**ARTICLE**

# Freedom at Work: Understanding, Alienation, and the AI-Driven Workplace

Kate Vredenburgh 

Department of Philosophy, Logic, and Scientific Method, The London School of Economics, London, United Kingdom
Email: K.Vredenburgh@lse.ac.uk

**Abstract**
This paper explores a neglected normative dimension of algorithmic opacity in the workplace and the labor market. It argues that explanations of algorithms and algorithmic decisions are of noninstrumental value. That is because explanations of the structure and function of parts of the social world form the basis for reflective clarification of our practical orientation toward the institutions that play a central role in our life. Using this account of the noninstrumental value of explanations, the paper diagnoses distinctive normative defects in the workplace and economic institutions which a reliance on AI can encourage, and which lead to alienation.

## 1. Introduction

Imagine that you live in an unjust society that counts trustworthy, omniscient, and unusually straightforward oracles among its members. You ask an oracle whether your society will ever be just and are told that you live in fortunate times: within the next five years, sweeping institutional changes will transform your society into a just one. Five years pass, and there are indeed sweeping institutional changes. Since your oracle is trustworthy, you know that your society is now just. But you do not know why it is just, for it turns out that justice has been achieved by developing and implementing a centralized algorithmic decision-making system to allocate benefits and burdens, administer the laws, and so on. The system is too complex for you—or anyone else, oracles aside—to understand its underlying structure or the decisions it makes.

The burgeoning literature in computer science, law, and philosophy on the explainability of artificial intelligence has focused on various ways that the ability to explain artificial intelligence is of instrumental value. Explainable AI, for example, is argued to be valuable for recourse, or for contesting decisions, both of which are necessary for institutions to be legitimate and fair (Ventkatasubramanian and Alfano 2020; Vredenburgh 2021). From the perspective of that literature, you would have been wrong to trust that the oracle is omniscient. If your society is indeed objectionable due to its opacity, it must be unjust.

However, I'd like you to suppose for a moment that the imagined society is indeed perfectly just. Do you still have the intuition that the above society is objectionable in some way?

Many, I think, do. I will explain and defend this intuition in terms of the noninstrumental value of explanation. Explanations of the structure and functioning of organizations and social institutions are of noninstrumental value because they form the basis for reflective clarification of the all-things-considered practical orientation we each cannot help but take toward our own

social world.[1] A practical orientation may range from affirmation or identification with the social world to rejection or opposition; from tacit or inchoate to reflectively articulated; it may take as its object not just the social world as a whole, but also particular institutions within it; and, crucially, it is the sort of attitude for which there is a right kind of reason. The normative character of the social world is what makes ways of relating to it appropriate.

Since taking the proper practical orientation toward the social world requires understanding its normative character, it requires normative explanation. Possession of these explanations is not an instrument by which we orient ourselves to the social world, but part of what such orientation consists in.

This paper focuses on the value of normative explanation in the workplace, a domain in which the transparency of institutional structure and functioning has not received as much attention as, say, the political sphere. But I take it to be worth focusing on not only on philosophical grounds, but also because of the ways in which recent technological developments are, dramatically and sometimes uniquely, making work unexplainable.

Section 2 introduces the major conceptual machinery used in the paper, that of a practical orientation, and explains its relationship to social freedom and alienation. Section 3 argues that explanations of the structure and function of workplaces and economies that rely on AI for decision-making are of noninstrumental value. Sections 4 through 6 argue that economic institutions and workplaces that use AI for decision-making are particularly vulnerable to undermining social freedom by limiting normative explanations. Section 4 examines one mechanism that limits normative explanation, technical opacity. I argue that technical opacity does not pose the largest threat to the availability of normative explanations in the workplace and economic institutions. Instead, mechanisms of worker isolation (section 5) and control (section 6), which have been expanded and transformed by AI, pose a greater threat.

## 2. Practical orientation

Why think that a society that is just is normatively lacking? A way into this thought is through Rawls's concept of a well-ordered society, particularly the requirement of publicity (Rawls 2000). Exercises of coercive power violate individuals' autonomy unless they can endorse a government as legitimate. However, given facts about reasonable pluralism, the government must be justified in public terms, and society must be regulated over time by public conception of justice, if such a government is to respect individual autonomy (Schouten 2019).

Rawls' requirement of publicity is rooted in a tradition of political philosophy that goes back at least to Hegel. In this paper, I will be concerned with Hegel's concept of social freedom—or being reconciled with one's social world. To be reconciled with one's social world is, for Hegel, to be at home in it—to be no longer alienated from society or oneself, but instead to see one's social world as worthy of endorsement (Hardimon 1994). Social freedom has both a subjective and an objective component. The objective component of social freedom is whether one's institutions secure the conditions of freedom for all. For Hegel, these conditions center around self-determination and self-realization. But it is Hegel's account of subjective social freedom, not his substantive account of objective social freedom, that is central to the arguments of this paper.

I take on board Hegel's thought that one must also experience one's actions as free in order to be fully free. Experiencing one's actions as free has two aspects. First, one must experience them as self-determined. For action to be self-determined within the coercive and constraining institutions of

---

[1]Following Korsgaard (1983), I distinguish between final, or noninstrumental, and instrumental goods. Normative explanations of one's social world and the understanding they engender are valuable for their own sake, or finally valuable. But they are not valuable "in and of themselves"; they get their value from the value of relating to one's social world in a certain way, given that one lives in society with a sufficiently complex economy. I will use *noninstrumental* because the term may be more familiar to readers.

modern society, and experienced as such, one must be able to appropriately identify with and affirm the roles one is required to play by those institutions. Second, one must experience the social world as conducive to my practical agency, or self-realization.[2]

Thus, it is a mark of freedom to be able to identify with and affirm the institutions that shape the contours of your life. However, you may not be so fortunate as to live under institutions that merit identification and affirmation. In such conditions, it is itself an important kind of freedom simply to have an accurate *practical orientation* to your social world, whatever its valence—that is, an orientation that is fitting to the actual normative character of the social world. The intuition here resembles the thought that if someone is pretending to be your friend for personal gain but, in truth, does not care about you, it is better to know this and adjust your attitude toward the relationship accordingly, than to be in a pleasurable state of deception.

A practical orientation is a reflective attitude whose object is the major determinants of the structure and normative character of one's social world, such as institutions, norms, and organizations.[3] It is an attitude for which there can be the right kind of reason, namely, whether one's social world has the normative character that one takes it to have. An educational system designed to promote equality of opportunity, for example, licenses attitudes such as endorsement from its teachers, whereas one whose function is to uphold unjust class structures licenses attitudes such as rejection and opposition. More generally, in a society that secures the conditions of justice and individual freedom, a practical orientation of affirmation and identification is appropriate, as we saw above. In unjust and unfree societies, a wider range of attitudes are called for. I will call a practical orientation *appropriate* when it successfully reflects the normative character of the social world.

For individuals to have an appropriate practical orientation toward their social world, they must understand its normative character, for at least two reasons. First, one's practical orientation is a reflective attitude toward one's social world, and it embodies deliberative autonomy in part by being based on accessible reasons rather than developed by luck or due to unreflective habituation. When one understands some phenomenon, one can articulate the reasons that facts of interest obtain (Zagzebski 2001). Second, one's practical orientation guides action. To do so, the individual needs to understand her social world rather than know a set of disjointed facts about it. Understanding organizes information in some domain, allowing one to make inferences about new phenomena in that domain (Elgin 1996). Without understanding, one's practical orientation will be an unreliable guide to action.

A practical orientation is also a practical attitude—it is an attitude that is aimed at realization in one's social world (Neuhouser 2000, 111). An orientation of indifference, for example, may lead an individual to unreflectively conform to the prevailing norms, whereas an attitude of rejection may lead to protesting or opting out of certain social arrangements. A practical orientation thus is not a purely theoretical attitude consisting only in a set of beliefs about the normative character of the social world. Rather, it consists in a way of relating to the social world in light of its normative character, one that often centers around how one relates to the social roles that make up that social world (Hardimon 1994, 17). Thus, it can embody a species of practical freedom as well.

Societies in which individuals are not fully free are societies in which they are alienated. Individuals are alienated in part because their institutions do not guarantee the conditions for their freedom, even if they do not realize that they are living under such unfree conditions. Hegel

---

[2]Thus, the subjective component of being at home is a sort of "satisfaction of the will" in virtue of which it represents a species of freedom (Neuhouser 2000, 111). It is the attitudinal recognition of one's state of positive freedom, i.e., "the enduring assurance that one inhabits a world whose basic framework makes it capable in principle of accommodating one's most fundamental practical ends" (111).

[3]This paper assumes that possessing an appropriate practical orientation partly requires understanding the normative character of one's social world. This claim is compatible with the claim that the attitudes and actions that constitute an individual's practical orientation are deeply shaped by one's social world (Haslanger 2019).

calls this type of alienation "objective alienation" (Hardimon 1994, 119–21). What is more distinctive about Hegel's framework is the account of so-called subjective alienation. Individuals are also alienated when they are systematically prevented from grasping the normative character of their social world regardless of the content of its normative character. In other words, individuals are prevented from developing an appropriate practical orientation. It is this second type of alienation that is the focus of this paper.

## 3. A practical orientation at work

In this section, I argue that explanations of the structure and functioning of workplaces and economic institutions that use AI for decision-making are noninstrumentally valuable. While this paper is particularly interested in the noninstrumental value of explanations of workplaces that use AI, the arguments of this section apply to the workplace and economic institutions generally. The focus on AI is important because AI makes workplaces more vulnerable to limiting the normative explanations required to form an appropriate practical orientation, as we will see in sections 4 through 6.

   Given the discussion of the previous section, one might be puzzled that the arguments below target explanation, not understanding. However, I assume a constitutive connection between normative explanation and understanding the normative character of one's social world. A normative explanation just is an explanation of a normative fact, partly in virtue of other normative facts. And, to understand the normative character of one's social world just is to grasp a correct normative explanation of its character.[4] When grasped, such explanations constitute the requisite self-understanding by which we orient ourselves in the social world. If this assumption is correct, then phrasing the argument in terms of understanding rather than explanation does not make a difference.

   Furthermore, I focus on explanations because I am interested in the normative defects that can arise when institutions use AI for decision-making. It would be too strong a moral requirement on institutions that they engender understanding in those subject to AI decisions: such institutions may violate an individual's personal prerogative to pursue projects in other domains, or be intolerably costly, given the different knowledge and cognitive capacities of individuals. However, it is a plausible moral target that everyone has access to the conditions that tend to enable them to develop an appropriate practical orientation toward their social world. And, one such condition is the availability of normative explanations.

   The argument for the noninstrumental value of normative explanations of one's workplace and economic institutions starts from the Hegelian commitment that understanding the structure and functioning of one's social world is noninstrumentally valuable because it allows one to form an appropriate practical orientation toward it. Political philosophers in the Hegelian tradition take the social world to be made up of society's basic institutions and the social norms and practices in those institutions (Hardimon 1994). However, in this paper, I take the social world to be made up of both the basic institutions of one's society and the local organizations and norms that structure one's political, economic, personal, and civic life. A practical orientation toward one's local context is often the means by which one forms a practical orientation toward one's social institutions. I learn about the justness of my society's educational institutions through attending school, and through my friend's experiences at school. However, my local context is not merely instrumentally useful; it partly constitutes my practical orientation to the educational system. This is because my practical orientation shapes and is shaped by my social role as a student, and social roles are both globally and

---

[4]Strevens (2013) argues that scientific understanding is produced by grasping a correct scientific explanation.

locally defined.[5] My school district may attribute certain rights and duties to the role of a teacher that are uncommon in my society or vice versa.

One's workplace and economic institutions are central to one's social world. Work is both time consuming and demanding: people spend a huge portion of their lives at work, and nearly all of the work one can do in modern societies is physically, emotionally, and intellectually demanding. But the workplace is also a site for many of the goods that people have reason to want (Gheaus and Herzog 2018). To form an appropriate practical orientation toward one's workplace and economic institutions requires understanding whether they indeed provide the goods that people have reason to want, for themselves and for others. And, since AI increasingly determines the structure and functioning of workplaces and economic institutions, explanations of automated decision-making are noninstrumentally valuable.

To further understand the argument and its implications for societies that rely on AI in economic decision-making, we need to dig a bit deeper into why understanding—and thereby normative explanations—is necessary for an appropriate practical orientation. Understanding is necessary for epistemic reasons. It can be difficult to know whether one's social world lives up to the requirements of justice, freedom, and solidarity. That is particularly true of the modern economy, which is complex and contains much burdensome work. Often, workers are not in a good epistemic position to directly perceive the normative character of their economic system as well as their own work. This point is naturally supported by a Marxian account of capitalist economic production, where, because of how economic production is socially structured, the normative character of economic relations—e.g., that workers are exploited—is different from how they appear—e.g., that workers are fully compensated for their labor (Cohen 2001). Even if one does not follow Marx, however, there are good reasons to take economic institutions and workplaces to be opaque, and this opacity to ground the noninstrumental value of understanding them. Because of the division of labor, knowledge is distributed throughout an organization, preventing individuals from directly perceiving the normative character of their own work (Herzog 2018). And, even a workplace that is good for most can have work that is demanding, boring, or dangerous for some; thus, the perceptible nature of one's work can be unreliable evidence for the normative character of one's workplace. Market-based economic institutions are also opaque because they are complex: price signals, for example, aggregate information from heterogenous individuals so that individuals can act on that information without understanding the determinants of the price (Hayek 1948). Since the normative character of one's workplace and economy is not often immediately apparent to workers, they require explanation-induced understanding.

Normative explanations are also required for practical reasons. In the economy and one's workplace, individuals act within socially circumscribed roles—teacher, supermarket clerk, police officer, working class, employed or not—that are both institutionally and locally defined. For individuals to play their role well, it is not enough for individuals to know what their social role requires of them; they need to understand the normative character of what they do and their institutions. That is because social roles do not completely specify what one ought to do in all the circumstances that one will face qua role occupier (Zheng 2018). In light of the underspecification, individuals ought to fill in those role obligations in a way that reflects their own moral understanding of how to occupy the role well (Cohen 1967; Zheng 2018). To do so, they ought to act from a practical orientation grounded in an understanding of their social world's normative character. Thus, having normative explanations of the structure and functioning of one's workplace and economic institutions is noninstrumentally valuable because they are constitutive of the practical orientation that allows one to play one's roles well.

---

[5]I take social roles to be sets of predictive and normative expectations that apply to individuals in virtue of the relations they stand in with others and whose violations are backed by sanctions (Zheng 2018).

We are now in a position to see what kinds of normative explanations of the structure and functioning of one's social world are noninstrumentally valuable constituents of an individual's practical orientation. Normative explanations do not merely tell people what practical orientation to adopt; after all, a practical orientation embodies a kind of deliberative and practical freedom. Instead, they put individuals in a better epistemic position to take a stance on the normative character of their social world, and to act out of that stance. Thus, normative explanations should explain the normative character of the social world in terms of how it is. And, to explain how the social world is just is to explain how parts of it function—e.g., what role they play in some larger system—and how it is structured—e.g., what positions and relations make up the part of the social world of interest (Haslanger 2020).[6] Such explanations enable individuals to reflect on the normative character of the social world because they better understand how it works.

Especially important are what I will call *affirmative* and *undermining* explanations. Suppose that members of your society are forced to spend much of their childhood in school. If you possess an *affirmative normative explanation* of that fact—one which explains why this fact makes the institution have a normative property you have reason to want—your student days will be lived more freely. The point is not that you are, in fact, free, because the state may legitimately mandate schooling, and understanding this allows you to understand the situation you are in. Nor is it (only) that you will feel less constrained. Understanding why it is good that children be made to spend so much time in school, and hence the point of being a student, allows you to relate to your school and educational system freely in your daily interactions with others. For example, it enables you to fill out the indeterminate role of student in a context-specific way, and to take meaning from it that you otherwise could not.

Of course, some economic arrangements are not endorsable. In such cases, individuals ought to take a practical orientation toward those arrangements that is rooted in an *undermining* explanation. For example, if mandatory schooling ought to advance substantive equality of opportunity, then understanding how the quality of education in a society depends on the race and class of the student, together with the role of education in social reproduction, will guide individuals in settling on what attitude to take toward the educational system. Such undermining explanations are especially important in unjust societies in which the injustice is hidden. Indeed, unjust institutions are often stable because they obscure the injustice of their functioning, especially through dominant group practices of perpetuating ignorance (Cohen 2001; Mills 2017). In such cases, a normative explanation may undermine an institution by revealing its function. Such explanations may contradict the widespread beliefs and cognitive habits of many who live under such institutions and are necessary to reveal injustice.

This section has defended the claim that normative explanations of the structure and functioning of one's workplace and economic institutions are noninstrumentally valuable. We will now turn to the topic of how AI makes the modern workplace more vulnerable to systematically limiting the availability of the normative explanations that individuals need to develop a practical orientation toward their workplace and economic institutions. The next three sections examine three different sources of opacity: technical (section 4), worker isolation (section 5), and managerial control (section 6).

## 4. Alienation and opacity: Technical opacity

The first mechanism by which workplaces become opaque to workers is *technical opacity*. Technical opacity has received the most attention in philosophy (Creel 2020; Zerilli et al. 2019), computer

---

[6]This claim is compatible with both an ontological grounding—the social world is made up of social structures that do not reduce to individuals and their interactions—or an epistemic one—the social world is only made up of individuals and their interactions, but it is easier to understand the large-scale coordination of individual actions by studying the structure and function of workplaces and institutions.

science (Doshi-Velez and Kim 2017), and the law (Barocas and Selbst 2018). Some algorithmic systems are opaque to interested parties because the data and trained model are kept secret, backed by trade secrecy protections, or because the interested parties do not have the relevant technical expertise (Burrell 2016). However, the concern about technical opacity is a concern about in principle explainability, i.e., that, in principle, some algorithmic outputs cannot be explained in a way that would be understandable to even an expert.[7] Opaque algorithms thus seem to pose a devastating threat to the ability of individuals to understand their workplace, and, thus, to develop an appropriate practical orientation toward it.

This thought requires some unpacking, beginning with the properties of algorithms that make them opaque. Many of these opaque algorithms are developed using techniques from machine learning. Machine learning utilizes vast data sets to find surprising correlations that are used to tackle complex problems. Consider the problem of spam filtering. Email users are often—but not always!—good at recognizing spam, but they would be hard pressed to articulate a rule to reliably classify spam. To tackle this problem, machine-learning methods can be used to construct models with thousands of variables, often connected by a complicated, nonlinear function. The complexity of the resulting models makes them effective at filtering spam, but also extremely difficult for human beings to understand given our cognitive limitations.

Complexity, of course, is not always a barrier to understanding. The natural world is undeniably complex, yet the sciences have developed methodologies for the discovery of its laws and causal structure. However, there are at least two techniques that are central to understanding the natural world but look to be unavailable in the case of complex algorithms. Scientists construct simplified models by idealizing—e.g., deliberating introducing false statements about a target system—and abstracting—e.g., omitting certain properties of a system. Idealizing and abstracting allow scientists to simplify models by reducing the number of variables. They can thereby highlight the important explanatory relationships in a system, which are often causal.[8]

Idealizing and abstracting are made difficult by the complexity of machine learning algorithms. Their complexity makes it difficult to isolate important variables and to construct simple equations that capture counterfactual dependencies between those variables.[9] This inability to pick out a smaller set of explanatorily relevant variables and simple relationships between those variables is a key source of the lack of in principle explainability.

Part of the explanation for this failure is machine learning's detection of correlations rather than causation. Even if one could construct such equations, machine learning's reliance on correlations would leave an expert human user no more enlightened as to why—in any explanatory sense of "why," especially a causal sense—the model outputs the value that it does. As Barocas and Selbst (2018) discuss, machine learning is often used to generate predictive models because decision-makers do not have robust, predictively powerful causal generalizations; if they did, models developed through machine learning would be a waste of resources. But, because techniques from machine learning for model generation are used precisely in those areas where modelers have struggled, the predictively powerful correlations that they exploit tend to be neither causal nor intuitive. Why would someone's facial movements be predictive of their employability, for example?[10]

---

[7]Doshi-Velez and Kim (2017, 2) gloss explainability as "the ability to explain or to present in understandable terms to a human."

[8]For different accounts of the nature and value of idealization and abstraction in the sciences, see Potochnik (2017), Strevens (2008), and Weisberg (2013, chap. 6).

[9]Barocas and Selbst (2018) cite linearity, monotonicity, continuity, and dimensionality as four properties of complex machine-learning models that ground their complexity.

[10]Companies such as HireVue offer algorithmically driven assessments of the employability of job applicants based on data from video interviews. This hypothetical example is not intended to suggest that actual models developed by companies like HireVue are predictively accurate.

A caveat is in order here. The relative opacity of different types of algorithms depends on both human psychology and advances in computer science. Computer scientists have developed techniques to increase explainability by creating simpler approximations of the model,[11] or by providing local counterfactual explanations, which show how perturbing some input will change the model's prediction.[12] Of course, the deep issue that machine-learning algorithms produce classifications and predictions based on correlations rather than causation remains, without explicit causal modeling.

Technical opacity, however, is neither a devastating blow for the use of complex AI in the workplace, nor the most serious threat to the availability of normative explanations. The deployment of technically opaque algorithms can be compatible with workers understanding their workplace and economic institutions. That is because the required normative explanations will not, by and large, cite details of the specific algorithmic criteria behind particular decisions. Many undermining explanations, for example, do not require details of how the algorithm converts inputs into outputs. Instead, a normative explanation in terms of the end for which the AI system has been designed or the function that it plays in the social world will be a sufficient normative explanation. That is because, firstly, if a decision aid or institution has a bad end, one need not know the details of how it operates to know why it should not exist at all. For example, if incarceration should not depend on access to bail money at all, then it is not noninstrumentally valuable to be able to explain why an algorithm set your bail at a particular sum. Likewise, taking the appropriate practical orientation toward your society's failure to perform some important function does not always require causal understanding of how it operates instead. One does not, for instance, need to know the details of how the US health insurance industry sets the price of coverage to know why it is wrong that healthcare is not widely available.[13] And indeed, causal explanations of an institution's functioning in terms of causal-historical details can often be misleading. This is particularly true of explanations of an institution or organization in terms of individual intentions, as an institution can perform a function that no individual nor group agent intends.[14]

Sometimes, the details of how the rules of algorithmic systems structure one's workplace or economic institutions will be important to form an appropriate practical orientation. Taking a practical stance on one's workplace requires understanding the rules, aims, and practices in the workplace—how one's fellow workers are treated, how one's workplace contributes to social reproduction and human flourishing (or not), and so on. As argued above, understanding the social world sometimes requires understanding how its structural and functional components work. In the case of AI, this requires knowledge of the abstract rules of the AI system—so-called *functional transparency*, or the rules connecting inputs to outputs.[15] For example, when the US state of Indiana automated its welfare system, the number of denied applications doubled to over one million in three years. Understanding that the system labels any errors in applicant paperwork as "failure to cooperate," leading to automatic cessation of benefits within a month, can help reveal that the system's function is to police and punish low-income residents of Indiana (Eubanks 2018).

---

[11]See Bastani, Kim, and Bastani (2017) for one attempt to use machine learning to develop a technique to approximate a more complex model using a simpler model.

[12]See Ross, Hughes, and Doshi-Velez (2017) for an example of this approach to increasing explainability, which aims to learn model-agnostic and domain-general decision rules that show how perturbing an input changes a prediction.

[13]Having a causal-historical explanation of why the twentieth-century movement for universal healthcare in the US did not succeed, or why the US has done little to combat climate change, does constitute valuable understanding of a problematic feature of the social world. But this problem is distinct from the problems of the lack of access to healthcare and the threat of climate change. It is rather what Jaeggi (2018, chap. 4) calls a second-order problem—a problem with how a society handles problems.

[14]Haslanger (2020) uses the example of a local school district's policy that students who are late to class more than nine times a term fail. The policy was intended to increase student attendance, making the goods of education available to all. Instead, the policy adversely impacted lower income students that rely on public transport, which is often late.

[15]See Creel (2020) for the distinction between functional and other kinds of algorithmic transparency.

Functional transparency, however, is compatible with opacity about how the algorithm is realized in code, or how a particular output was produced based on the input data (Creel 2020). So, technical opacity is not a strong barrier to the ability of individuals to develop an appropriate practical orientation toward a workplace or economy that embeds AI in decision-making.

By focusing on technical opacity, much of the academic and political fervor over opaque AI has been misdirected. Generally, technical opacity need not undermine individual's ability to form an appropriate practical orientation toward their workplace. But, the threat of algorithmic opacity to individuals' social freedom does not stop with technical opacity. As work in sociology and economics shows, other properties of AI systems also tend to limit the availability of normative explanations. Workplaces become opaque to workers due to a loss of control and isolation. The use of AI in the workplace and wider economy makes these institutions more vulnerable to limiting the availability of normative explanations because of core properties of data gathering and AI, such as extensive surveillance in the workplace, learning, the use of proxies, scale, and matching.

## 5. Alienation and opacity: Isolation

Sections 5 and 6 examine how opacity is created when opaque AI is embedded in modern capitalist workplaces and economies. They identify two categories of mechanisms that undermine the availability of normative explanations: isolation and loss of control. Isolation and loss of control may be bad in and of themselves. However, in this and the next section, I am interested in how isolation and loss of control hinder worker understanding of their workplace and economic institutions. In other words, I am interested in their downstream effects, especially regarding the production of subjective alienation.

In this section, I focus on how isolation, especially the isolation produced by AI-enabled hyperspecialization and physical isolation, can undermine the availability of normative explanations and produce alienation. The phenomenology of this type of alienation is the phenomenology of the automaton who carries out tasks at work without understanding why they are doing what they're doing nor do they understand the conditions in which others work. Cohen (1996–1997) calls this type of alienation "ontological," as man acts as an unreflective productive machine.

While opacity due to the division of labor is an issue in centralized organizations, opacity due to worker isolation is most dramatic in the gig economy.[16] In the gig economy, platforms such as Uber, Amazon's Mechanical Turk, and TaskRabbit match individual laborers to tasks at certain price points.[17] These platforms use an Application Programming Interface (API) that defines a list of instructions that the program will accept, as well as how each instruction will be executed. Using APIs, businesses or private individuals can outsource projects for so-called "human computation," as long as those tasks can be broken down into discrete microtasks.

Take, for example, work done by Ayesha, a gig worker in Hyderabad who uses CrowdFlow to do paid tasks for companies such as Uber (Gray and Suri 2019, xv–xvi). Uber's Real-Time ID check software uses AI to check whether identity check selfies match the photo ID on record. AI flags any discrepancies between photographs—say, because a driver, Sam, has shaved off his beard recently, but has a beard in his photo ID—and a task worker like Ayesha receives those photographs and is

---

[16]According to the World Bank (2019), 6 percent of the world's labor force is part of the gig economy. One in three adults in the United States in 2019 earned money from gig work, but only one in ten are "regular" gig workers who work more than twenty hours a week, and only 13 percent of adults did so through an online platform (Board of Governors of the Federal Reserve System, 2020). Participation in the gig economy is higher in developing countries.

[17]As Gray and Suri (2019, chap. 2) stress, the practice of hiring individuals for a discrete project, as well as the persistence of human labor despite automation of certain work processes, are not new phenomena. Indeed, the political gains of robust legal protections by unions for certain kinds of employment—mainly full time, factory employment, not contract work—is something of a historical anomaly of the twentieth century.

paid to judge whether it is Sam in both. Workers compete for such tasks, build up a record of successful task completion, and receive payments, all mediated through the platform's API.

Modern computing has thus enabled so-called hyperspecialization, where the labor required to create a consumer good is broken down into many different tasks performed by individuals who specialize at those tasks. Thanks to modern computing's ability to send information at basically no cost, it is possible to divide up the production of intangible, knowledge-based goods even more finely and to coordinate the output of those tasks to produce the good. Instead of the eighteen separate tasks in Adam Smith's pin factory, we now have humans taking discrete chunks from an audio file and transcribing them.

The hyperspecialized division of work into microtasks limits the availability of the normative explanations required to develop a practical orientation toward one's work. Because larger tasks are broken down into microtasks, individuals are prevented from knowing what larger task they are contributing to.[18] Labelers of images for an image database may not even know that they are labeling images for an image database, much less what the database is for. Furthermore, the performance of microtasks often does not give workers access to evidence about the relevant normative properties of the larger task, such as whether images from some social groups or geopolitical regions are inappropriately overrepresented in the database. Algorithmically powered hyperspecialization is an epistemic barrier in the workplace: it prevents individuals from understanding what exactly they are doing, which prevents them from forming an appropriate practical orientation toward it.

Not only do platforms enable hyperspecialization, platform design enables platform companies and the businesses that use platforms to keep the structure of work conditions opaque. Technology has spurred the gradual dismantling of traditional employment in favor of platform-mediated contingent work. Platforms are generally designed so that the rules that structure work and the reasons behind decisions are opaque to workers (Gray and Suri 2019). Such design choices serve the interests of platforms and companies by, say, avoiding costly adjudication over nonpayment on the basis of a judgment of the quality of work, or suspension from the platform.

Furthermore, APIs generally do not build in a way for workers to communicate, and gig workers are usually physically isolated and are not working in teams. Thanks to hyperspecialization and the ability to combine work tasks remotely into a single product, workers no longer need to work on (colocated) teams for work to proceed efficiently. And because of surveillance capabilities, algorithms can be used to direct workers, even in real time, which also reduces the need for worker to work in teams, or for a manager to communicate instructions and feedback to them. For example, workers using chat channels can be monitored in real time and automatically nudged by a chatbot to take a poll about next steps (Zhou, Valentine, and Bernstein 2018). They may complete a task without ever communicating with a manager or fellow worker (Gray and Suri 2019). This social isolation prevents workers from sharing information about work conditions, or querying a manager or other relevant authority. Thus, this social isolation prevents workers from having access to the normative explanations that would help them develop an understanding of the structure and functioning of their workplace or the gig economy.[19]

---

[18]This is a general organizational problem (Herzog 2018).

[19]Although, as Gray and Suri (2019, 124–29) discuss, there is more collaboration between platform workers than one might expect. Workers in India, for example, who do not have a government-issued identification that matches a home address—necessary for working on MTurk—sometimes collaborate with those with a functioning MTurk account who no longer work on the platform, sharing profits in exchange for platform access. More experienced workers guide friends to trustworthy platforms, share tips about tasks via messaging apps or online forums, and collaborate on tasks with partners or friends. Worker connection is not merely to facilitate higher earnings since workers often nourish connections and share information that come at a cost to their own earning potential. This connection instead illustrates "workers' need for connection, validation, recognition, and feedback" (138).

## 6. Alienation and opacity: Control

Alienated relations generally involve a loss of power and an attendant feeling of a loss of control (Jaeggi 2014). Economic conceptions of alienation, for example, often locate alienation in a loss of control over the means of production. Here I am interested in a more general loss of control over the social conditions that set out the possibilities in which one acts. I will discuss two ways in which AI can reduce the availability of normative explanations by reducing worker control: (1) by leading to workplaces whose rules "take on a life of their own" and (2) by enabling new forms of managerial control.

Alongside ontological alienation, this section also focuses on the phenomenology of the divided self, another hallmark of alienation. Cohen (1996–1997) calls this sort of alienation "psychological," because the agent has contradictory judgments and values, and because what she does goes against some of those judgments and values. Psyche and society are, as Cohen says, at odds with each other.

A hallmark of a general loss of control is the feeling that an institution has taken on "a life of its own." An example is standardized testing in the United States secondary school system. While standardized testing can serve valuable ends such as social mobility, it can also lead to the phenomenon of "teaching to the test." Teachers in school systems with heavy standardized testing are often frustrated that the rich set of ends that education can realize are narrowed to a single end—performance on the test. This external redefinition of what it is to be a good teacher—enabling students to be successful on a standardized test—induces feelings of loss of control.

Institutions that utilize AI for decision-making are especially prone to becoming institutions that take on a life of their own. This tendency is grounded in four properties of AI: learning, a formal language, optimization for a small set of goals, and scale.

The "learning" of machine learning refers to the use of an algorithm to build a model to perform a specified task of interest based on training data. For example, a researcher may want to build a model to detect COVID-19 in chest radiographs. They could use a machine-learning algorithm to learn a predictive model based on examples of chest radiographs from patients with and without COVID-19. Learning allows artificial agents to narrow down the space of hypotheses in response to experience. It is a particularly useful method for building models when scientists or decision makers have a poor approximation of the actual function that generates the observed data. For example, a company may employ a hiring algorithm if it does not have reliable rules or heuristics to select job applicants that would be productive employees.

Learning thus can—and often will—produce models that change the decision rules since its increased classificatory and predictive accuracy is due to learning more and new patterns in the data. And, more mundanely, machine learning generates models that can take over tasks from human workers. Learning thereby redefines the relevant role, rules, and values in the workplace, or in the economy, if the model is used widely. This change can happen directly if those rules are known, or the introduction of automated decision-making creates new tasks for workers. It can also happen indirectly in cases where the rules are opaque. If a hiring algorithm uses a quality $q$ to predict worker retention, then hiring workers with that quality will change the types of people in the workplace and may produce a shift in how individuals understand and perform their roles.

Institutions that use AI will also be more likely to take on a life of their own because AI uses a formal language and because of optimization. The learning done by artificial agents relies on data that can be processed by the learning algorithm in a formal language. The data must also be available. Both of these requirements plague data science projects in the workplace, and data scientists must work with managers or administrators to define the task in such a way that a predictively useful model can be learned on the basis of existing data (Passi and Barocas 2019). To satisfy these requirements, data scientists often define the target variable in such a way that it acts as a proxy for the underlying variable of interest. Say, for example, that an employer wanted to learn a model to predict which job applicants will stay at the company as an input to hiring. Even that straightforward target would need to be operationalized as, say, the task of predicting which job applicants will stay at the company for at least five years.

Furthermore, AI-based decision systems optimize for a single or small set of goals. To run with the hiring example, the imagined AI system optimizes for employee retention for five years, thereby ignoring other goals in hiring.[20] This imposes a single, homogenous notion of what it is to be a good employee.[21] And finally, because AI systems can be implemented at scale, they can standardize decision-making across a large workplace or the economy.[22]

I do not take optimization, using proxies in one's decision-making, or using a single decision model at scale to be intrinsically bad. But doing so tends to engender an instrumental mindset within institutions that can lead goals and metrics to take on a life of their own. In such cases, decision makers come to value a single goal. In addition, managers and workers often value the proxy in itself rather than the end it represents. The US News and World Report college rankings are an example of how proxies can replace ends as the site of value. Those rankings, which inform students' decisions about which universities to apply to, are based on proxies for, among other things, student welfare, such as the number of athletic facilities. Many universities have responded by building more and nicer athletic facilities to rise in the rankings without, it seems, a regard for the underlying value, as those resources could often better improve student welfare if directed elsewhere.

When institutions take on a life of their own, the availability of normative explanations is undermined. It will also tend to result in psychological alienation, as AI's redefinition of institutional rules or roles is discontinuous with the rules and role definitions it replaces. The surprising patterns in the data discovered in the learning process are exploitable for purposes of prediction but are often out of step with individuals' interpretation of their roles. Furthermore, some learning processes, such as unsupervised learning, can produce algorithms that are unintelligible to agents because they contain concepts and correlations that are highly gerrymandered or semantically uninterpretable. Thus, individuals end up with conflicting attitudes, or conflicts between their attitudes and actions, because they are not able to integrate the institutional roles within their broader practical orientation toward their social world.[23] In such situations, normative explanations are psychologically unavailable because individuals are unable to grasp and integrate them. Of course, individuals may internalize the metrics and new rules defined by an AI system and allow those metrics and rules to guide their behavior, potentially resulting in ontological alienation.[24]

Managerial control[25] can also reduce the systemic availability of normative explanations. Advances in data collection and AI-based model building have dramatically reorganized the operations of firms and markets, and are often touted for increasing productivity and enabling learning and evidence-based decision-making. But this reorganization has also changed the landscape of organizational control within firms, allowing managers new and greater means to exercise control over workers, especially through directing, evaluating, and rewarding them using AI-powered tools (Kellogg, Valentine, and Christin 2020).

AI, in combination with managerial power, raises serious concerns about coercive control, but sociological studies suggest that bare coercive control tends to be ineffective (Rahman and

---

[20]Coyle and Weller (2020) discuss optimization in a policy context.

[21]Here we can draw on literature from sociology about the standardization imposed by rankings and other methods of quantification to support this point (e.g., Espeland and Sauder 2016).

[22]How particular workplaces respond to metrics or other quantified decision-making aids depends on their context (Christin 2018).

[23]Sociological research has shown that professionals resist new technologies that contradict their professional logic or do not enable what workers want to do (Kellogg 2021).

[24]In organizational sociology, one ideal type of organizational cultures is the *hegemonic* or *disciplinary* culture. A key aspect of such cultures is that rules and sanctions are internalized by employees and applied to themselves, even when they are not sure if they are being monitored (Sewell 1998; Kunda 2006). Organizational sociologists generally associated quantification and metrics with this ideal type (Foucault 1977).

[25]Rahman and Valentine (2021, 3) define managerial control as "the systems or practices that employee managers use to direct attention, motivate, and encourage workers to act in ways that support the organization's purposes."

Valentine 2021). Instead, managers and platforms often use indirect control mechanisms, such as automated nudging and opaque platform design. Such indirect control mechanisms are developed by platforms using findings from behavioral economics (Gino 2017). Uber, for example, exercises significant indirect control over driver behavior by not showing the ride destination or fare before drivers accept a ride, and encourages increased driver availability through misleading messaging about increased demand.[26] They also create meaningless badges and other goals to gamify driving, drawing on research from behavioral economics about how people are motivated by goals. Indirect control mechanisms are especially common in the gig economy because managers and platforms cannot rely on a shared firm culture or clear authority over workers. Since workers are not legally designated as employees of the platform or contracting firm, they need to find ways to incentivize workers to do what they want. Algorithmic nudging thus creates a new form of managerial control.

Nudging also reduces the availability of normative explanations. Nudges remove the human agents who might provide such explanations—managers. Of course, many managers issue orders that are not explained, but workers can still identify whom to go to for such explanations. Furthermore, since some algorithmic direction is manipulative, it bypasses reflection on whether the directed task ought to be performed. Such algorithmic nudges reduce the availability of normative explanations by reducing their cognitive salience to workers. In this way, nudges create ontological alienation. Of course, algorithmic nudges can be obvious and frustrating for workers, creating psychological alienation as well.[27]

Learning, the use of a formal language, and the hyperdivision and real-time direction of work all hinder individuals from developing an appropriate practical orientation by making normative explanations less available.

## 5. Conclusion

This paper uses Hegel's concept of a practical orientation and its connection to freedom and alienation to argue that explanations of the structure and functioning of one's workplace are noninstrumentally valuable. It also diagnosed three mechanisms by which AI tends to make normative explanations systematically unavailable: technical opacity, loss of control, and isolation.

To conclude, I want to situate this paper in wider debates about the values that workplaces and economic institutions ought to embody. Hasn't our normative attention been misdirected, you might ask, by focusing on issues of transparency? AI is making many people's jobs even worse than they already were. Service workers are now at the beck and call of automated scheduling software that predicts customer demand in real time and schedules work on that basis—usually the night before and regardless of whether the worker has a ride to work or can arrange care for dependents. The evaluation and discipline of workers is increasingly mediated by extensive, real-time collection of data about, for example, their keystrokes on work computers, length of in-person interaction with customers in retail jobs, or physical movements in warehouse work.[28] The comprehensiveness of this surveillance is technologically impressive, legally permitted, and morally objectionable.

No amount of transparency can make it rational to identify with a job that has an objectionable purpose, or no purpose at all. However, the noninstrumental value of explanation in AI-structured workplaces matters, despite the ubiquity of work not worth affirming, for three reasons. First, as I have argued throughout the paper, there is value in having a practical orientation to your social world that befits its actual normative character. Second, it bears on reflection about the role that

---

[26]Rosenblat and Stark 2016. Uber is a ride-sharing company that matches drivers of privately owned vehicles with riders willing to pay the rate set by the company for being transported.

[27]Some workers may have an appropriate practical orientation to gig or other work. However, this paper's target is the tendency of AI to reduce the availability of normative explanations that are required to develop a practical orientation to one's workplace or economic institutions.

[28]See Kellogg, Valentine, and Christin (2020, 371) for references.

artificial intelligence might play in a better world. For example, platform-mediated work creates a triadic relationship between workers, managers, and platform workers, offering opportunities for platforms to align with workers or to otherwise reconfigure power relations in the workplace (Kellogg 2021). Third, understanding how social structures shape your work life is part of understanding yourself to share a structural position with others, which can serve, in turn, as a basis for collective action. Thus, there are normatively weighty reasons to ensure that workplaces and economic institutions make normative explanations available to workers, especially those that utilize AI.

**Kate Vredenburgh** is an Assistant Professor in the Department of Logic, Philosophy, and Scientific Method at the London School of Economics.

## References

Barocas, Solon, and Andrew Selbst. 2018. "The Intuitive Appeal of Explainable Machines." *Fordam Law Review* 87 (3): 1085–139.

Bastani, Osbert, Carolyn Kim, and Hamsa Bastani. 2017. "Interpretability via Model Extraction." arXiv:1706.09773.

Board of Governors of the Federal Reserve System. May 2020. "Report on the Economic Well-Being of U.S. Households in 2019." https://www.federalreserve.gov/publications/default.htm.

Burrell, Jenna. 2016. "How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms." *Big Data and Society* 3: 1–12. https://doi.org/10.1177/2053951715622512.

Christin, Angèle. 2018. "Counting Clicks: Quantification and Variation in Web Journalism in the United States and France." *American Journal of Sociology* 123 (5): 1382–415. https://doi.org/10.1086/696137.

Cohen, G. A. 1967. "Beliefs and Roles." *Proceedings of the Aristotelian Society* 67: 17–34.

Cohen, G. A. 2001. *Karl Marx's Theory of History: A Defense*. Expanded ed. Princeton, NJ: Princeton University Press.

Coyle, Diane, and Adrian Weller. 2020. "'Explaining' Machine Learning Reveals Policy Challenges." *Science* 368 (6498): 1433–34.

Creel, Kathleen A. 2020. "Transparency in Complex Computational Systems." *Philosophy of Science* 87 (4): 568–89. https://doi.org/10.1086/709729.

Doshi-Velez, Finale, and Been Kim. 2017. "Toward a Rigorous Science of Interpretable Machine Learning." arXiv:1702.08608.

Elgin, Catherine. 1996. *Considered Judgment*. Princeton, NJ: Princeton University Press.

Espeland, Wendy Nelson, and Michael Sauder. 2016. *Engines of Anxiety: Academic Rankings, Reputation, and Accountability*. New York: Russell Sage.

Eubanks, Virginia. 2018. *Automating Inequality*. New York: St. Martin's Press.

Foucault, Michel. 1977. *Discipline and Punish: The Birth of the Prison*. New York: Knopf.

Gheaus, Anca, and Lisa Herzog. 2018. "The Goods of Work (Other Than Money!)." *Journal of Social Philosophy* 40 (1): 70–89. https://doi.org/10.1111/josp.12140.

Gino, Francesca. 2017. "Uber Shows How Not to Apply Behavioral Economics." *Harvard Business Review*. https://hbr.org/2017/04/uber-shows-how-not-to-apply-behavioral-economics.

Gray, Mary, and Siddharth Suri. 2019. *Ghost Work*. New York: Houghton Mifflin Harcourt.

Hardimon, Michael. 1994. *Hegel's Social Philosophy: The Project of Reconciliation*. Cambridge: Cambridge University Press.

Haslanger, Sally. 2019. "Cognition as a Social Skill." *Australasian Philosophical Review* 3 (1): 5–25.

Haslanger, Sally. 2020. "Failures of Methodological Individualism: The Materiality of Social Systems." *Journal of Social Philosophy* 0 (0): 1–23. https://doi.org/10.1111/josp.12373.

Hayek, Friedrich von. 1948. *Individualism and Economic Order*. Chicago: University of Chicago Press.

Herzog, Lisa. 2018. *Reclaiming the System*. Oxford: Oxford University Press.

Jaeggi, Rahel. 2014. *Alienation*. New York: Columbia University Press.

Jaeggi, Rahel. 2018. *Critique of Forms of Life*. Cambridge, MA: Harvard University Press.

Kellogg, Katherine, Melissa Valentine, and Angèle Christin. 2020. "Algorithms at Work: The New Contested Terrain of Control." *Academy of Management Annals* 14 (1): 366–410.

Kellogg, Katherine. 2021. "Local Adaptation without Work Intensification: Experimentalist Governance of Digital Technology for Mutually Beneficial Role Reconfiguration in Organizations." *Organization Science* 0 (0). https://doi.org/10.1287/orsc.2021.1445.

Korsgaard, Christine. 1983. "Two Distinctions in Value." *The Philosophical Review* 92 (2): 169–95.

Kunda, Gideon. 2006. *Engineering Culture: Control and Commitment in a High-Tech Corporation*. 2nd ed. Philadelphia: Temple University Press.

Mills, Charles. 2017. "White Ignorance." In *Black Rights/White Wrongs: The Critique of Racial Liberalism*, 49–72. Oxford: Oxford University Press.

Neuhouser, Fred. 2000. *Foundations of Hegel's Social Theory: Actualizing Freedom*. Cambridge, MA: Harvard University Press.

Passi, Samir, and Solon Barocas. 2019. "Problem Formulation and Fairness." *Conference on Fairness, Accountability, and Transparency (FAT* '19)*. arXiv:1901.02547.

Potochnik, Angela. 2017. *Idealization and the Aims of Science*. Chicago: University of Chicago Press.

Rahman, Hatim, and Melissa Valentine. 2021. "How Managers Maintain Control through Collaborative Repair: Evidence from Platform-Mediated 'Gigs'." *Organizational Science* 32 (5): 1300–26. https://doi.org/10.1287/orsc.2021.1428.

Rawls, John. 2000. *Lectures on the History of Moral Philosophy*. Cambridge, MA: Harvard University Press.

Rosenblat, Alex, and Luke Stark. 2016. "Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers." *International Journal of Communication* 10: 3758–84.

Ross, Andrew, Michael Hughes, and Finale Doshi-Velez. 2017. "Right for the Right Reasons: Training Differentiable Models by Constraining Their Explanations." arXiv:1703.03717.

Schouten, Gina. 2019. *Liberalism, Neutrality, and the Gendered Division of Labor*. Oxford: Oxford University Press.

Sewell, Graham. 1998. "The Discipline of Teams: The Control of Team-Based Industrial Work through Electronic and Peer Surveillance." *Administrative Science Quarterly* 43 (2): 397–428.

Strevens, Michael. 2008. *Depth*. Cambridge, MA: Harvard University Press.

Strevens, Michael. 2013. "No Understanding without Explanation." *Studies in the History and Philosophy of Science* 44: 510–15.

Venkatasubramanian, Suresh, and Mark Alfano. 2020. "The Philosophical Basis of Algorithmic Recourse." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* 20)*: 284–93.

Vredenburgh, Kate. 2021. "The Right to Explanation." *Journal of Political Philosophy* 0 (0): 1–21. https://doi.org/10.1111/jopp.12262.

Weisberg, Michael. 2013. *Simulation and Similarity*. Oxford: Oxford University Press.

World Bank. 2019. *The World Development Report (WDR) 2019: The Changing Nature of Work*. Washington, DC: World Bank. https://doi.org/10.1596/978-1-4648-1328-3.

Zagzebski, Linda. 2001. "Recovering Understanding." In *Knowledge, Truth, and Duty: Essays on Epistemic Justification, Responsibility, and Virtue*, edited by Matthias Steup, 235–52. New York: Oxford University Press.

Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2019. "Transparency in Human and Algorithmic Decision-Making: Is There a Double Standard?" *Philosophy and Technology* 32: 661–83.

Zheng, Robin. 2018. "What Is My Role in Changing the System? A New Model of Responsibility for Structural Injustice." *Ethical Theory and Moral Practice* 21: 861–85.

Zhou, Sharon, Melissa Valentine, and Michael Bernstein. 2018. "In Search of the Dream Team: Temporally Constrained Multi-Armed Bandits for Identifying Effective Team Structures." Paper presented at the *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3173574.3173682.