Liberalism and Automated Injustice

forthcoming in *Elgar Research Handbook of Liberalism* (ed. Duncan Ivison)

Chad Lee-Stronach

1 Introduction

It is well understood that our lives are substantially steered by the decisions of social institutions, "beginning with admission or nonadmission to nursery school and ending with admission or nonadmission to the nursing home"(Elster, 1992, p. 2). The fact that we are subjected to such institutions grounds our claims that these institutions be, in various ways, *just* in their decision-making. In a liberal democratic society, this broadly requires that they are consistent with a constellation of values, including individual freedom from unjustified coercion, social and distributive equality, and morally autonomous choice in personal and political matters (Dworkin, 2000; Kolodny, 2023; Rawls, 2001). So far, so familiar.

Less well understood is the recent but increasingly ubiquitous fact that our social institutions now rely heavily on data-driven, machine-executed algorithmic decision processes (in short, computing technology) to make decisions that steer our lives. Such is the extent of this change that is dramatic though not inaccurate to say that “we are all now subjects of the empire of algorithms” (Daston, 2022, p.7). In light of this fact, we must ask: under what conditions, if any, are automated social institutions consistent with a liberal-democratic conception of justice?

There is, as yet, no clear answer to this question. In an effort to approach such an answer, in this chapter I set out a conceptual schema for locating and explaining putative injustices of automated decision-making in social institutions. The strategy being that to understand what justice requires, it is helpful to first learn to recognise and explain instances of injustice (Sen, 2009).

I will begin with a broad, first-pass categorisation of injustices involving automated social institutions (**Part 2**). I then present a conceptual schema that offers a more fine-grained and systematic explanation of such injustices (**Part 3**). This heuristic framework reveals various distinct sites of injustice in information processing systems. I suggest that these sites may be unjust by virtue of their intrinsic (**Part 4**) or extrinsic features (**Part 5**). This schema does not give a full explanation of the phenomena at hand, but it does, I think, provide a useful framework for understanding what justice requires from an increasingly automated social world.

2 Beyond "Automated Injustice"

Liberal political philosophy provides a rich and systematic conceptual schema for identifying and explaining various types of injustice. A social institution can be unjust when it (among other things) undermines individuals' rights, treats them with adequate respect, prevents genuine consent to power being exercised over them, disables their moral and political freedoms, undermines appropriate social relations of equality, fails to provide justification for the exercise of power or obscures the mechanism and agents of power that affect people's vital interests.[1] I will take this broad liberal conception of what constitutes injustice for granted in what follows.

Our question: how, if at all, does a social institution's use of computing technologies to make decisions realise any of these types of injustice? I will suggest that to answer this question, we need to more precisely define the nature and role of the computing technologies within the institution. The schema I will introduce in **Part 3** aims to do just this.

Suppose, however, that we choose to instead give only a high-level, placeholder description for the details of injustices involving computing technologies. Suppose we call it *automated injustice*, focusing on the fact that the injustice is grounded in the social institution's decision procedures being at least partly executed by computing machinery, rather than moral persons. While automation is a salient feature of present concerns about the role of computing technology, and specifically burgeoning artificial intelligence, in social institutions, we will see that it conflates or otherwise obscures distinctive types of injustices that require significantly different normative evaluations and responses.

To illustrate, suppose that we attempt to analyse injustices involving these technologies using only the placeholder concept. The following, one might argue, are all instances of automated injustice:

Law Enforcement: Law enforcement departments' decisions as to how to geographically allocate police personnel are in part guided by a company whose software makes statistical predictions based on police data, such as arrests. Judge's decisions as to whether and under what conditions to release a defendant on bail are influenced by the output of proprietary software that aims to predict 'recidivism risk' of defendants.

Voting: Individual decisions as to whether and how to vote are influenced by exposure to highly personalized political advertisements on social media. This degree of personalization is made possible by the social media company sharing personal data to a private company working on behalf of a particular political candidate.

Personal Use: Individuals searching for information or entertainment use an online search platform whose output of 'relevant' results may be skewed by advertising interests and sometimes presented via manipulative or deceptive user-interface.

Education: An education provider's decision about whether to admit a student is determined by software whose basis of determination is inaccessible to the operator.

Social Engagement: Users of social media are primarily exposed to content that conforms to their existing viewpoints and undermines opposing viewpoints, due to recommendation algorithms that aim to optimise 'user engagement'.

Welfare: A government welfare department fully automates the determination and issuance of debt notices for calculated welfare overpayment. The recipients of notices are not provided details of the calculation or means for challenging it.

Each of these examples involves computing technology, however they are importantly different in their technical, social, political, and moral properties. For example, **Social Engagement** involves machine learning algorithms (very broadly, dynamically updating optimal pattern prediction techniques) (Hardt and Recht, 2021), based on individualised data of users of online social networks such as psychoanalytic profiles, behaviour tracking (e.g. by scroll tracking), and targeted messaging interventions deployed via user interfaces on their personal devices; **Welfare** involves a decision process governed by the outputs of simplistic arithmetic formula.[2] Although both examples involve automated decision-making of some kind, it is not at all obvious that they share any morally relevant properties, nor that the appropriate interventions, if any, should be similar for each. Insofar as there is injustice of some kind at play in these examples, we can only vaguely say that is it located *somewhere* in relation to the system. Furthermore, and as a result, we cannot say exactly it is about these examples that constitutes or causes the injustice? Our placeholder's appeal to "automation" obscures more than to illuminates. Next, I will aim to use the schema excavate a more precise analysis of these injustices.

Specifically, I will next suggest that instances of "automated injustice" are better understood as distinct types of injustices that can be located across various levels of description of information processing systems. For brevity, we can call these injustices instances of *informational injustice* (Mathiesen, 2015). I claim that by casting the explanatory resources of liberal political philosophy through the lens of this schema provides a more useful heuristic for locating and explaining putative instances of "automated justice".

3 A Schema

The above examples of automated injustice all involve machine-executed information processing systems to guide, to varying degrees, decision-making within social institutions. In these contexts, the systems take socially-relevant information as input and systematically yield a socially-relevant decision as output, subject to the system's background goals and constraints. What may be called "automated injustice" can more precisely be explained by examination of the components of the information processing systems embedded in the social institution (intrinsic informational injustice) as well as the relation between this system and the social contexts surrounding its design, development, and deployment within a social institution (extrinsic information injustice). This proposal is clearly an oversimplification, but I think it can be a useful heuristic framework for understanding the moral dimensions of this fast-evolving and urgent social problem.[3]

More specifically, the proposal is that the intrinsic features of an automated information processing system can be identified and explained at various levels, following a conceptual schema developed by David Marr in his examination of how human vision represents an environment and reliably infers properties of the environment from that representation (Marr, 1982, p. 68). According to this schema, an information processing system must be explained in terms of twin processes of representation and computation that proceed across three levels, descending from abstract specification to physical realisation.[4]

At the first, highest level of abstraction, there is the specification of the content, or subject matter, of the information and the computational problem to be solved, which may involve explicitly specifying the objectives and constraints that apply to the system. This concerns *what* information is to be processed and *why*. At the intermediate, second level, there is on the representation side the selection of data and data types, and on the computation side the specific algorithms used to process that data to arrive at an output that is consistent with the problem specification. At the third level, there is have the physical realisation of the representation and computational processes, via (in the case of computing) implementation through the code, software, hardware, and broader sociotechnical infrastructure.

I shall contend that injustices that are located within some component of the schema can be described as an *intrinsic* informational injustice.[5] In such cases, we can identify injustice within the information processing system itself, in its role within the social institution more generally. There will also be cases where the injustice only exists in its (often complex) relation to a social or political context. These latter cases may be *extrinsic* informational injustice.

Drawing the bright line between these categories is clearly difficult, and indeed one may doubt the coherence of the intrinsic category, given that any such system will inevitably be embedded in, interact with, and be influenced by, a particular sociotechnical environment. The usefulness of the intrinsic injustice category will be in its ability to carry reliable (if nevertheless defeasible) inferences about the moral justifiability of particular information processing systems *across* social contexts.[6]

4 Intrinsic Informational Injustice

We will now see that examining cases of automated injustice using the above schema reveals various sub-categories of injustice. Some of these categories will be reminiscent of those from other domains; others do not seem, on the face of it, to be reducible to one another or some other category. There may also be cases where the categories bleed into each other, or dependent on interpretation, not forming a clear partition. In any case, these categories serve as a useful tool for determining whether and where injustice exists within a social institution's automated decision process. Although the categorisation need not be committed to any particular substantive account of justice, for concreteness I will explain how the cases in questions are unjust in these various ways to according to liberal-democratic values.

4.1 Content Injustice

The idea of *content injustice* refers to the justice of an institution making a decision concerning a particular subject matter.[7] I will assume that there are, from the point of view of liberal-democratic values, subject matter constraints on social institutions: there are matters about which a specific type of social institution should and should not base its decisions.[8]

Consider a case where an employer uses an automated system to determine the sexual orientation of applicants. Even if this prediction did not ultimately affect the hiring decision, it seems that it should not be part of the social institution's decision problem to begin with. There is no legitimate purpose. Generalising, we may notice that content justice may be circumscribed first by individual's rights and second by the impersonal justification of that social institution. In the first case, rights-based content injustice –

where an information processing system makes judgements about an individual – may simply be instances of discrimination, privacy violation, or gaining undue advantage or power over others. In the second case, even if there is no rights violation, the content injustice may be due to the social institution using an information processing system on subjects that are outside the justification of its existence or maintenance.

A different kind of content injustice arises when legitimate content is not included in the decision-making system. In such cases, there is an undue narrowness or failure to respond to the reasons that exist concerning the appropriate determination for some individual. Suppose a public health agency uses software to guide its allocative decisions reduce the incidence of a communicative disease, but which fails to account for the affects of the those subject to its decisions, such as students affected by school closures. The injustice in the decision, if it exists, is explained by a failure to be sufficiently responsive to the moral considerations that bear on the decision.

This is, admittedly, merely a sketch of an account. Although the schema appears to reveal a distinct category of justice, it faces objections. In particular, it may be objected that outside the beliefs and intentions of the individuals who created and deployed the system, there is no content to the system. There is no such thing as *intrinsic* content injustice. In response, granting that the meaning of these technologies is socially constructed, it nevertheless appears that we can identify the content of the system by considering its components and functions within the social context: such content is, after all, demanded by the patent law concerning such technologies.[9] Although it may be difficult to precisely determine the content of an information processing system within a social institution, it does not follow that there is no fact of the matter.

4.2 Computational Injustice

For a given subject matter for decision-making, there may be various ways of specifying the goals and constraints of the decision itself. In the present context, the question is: what is the nature of the problem to be solved by the social institution? The answer to this question may be more revealing of injustice than any other component: justice may be "likely to be understood more readily by understanding the nature of the problem being solved than by examining the mechanism (and the hardware) in which it is embodied." (Marr, 1982, p. 27) We cannot simply examine the downstream components, such as the algorithm, data, presentation and artifact, because they will generally be consistent with multiple problem specifications.

Indeed, it seems that many examples of automated injustice are more precisely instances of computational injustice. Recall the case of **Welfare**: suppose that the problem specification is to make determinations about whether to claim repayment from individuals according to a rule that maximises the sum of money from individuals deemed to have been 'overpaid' by the public welfare system.[10] This problem specification merely requires the algorithm that minimises false negative errors (where no demand for repayment is made, but individual was overpaid). This specification is unjust because it gives no weight a morally relevant type of error: false positives (where the claim for repayment is made, but the individual was not overpaid). Even supposing that a government department can legitimately claim repayment for when it alone perceives it is has paid individuals more than they are eligible, determinations on whether to make that claim for any particular individual must give appropriate weight to that individual's interests in avoiding being subjected to a mistaken claim.

Also at this level of computational injustice, we can locate the debate over the justifiability of imposing 'predictive fairness' constraints on statistical decision-making software, with the aim of ensuring equal distributions of predictive error between particular population sub-groups, such as race. A challenge for such constraints is the fact that tend to permit algorithms that seem to ensure a type of parity across groups at the cost of disadvantaging particular individuals within those groups and being suboptimal from the point of view of social welfare (Corbett-Davies and Goel, 2018). At issue is not any particular algorithm, data type, or implementation; rather, the question is whether the decision problem appropriate specifies the fairness and welfare considerations that the system is aiming to realise.

Deeper questions exist in this area. In particular, we may wonder whether social institutions can legitimately treat their determinations about individuals as predictive optimization problems (Reich et al., 2021). One may argue, particular in contexts of law and findings of liability, whether the appropriate framing of the decision problem is one of inference to the *best* explanation, which may encompass epistemic and social values that lay outside what can be achieved by a purely predictive framing of the task.

4.3 Representational Injustice

Given the specified content and computational problem, we turn to the formal representation of this information within the system. We want to know what information the system explicitly or implicitly represents and how it makes these representations.[11] In the context of a social institution, we want to understand the justifiability of particular data or data types being used to make decisions about individuals.

A commonly discussed example of this injustice concerns underinclusive data sets. Machine learning algorithms, which generate and make decisions on the basis of a statistical model based on its training data, rely on a robust corpus of data to perform effectively outside its training context. Underinclusiveness in the data in this way causes performance problems in the algorithms. Facial recognition software whose algorithms are trained on data sets containing predominantly images of white male faces perform poorly for individuals who do not present those features, particularly intersectional identities such as being both black and female Buolamwini and Gebru (2018).

On the question of social categories, a further concern is that they may be misrepresented when explicitly represented as mathematical objects, such as random variables in a statistical decision problem (Hu, 2023). If, for example, race is understood to be social constructed based on social positions that are caused by or correlated with particular benefits and burdens, then particular inferences may be discriminatory if they robustly track the statistical correlation, even if they do not invoke race as such Dwork et al. (2012). The choice of data representation implies metaphysical commitments concerning social categories, which in turn have normative and descriptive implications of the information processing system used by the social institution.

In the context of social decisions, the algorithmic structures rely on a reduction of the social decision to a quantitative representation that is amenable to optimal pattern classification techniques. There may be, in practice, upwards interaction between the content and the data selection. It may be that because we, the creators of the information processing system, expect to have only particular data available, that we constrain the subject matter or computational problem accordingly. This is consistent with the schema presented here.

4.4 Algorithmic Injustice

The content, computation, and representation constrain the choice of algorithm that will be used to solve the information processing problem. In its broadest understanding, an algorithm is a set of instructions for solving a problem (Daston, 2022, pp. 126-7). More precisely for the current context of discussion is the narrower, technical definition of an algorithm as a finite set of rules which, for a given set of inputs, give a sequence of unambiguously specified operations that produce an output that solves all instances of a specific class of problem (Knuth, 2022). An algorithm thus provides a mechanical means of solving a particular kind of problem, one whose correctness does not on the background knowledge or abilities of the agent that executes it. Thus, it may be executed by an agent – such as the original calculating punch cards – or by a machine running a program that implements the algorithm (Jones, 2016). The contemporary concern about 'algorithmic justice' can be understood as a concern about individuals being subject to precisely defined, usually quantitatively rendered, non-discretionary decision procedures.

Examples of algorithmic injustice often involve the selection of an algorithm that systematically performs worse for particular individuals.[12] The interaction between data and the algorithm may generate feedback loops that exacerbate disadvantage (Ensign et al., 2017). Yet there may be other examples where injustice exists in the procedure itself, as opposed to the outcome: there are contexts where decision procedures that should allow for those subject to the decision to provide input, rather than the procedure making a unilateral determination. Moreover, there may be objections as such to the selection of algorithms that are inherently opaque in their operation and basis for their determinations (Creel, 2020).

4.5 Artifactual Injustice

We have so far explored injustices that originate in the abstract specification and representation inherent in an institution's decision-making process. Now we can turn to the implementation of this process. Artifactual injustice can be said to occur when the physical realisation – by way of code, software, and hardware – that executes the decision process is morally objectionable to the individual who is subject to it.

A simple example of artifactual injustice is otherwise well-meaning decision procedures that, when implemented, fail to operate satisfactorily for particular individuals due to errors in the software and hardware engineering. Consider technologies – such security cameras, videogame consoles, and handsoap dispensers – which rely on near-infrared sensors that must be appropriately calibrated to effectively detect darker skin tones, as opposed to just lighter ones (Liao and Huebner, 2020, n. 1). A deeper sense of artifactual injustice can be unearthed by exploring how technical specifications can constitute and reproduce forms of social order that may be inherently unjust (Benjamin, 2019; Winner, 1980).

4.6 Presentational Injustice

The presentation of the algorithm and its data refers to what and how information is conveyed about the decision process. This is a matter of justice insofar as the decisions should be explained, justified, or open to deliberation and response by those subject to them (Waldron, 1987). We ask at this implementation level: what information is made available about the specifications of the content, computational problem, data, algorithm, code, software and hardware?

On this point, we may identify many of instances of real-world automated injustices. Education providers not understanding the basis of the software's assessment of an applicant (O'Neil, 2016; Vredenburgh, 2021). Law enforcement and judicial officials not having access to the basis or justification of determinations that affect their decisions (Rudin et al., 2020). The welfare system that does not explain the basis of its automated assessment to the recipient of the notice. And so on. These presentational requirements are plausibly not merely instrumentally valuable but constitutive of relations of justified coercion by authorities and relations of social equality (Lazar, 2023).

5 Extrinsic Informational Injustice

There are various challenges one might make to the schema presented so far. Chief among them is the objection – analogous to the embodied cognition critique to computational models – that what is broadly called 'automated injustice' is not best understood in terms of discrete levels of analysis. This objection asserts that to understand these systems and their importance, we must understand the sociotechnical context in which these systems are designed, validated, deployed and monitored in practice. On this view, automated injustice is never intrinsic: it is always constituted by interacting and contested features of social schemas and practices; it is best understood as sociotechnically embodied. There is only one level of analysis: the pipeline of design from initial problem specification and data gathering to its deployment in practice, within a particular social context (Fazelpour and Danks, 2021).

There is certainly truth to this objection. There are injustices that are not inherent in the information processing system but rather only exist due to the relation of that system to a social context. To incorporate this, we may subsume the scheme in social contexts that inform and are informed by each element of the system. In this enriched schema, each component of the information processing system is nested within an environment containing natural and non-natural persons, who interact to affect or be affected by that component of the system. In this way, we may capture what can be called *extrinsic* information injustice, whereby a social institution is unjust by virtue of how it information processing system causes or constitutes unjust social relations in a particular social context. This is undoubtedly an important dimension of justice. Taking it into account may allow us to better recognise, broadly speaking, the relevance of non-ideal conditions to conceptions of justice (Anderson, 2013; Elster, 1992; Walzer, 1983).

To incorporate extrinsic informational injustice, it will help to articulate how information systems can contribute to particular social structures or systems in ways that that may be unjust (Haslanger, 2022). A structural injustice concerns the organisation and representation of values, actions, beliefs of agents within a social practice, such as an institution. A systemic injustice "occurs when an unjust structure is maintained in a complex system that its self-reinforcing, adaptive, and creates subjects whose identity is shaped to conform to it" (Haslanger, 2022, p. 22). Given this distinction, we can see that the task of explaining systemic injustice is considerably more difficult than that of structural, which itself is more difficult than analyses of intrinsic injustice. In the systemic case, there may be no fact of the matter where exactly the injustice resides; it, rather, emerges from complex interactions of agents and structures within a sociotechnical system (Fazelpour et al., 2022). This diagnostic difficulty aside, there does seem to be a meaningful notion of extrinsic injustice that is relevant to explaining cases of automated injustice.

Within this enriched schema, we can see that extrinsic informational injustices will track claims that are familiar to liberal-democratic political theory, more generally: broadly, who has the power? Who has the authority? How are social relations created or modified by these technologies? Are the benefits and burdens of these systems distributed fairly?

Using the schema for intrinsic informational injustice, we can ask more pointed questions. We may ask:

Content Injustice: Who were the deliberators determining the subject matter for decision? Were they duly empowered and accountable to those subject to the decision?

Computational Injustice: Who defines the objective? Who defines the constraints? How does this problem specification compare and interact with other institutions?

Representational Injustice: How and by whom is the data defined, collected, cleaned and managed?

Algorithmic Injustice: Who owns, operates and is subject to the algorithm? Can the algorithm be effectively audited?

Presentational Injustice: What do people need to understand to be able to engage with the system? What are the rational limitations or psychological tendencies of these systems?

Artifactual Injustice: Who builds, controls, and benefits from these systems? Can those subject to these systems hold them to account? Can they avoid them without bearing undue cost?

These extrinsic moral features of information processing systems are crucial, though perhaps more familiar than the intrinsic features discussed earlier. In any case, it seems that the combination of intrinsic and extrinsic features, cast along the levels of analysis, sharpens our understanding of automated injustice.

6 Conclusion

The widespread use of automated decision-making in social institutions presents dangers to liberal democratic societies. The philosophical challenge is to understand the nature of this danger. In doing so, it seems that we may gain new insight into the nature and variety of social institutional injustice. The schema I have presented offers a way of distinguish broad categories of injustice (intrinsic and extrinsic) and the processes and levels of analysis within each (content, computation, representation, algorithm, presentation, artifact). This scheme, it is hoped, may serve as a rough-and-ready map for locating injustice an increasingly automated social world.

Notes

[1] Thanks here to Duncan Ivison for tying these together.

[2] Taking the difference between monthly average of the subject's taxable income and comparing that to a threshold of monthly income above which one is ineligible for welfare

[3] Here I take comfort from Goodman: "[C]onscious and cautious over-simplification, far from being an intellectual sin, is a pre-requisite for investigation. We can hardly study all at once all the ways in which everything is related to everything else." (Goodman, 1983, p. xx).

[4] The higher levels of the schema may be understood, as a first approximation, as multiply realisable by the lower levels, in what may seem like a supervenience relation. However, given that these lower levels may instantiate multiple higher level processes, and also that designers of these systems may restrict the aims and means of the system to that which is convenient or other feasible, this first approximation is misleading because

any particular lower level may multiply realise various higher level processes. This many-to-many mapping between levels is perhaps characteristic of the dual-use problems of technology, more generally.

[5] For a alternative approach to applying this schema to automated injustice, see: (Kasirzadeh and Klein, 2021).

[6] (Jackson, 2010, p. 136)

[7] On the (often overlooked) role of content within Marr's schema, see: (Ritchie, 2019). On different accounts of subject matter (the exact choice of which does not bear on the discussion), see e.g.: (Lewis, 1988; Yablo, 2014).

[8] Strictly speaking, there are four possibilities to explore. These include when the content is: 1) wholly outside the legitimate domain of decision-making; 2) wholly inside the legitimate domain, but does not fully cover it; 3) wholly covering the legitimate domain and some illegitimate illegitimate; 4) partially legitimate and partially illegitimate. To simplify, I will simply look at the inclusion of illegitimate content and the non-inclusion of legitimate content.

[9] For example, whether Twitter technical architecture is of "a device independent message distribution platform" does not depend on the attitudes of those who own or use it.

[10] For discussion of a real case of this kind, coined 'Robodebt' see: (Broad, 2018).

[11] Here I go beyond Marr's focus on explicit representation: "a formal system for making explicit certain entities or types of information, together with a specification of how the system does this." p. 20

[12] In the context of medical diagnostics and "individualised medicine", see: (Vyas et al., 2020).

References

Anderson, E. (2013). *The Imperative of Integration.* Princeton University Press.

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code.* Polity Press.

Broad, E. (2018). *Made by Humans: the AI Condition.* Melbourne University Press.

Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, 81:1–15.

Corbett-Davies, S. and Goel, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science*, 87(4):568–589.

Daston, L. (2022). *Rules: A Short History of What We Live By.* Princeton University Press, Princeton & Oxford.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. *ITCS 2012 - Innovations in Theoretical Computer Science Conference*, pages 214–226.

Dworkin, R. (2000). *Sovereign Virtue: the Theory and Practice of Equality.* Harvard University Press, Cambridge, Massachusetts.

Elster, J. (1992). *Local Justice: How Institutions Allocate Scarce Goods and Necessary Burdens.* Russell Sage Foundation, New York.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. (2017). Runaway Feedback Loops in Predictive Policing. pages 1–12.

Fazelpour, S. and Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, (May):1–16.

Fazelpour, S., Lipton, Z. C., and Danks, D. (2022). Algorithmic Fairness and the Situated Dynamics of Justice. *Canadian Journal of Philosophy*, 52(1):44–60.

Goodman, N. (1983). *Fact, Fiction, and Forecast.* Cambridge, Massachusetts, 4th edition.

Hardt, M. and Recht, B. (2021). *Patterns, predictions, and actions: A story about machine learning.*

Haslanger, S. (2022). Systemic and Structural Injustice: Is There a Difference? *Philosophy*, pages 1–27.

Hu, L. (2023). What is “Race” in Algorithmic Discrimination on the Basis of Race. *Journal of Moral Philosophy*, pages 1–23.

Jackson, F. (2010). *Language, Names, and Information.*

Jones, M. L. (2016). *Reckoning with Matter.* University of Chicago Press, Chicago and London.

Kasirzadeh, A. and Klein, C. (2021). *The Ethical Gravity Thesis: Marrian Levels and the Persistence of Bias in Automated Decision-making Systems*, volume 1. Association for Computing Machinery.

Knuth, D. E. (2022). *The Art of Computer Programming.* Number 1 in processing. Addison-Wesley Publishing Company.

Kolodny, N. (2023). *The Pecking Order: Social Hierarchy as a Philosophical Problem.* Harvard University Press, Cambridge, Massachusetts.

Lazar, S. (2023). Governing the Algorithmic City. In *Tanner Lecture: AI and Human Values*, pages 1–47.

Lewis, D. (1988). Statements partly about observation. *Philosophical Papers*, 17(1):1–31.

Liao, S.-y. and Huebner, B. (2020). Oppressive Things. *Philosophy and Phenomenological Research*, 103(1):92–113.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information.* MIT Press.

Mathiesen, K. (2015). Informational justice: A conceptual framework for social justice in library and information services. *Library Trends*, 64(2):198–225.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishing Group, New York, NY, USA.

Rawls, J. (2001). *Justice as Fairness: A Restatement.* Harvard University Press, Cambridge, Massachusetts and London, England.

Reich, R., Weinstein, J., and Sahami, M. (2021). *System Error: Where Big Tech Went Wrong and How We Can Reboot.* Harper Collins, New York.

Ritchie, J. B. (2019). The content of Marr's information-processing framework. *Philosophical Psychology*, 32(7):1078–1099.

Rudin, C., Wang, C., and Coker, B. (2020). The Age of Secrecy and Unfairness in Recidivism Prediction. *Harvard Data Science Review*, 2(1):1–53.

Sen, A. (2009). *The Idea of Justice.* Harvard University Press, Cambridge, Massachusetts.

Vredenburgh, K. (2021). The Right to Explanation. *Journal of Political Philosophy*, 0(0):1–21.

Vyas, D. A., Eisenstein, L. G., and Jones, D. S. (2020). Hidden in Plain

Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *New England Journal of Medicine*, 383(9):874–882.

Waldron, J. (1987). Theoretical Foundations of Liberalism. *The Philosophical Quarterly*, 37(147):127–150.

Walzer, M. (1983). *Spheres of Justice: A Defense of Pluralism and Equality.* Basic Books, New York.

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, 109(1):121–136.

Yablo, S. (2014). *Aboutness.* Princeton University Press.