big Data, Artificial Intelligence and Machine Learning." In *Big Data and the Democratic Process* ed. Kevin Macnish and Jai Galliot. Edinburgh University Press. (2020)

Amoore, L. (2011). Data derivatives: On the emergence of a security risk calculus for our times. *Theory, Culture & Society*, *28*(6), 24-43.

Amoore, L. (2014). Security and the incalculable. *Security Dialogue*, *45*(5), 423-439.

Amoore, L. (2020). *Cloud ethics: Algorithms and the attributes of ourselves and others*. Duke University Press.

Anderson, E. (2012) Epistemic Justice as a Virtue of Social Institutions. *Social Epistemology* Vol. 26, No. 2, April 2012, pp. 163–173

Basken, P. (2018). UT-Austin professors join campaign against faculty-productivity company. *Chronicle of Higher Education.*

Beeby, L. (2011). A critique of hermeneutical injustice. In *Proceedings of the Aristotelian Society (Hardback)* (Vol. 111, No. 3pt3, pp. 479-486). Oxford, UK: Blackwell Publishing Ltd.

Benjamin, R. (2019). Race after technology: Abolitionist tools for the new jim code. *Social Forces*.

Besteman, C., & Gusterson, H. (Eds.). (2019). *Life by algorithms: how roboprocesses are remaking our world*. University of Chicago Press.

Boschetti, F., Fulton, E., Bradbury, R., & Symons, J. (2012). What is a model, why people don't trust them, and why they should. In *Negotiating our future: Living scenarios for Australia to 2050, Vol. 2*. Australian Academy of Science.

boyd, Danah, and Kate Crawford. "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon." Information, communication & society 15, no. 5 (2012): 662-679.

Bratu, Christine, and Hilkje Haenel. "Varieties of Hermeneutical Injustice: A Blueprint." *Moral Philosophy and Politics* (2021).

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91).

Burgess, A. (2012). 'Nudging' Healthy Lifestyles: The UK Experiments with the Behavioural Alternative to Regulation and the Market. *European Journal of Risk Regulation*, 3-16.

Burrell, J., 2016. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), p.2053951715622512.

Butterworth, M., 2018. The ICO and artificial intelligence: The role of fairness in the GDPR framework. *Computer Law & Security Review*, *34*(2), pp.257-268.

Carlson, A. M. (2017). The need for transparency in the age of predictive sentencing algorithms. *Iowa L. Rev.*, *103*, 303.

Coady, D. (2010). Two concepts of epistemic injustice. *Episteme*, *7*(2), 101-113.

Coady, D. (2017). Epistemic Injustice as Distributive Injustice 1. In *The Routledge handbook of epistemic injustice* (pp. 61-68). Routledge.

Code, Lorraine. "Epistemic responsibility." In In James Kidd, José Medina, Gaile Pohlhaus (eds.) The Routledge Handbook of Epistemic Injustice, pp. 107-117. Routledge, 2017.

Code, Lorraine. "What Can She Know: Feminist Theory and the Construction of Knowledge." (1991): 177.

Collins, P. H. (2017). “Intersectionality and epistemic injustice.” In James Kidd, José Medina, Gaile Pohlhaus (eds.) *The Routledge handbook of epistemic injustice* (pp. 115-124). Routledge.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153-163.

Chriss, James J. "Influence, nudging, and beyond." *Society* 53, no. 1 (2016): 89-96.

De George, R. T. (2008). *The ethics of information technology and business*. John Wiley & Sons.

Desrosières, A. (1998). *The politics of large numbers: A history of statistical reasoning*. Harvard University Press.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc*, *7*(7.4), 1.

Dotson, K. (2011). Tracking epistemic violence, tracking practices of silencing. *Hypatia*, *26*(2), 236-257.

Else, Holly. "Row erupts over university's use of research metrics in job-cut decisions." *Nature* (2021).

Feller, A., Pierson, E., Corbett-Davies, S., & Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *The Washington Post*, *17*.

Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, *80*, 38.

Fricker, Miranda. Epistemic injustice: Power and the ethics of knowing. Oxford University Press, 2007.

Fricker, Miranda. "Evolving concepts of epistemic injustice." (2017): 53-60.

Glick, P., & Fiske, S. T. (1997). Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women. *Psychology of Women Quarterly, 21*, 119–135. https://doi.org/10.1111/j.1471-6402.1997.tb00104.x.

Green, B. (2020). Data science as political action: grounding data science in a politics of justice. *Available at SSRN 3658431*.

Grasswick, H. (2018). Understanding epistemic trust injustices and their harms. *Royal Institute of Philosophy Supplements*, *84*, 69-91.

Hagendorff, Thilo. "The ethics of Ai ethics: An evaluation of guidelines." Minds and Machines (2020): 1-22.

Hand, David J. "Principles of data mining." Drug safety 30.7 (2007): 621-622.

Harding, Sandra. *Whose science? Whose knowledge?*. Cornell University Press, 2016.

Haslanger, S. (1995). Ontology and social construction. *Philosophical topics*, *23*(2), 95-125.

Haslanger, S. (2000). Gender and race:(What) are they?(What) do we want them to be?. *Noûs*, *34*(1), 31-55.

Horner, J. K., & Symons, J. (2019). Understanding error rates in software engineering: Conceptual, empirical, and experimental approaches. *Philosophy & Technology*, *32*(2), 363-378.

Horner, J. K., & Symons, J. F. (2020). Software engineering standards for epidemiological models. *History and Philosophy of the Life Sciences*, *42*(4), 1-24.

Hübner, D. (2021). Two kinds of discrimination in AI-based penal decision-making. ACM SIGKDD Explorations Newsletter, 23(1), 4-13.

Hutchinson, B., & Mitchell, M. (2019, January). 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 49-58).

Jo, E. S., & Gebru, T. (2020, January). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 306-316).

Johnson, J. (1988). Mixing humans and nonhumans together: The sociology of a door-closer. *Social problems*, *35*(3), 298-310.

Kalluri, P. (2020). Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, *583*(7815), 169-169.

Kelleher, John D. and Brendan Tierney. Data Science. MIT Press, 2018

Keyes, O., Hutson, J., & Durbin, M. (2019, May). A mulching proposal: Analysing and improving an algorithmic system for turning the elderly into high-nutrient slurry. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-11).

Keyes, O. (2020). Automating autism: Disability, discourse, and Artificial Intelligence. *The Journal of Sociotechnical Critique*, *1*(1), 8.
Kidd, I. J., Medina, J., & Pohlhaus, G. (2017). *Introduction to the Routledge handbook of epistemic injustice* (pp. 1-9). Routledge.

Kitchin, Rob. The data revolution: Big data, open data, data infrastructures and their consequences. Sage, 2014.

Koopman, C. (2019). *How we became our data: A genealogy of the informational person*. University of Chicago Press.

Latour, B., & Venn, C. (2002). Morality and technology. *Theory, culture & society*, *19*(5-6), 247-260.

Lembcke, T.; Engelbrecht, N.; Brendel, A.; and Kolbe, L., (2019). "To nudge or not to nudge: Ethical considerations of digital nudging based on its behavioral economic roots". In *Proceedings of the 27th European Conference on Information Systems (ECIS)*, Stockholm & Uppsala, Sweden, June 8-14, 2019. ISBN 978-1-7336325-0-8 Research Papers. (https://aisel.aisnet.org/ecis2019_rp/95)

Lin, Y., Osman, M., & Ashcroft, R. (2017). Nudge: concept, effectiveness, and ethics. *Basic and Applied Social Psychology*, *39*(6), 293-306.

Leonelli, S. (2016). Locating ethics in data science: responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160122.

Leonelli, Sabina, "Scientific Research and Big Data", *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2020/entries/science-big-data/>.

Mason, R. (2011). Two kinds of unknowing. *Hypatia*, *26*(2), 294-307.

May, D. R., Li, C., Mencl, J., & Huang, C. C. (2014). The ethics of meaningful work: Types and magnitude of job-related harm and the ethical decision-making process. Journal of business ethics, 121(4), 651-669.

McKinlay, Steve. Trust and Algorithmic Opacity. In *Big Data and the Democratic Process* ed. Kevin Macnish and Jai Galliot. Edinburgh University Press. (2020)

McSherry, David. "Explanation in recommender systems." *Artificial Intelligence Review* 24.2 (2005): 179-197.

Medina, J. (2017). Varieties of Hermeneutical Injustice 1. In *The Routledge handbook of epistemic injustice* (pp. 41-52). Routledge.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35.

Neal, B. (2019). On the bias-variance tradeoff: textbooks need an update. *arXiv preprint arXiv:1912.08286*.

Nilsson, A., Erlandsson, A., Västfjäll, D., & Tinghög, G. (2020). Who are the opponents of nudging? Insights from moral foundations theory. *Comprehensive Results in Social Psychology*, 1-34.

Nissenbaum, Helen. "6. Puzzles, Paradoxes, and Privacy in Public." *Privacy in Context*. Stanford University Press, 2020. 103-126.

Noble, Safiya Umoja. Algorithms of Oppression: How search engines reinforce racism. NYU Press, 2018.

O'Neil, Cathy, and Rachel Schutt. Doing data science: Straight talk from the frontline. " O'Reilly Media, Inc.", 2013.

O'Neil, Cathy. Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books, 2017.

Origgi, G., & Ciranna, S. (2017). Epistemic injustice: the case of digital environments. In *The Routledge Handbook of Epistemic Injustice* (pp. 303-312). Routledge.

Ranchordás, S. (2020). Nudging citizens through technology in smart cities. *International Review of Law, Computers & Technology*, *34*(3), 254-276.

Rendsvig, R. & Symons, J. (2021) Epistemic Logic. *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition) Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/sum2021/entries/logic-epistemic/.

Rudin, C. (2019, July). Do Simpler Models Exist and How Can We Find Them?. In *KDD* (pp. 1-2).

Rudin, C., & Ustun, B. (2018). Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces*, *48*(5), 449-466.

Ruiz, A.G., 2019. White knighting: How help reinforces gender differences between men and women. *Sex Roles*, *81*(9), pp.529-547.

Saltz, J. S., & Stanton, J. M. (2017). *An introduction to data science*. Sage Publications.

Slater, P. (Ed.). (1980). *Outlines of a Critique of Technology*. Inklinks, Limited.

Simon, Judith. "Epistemic Responsibility in Entangled Socio-Technical Systems'." In Proceedings of AISB/IACAP World Congress 2012. 2012.

Simondon, G. (2017). *On the mode of existence of technical objects* (p. 59). Minneapolis: Univocal Publishing.

Spivak, G. C. (2003). Can the subaltern speak?. *Die Philosophin*, *14*(27), 42-58.

Symons, J., & Alvarado, R. (2016) "Can we trust Big Data? Applying philosophy of science to software." Big Data & Society 3, no. 2 : 2053951716664747.

Symons, J., & Alvarado, R. (2019). Epistemic entitlements and the practice of computer simulation. *Minds and Machines*, *29*(1), 37-60.

Symons, J., & Boschetti, F. (2013). How computational models predict the behavior of complex systems. *Foundations of Science*, *18*(4), 809-821.

Symons, J., & Horner, J. (2014). Software intensive science. *Philosophy & Technology*, *27*(3), 461-477.

Symons, J., & Horner, J. (2019). Why there is no general solution to the problem of software verification. *Foundations of Science*, 1-17.

Sunstein, C. R. (2014). *Why nudge?: The politics of libertarian paternalism*. Yale University Press.

Sunstein, C. R. (2016). People prefer System 2 nudges (kind of). *Duke Law Journal*, *66*, 121–168.

Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press.

Van den Hoven, Jeroen. "of Moral Wrongdoing." Internet ethics (2000): 127.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Weltz, J. (2019). *Over-Policing and Fairness in Machine Learning* (Doctoral dissertation, Pomona College).

Wexler, R. (2017). When a computer program keeps you in jail: How computers are harming criminal justice. *New York Times*, *13*.

Wexler, R. (2018). The odds of justice: Code of silence: How private companies hide flaws in the software that governments use to decide who goes to prison and who gets out. *CHANCE*, *31*(3), 67-72.

Winner, Langdon. Autonomous technology: Technics-out-of-control as a theme in political thought. Mit Press, 1978.

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

Yapo, Adrienne, and Joseph Weiss. "Ethical implications of bias in machine learning." (2018). URI: http://hdl.handle.net/10125/50557

Yong, E., 2012. Nobel laureate challenges psychologists to clean up their act. *Nature News*.