ABSTRACT

This paper explores a neglected normative dimension of algorithmic opacity in the workplace and the labor market. We argue that explanations of algorithmic decisions are of final value. Following Hegel, we take explanations of the structure and history of the social world to form the basis for reflective clarification of our practical orientation towards the institutions that play a central role in our life. Using this account of the final value of explanations, we diagnose distinctive normative defects in 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, its decision-making patterns, or the particular 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, e.g., enabling the accountability that is necessary for fair and legitimate institutions. 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, we’d like you to suppose for a moment—if you can—that the above imagined society is indeed perfectly just, and these values are realized. Do you still have the intuition that the above society is objectionable in some way?

We do. We explain and defend this intuition in terms of the non-instrumental value of explanation. In our view, explanations of the structure, functioning, and history of organizations and social institutions are of non-instrumental value because they form the basis for reflective clarification of the all-things-considered practical orientation we each cannot help but take towards 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.

It is a mark of freedom to be able to identify with and affirm the institutions that shape the contours of your life. But 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. Our 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 towards the relationship accordingly, than to be in a pleasurably state of deception. Since taking the proper practical orientation towards the social world requires understanding its normative character, it requires normative explanation. And since the normative character of the social world depends on how it actually functions, normative explanation depends on causal 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, we 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.

¹ Following Korsgaard (1983), we distinguish between final and instrumental goods.
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 automated decision-making in the workplace are of final value. Section 4 argues that economic institutions and workplaces that use AI for decision-making are particularly vulnerable to undermining social freedom by limiting normative explanations. Section 5 discusses why normative explanation is still finally valuable in the AI-structured workplace, despite the prevalence of objectionable work.

2. Practical orientation

Why think that a society that is just is normatively lacking? A way into this thought is through Rawls’ concept of a well-ordered society, in particular, the requirement of publicity. 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.²

In this paper, we are interested in whether this relationship between freedom and publicity holds not only of the basic structure, but also at the “meso-level” of bureaucracies, organizations, and social practices. We will argue that it does. In order to do so, we draw on a framework that we take to embody a view of that that is a generalization of the Rawlsian thought. That framework is Hegel’s account of social freedom, especially the concept of a practical orientation at its center. The concept of a practical orientation is useful to normative defects that arise in societies because their institutional and organizational ends and rules are not transparent to their members due to particular properties of AI – or, so we hope to convince you.³

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. One’s practical orientation is an attitude for which there can be the right kind of reason, namely, whether one’s social world has the normative character which the individual

² Schouten 2019.
³ While not the focus of this paper, we also take the concept of social freedom to be useful in diagnosing why certain ways of structuring the workplace using AI are objectionable. In particular, we have in mind here those workplace relations whose badness is not best explained using concepts such as the legitimate exercise of power or domination (see the discussion between Kolodny (2017) and Anderson (2017)).
takes it to have. But, it is also a practical attitude – it is an attitude that is aimed at realization in one’s social world.\(^4\) 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.\(^5\)

Call a practical orientation appropriate when it is responsive to the actual normative character of the social world. According to the Hegelian framework we adopt here, one ground of the value of an appropriate practical orientation is that having an appropriate practical orientation is partially constitutive of being fully socially free. We take on board Hegel’s thought that it is not only important for freedom that one’s institutions secure the conditions for freedom, namely, self-determination and self-realization. One must also experience one’s actions as free in order to be fully free.\(^6\) Experiencing one’s actions as free, and those actions actually being free, has two aspects. First, one must experience them as self-determined. In order for action to be self-determined within the coercive and constraining institutions of 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.\(^7\) A practical orientation just is this practical attitude towards one’s social world, in which an individual experiences her actions as self-determined and self-realizing.

Individuals can fail to be fully free. Societies in which individuals are not fully free are societies in which they are alienated. Individuals are alienated when and because their institutions do not guarantee the conditions for their freedom, even if they do not realize that they are living under such conditions; Hegel calls this type of alienation “objective alienation.”

---

\(^4\) Neuhouser 2003: 111.
\(^5\) Hardimon 1994: 17.
\(^6\) Hegel terms the former objective freedom, and the latter subjective freedom; together, they make up social freedom. We will sometimes use these terms when they are helpful, but we aim to avoid unnecessary terminology in this paper.
\(^7\) 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” (Neuhouser 2000: 111).
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. It is this second type of alienation that is the focus of this paper.

In the next section, we draw on the concept of a practical orientation to argue for the final value of a practical orientation in the workplace. We then turn to how institutions that rely on AI for decision-making are more vulnerable to undermining freedom in Section 4.

3. A practical orientation at work

In this section, we argue that explanations of automated decision-making in the workplace, labor market, and of the economy are of final value. Below, we give two arguments for this thesis, either of which suffices on its own. The second argument relies on more controversial premises than the first, but provides a strategy to argue for the final value of normative explanations in different domains of the social world.

Both arguments rely on 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.

Both arguments also rely on the necessity of understanding the normative character of one’s social world to have an appropriate practical orientation towards it. This premise follows from what it is to have an appropriate practical orientation. As defined, a practical orientation is a set of attitudes for which there can be the right kind of reasons. 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, for example, licenses attitudes such as rejection and opposition. Understanding the social world’s normative character is required for an appropriate practical orientation for two reasons. First, one’s practical orientation is a reflective attitude towards one’s social world, and the deliberative autonomy that it embodies aims to base the attitude on reasons. It is undesirable if individuals have an appropriate practical orientation merely as a matter of luck or unreflective habit. Second, practical orientations guide action. In order to do so, the individual needs to understand her social world, rather than know a set of disjointed facts.
about it. Without understanding, one’s practical orientation will be an unreliable guide to action.

With that common background in place, let’s move on to the first argument for the final value of normative explanations of algorithmic decision-making in one’s workplace. The argument begins from an empirical premise about what work is like: work is both time-consuming and demanding, but is also be the site of many of the goods we have reason to want. 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. The workplace is also a site for many of the goods that we have reason to want. Because people spend so much time at work, it is, for many, one of a few spheres in which they can develop skills and talents, be part of a community, gain social recognition, and make a social contribution.\(^8\)

The workplace can therefore be a site of self-determination and self-realization. Furthermore, in modern societies, it is one of the few sites thereof, for reasons given above. Since we assume the framework outlined in Section 2, we take it for granted that self-determination and self-realization are of value, and that an economic system that does not provide the conditions for self-realization and self-determination is criticizable on those grounds. To know whether the economic system provides such conditions, individuals require normative explanations of whether the structure and functioning of the economic system do indeed promote self-realization and self-determination.

Workers care, however, not just about whether the economic system tends to promote social freedom for its members. They also care about whether their own work can be so regarded. That is not just because they care about how their own life goes. It is also because a society can tend to provide the conditions to ensure freedom for its members, but fail to do so for some. Such a society would be criticizable for failing to promote freedom for all. But, it can be difficult to recognize when society has failed to do so. Work is often demanding, boring, or dangerous. Workers need to be able to identify when such conditions are a necessary part of getting an otherwise good job done, or when they are exploitative, coercive, or otherwise objectionable. Say, for example, I am a worker in an Amazon warehouse, and I am required to wear a bracelet that uses surveillance technology plus AI to track and optimize my movements. Is this a permissible part of a socially important job, or does it violate my right to

\(^8\) Gheaus and Herzog 2018.
privacy? Furthermore, societies that relegate some of its workers to exploitative jobs often rationalize such exploitation through a stabilizing ideology. And so, workers’ practical orientation must be in part constituted by an understanding of the normative character of their workplace. That allows them to understand how social conditions shape their work, and to learn from others how social conditions shape theirs.

The final value of normative explanations of one’s workplace follows fairly straightforwardly from the value of a practical orientation, once one appreciates the extent to which work governs individuals’ lives. Normative explanations of one’s work are of final value because they are necessary to develop an appropriate practical orientation towards one’s workplace, which seriously and unavoidable shapes one’s social conditions. The final value of normative explanations can also be derived from the value of contributing to a socially valuable institution. Our second argument starts from the claim that in a society, different institutions have different functions, which are partly parasitic on the sphere of activity to which they belong. The function of a society’s basic economic institutions, for example, is to coordinate the production of goods and services that allow society to reproduce itself over time and those within it to flourish.

The second step of the argument claims that an institution’s function determines individuals’ practical identities. In a well-functioning society, individuals are habituated such that contributing to the good of all through work becomes part of their practical identity. This claim is supported by the descriptive fact that, in order for society to reproduce itself, individuals within it need to act in such a way as to reliably uphold its functioning. Furthermore, individuals’ practical identities are partly socially determined because self-realization is a social enterprise: self-realization through playing a role well – such as being a good parent – partly consists in the standing individuals get in the eyes of others. In other

---

9 Unavoidably, unless one can afford to opt out of working.

10 The function we attribute to the economic sphere is based on Hegel’s claim that the function of civil society is to supply the required material goods (Neuhouser 2000: 128). Note that one does not need to be committed to institutions having a certain function, or even the goodness of an institution having a certain function, that is independent of social conditions. Nor need one assume that institutions have a single function for a functionalist analysis to be normatively enlightening (Haslanger 2020).

11 One could make the weaker claim that individuals need merely to act in such a way that robustly promotes the institution’s social function, not from a sense of its function (for this strategy of argument used as a defense of functionalism and rational choice theory, see Pettit 2000).
words, we have reason to want to work for the good of all in the economic realm because we have reason to want to be a contributor to society, and to be recognized as such.\textsuperscript{12}

The final step of the argument claims that individuals need to understand how what they do contributes to an institution’s function, in order for them to have a practical identity based on its function. Let’s say that the function of the economy is to produce the goods needed for all to flourish. Part of individuals’ practical identities is to work for the good of all. In order to develop and act out of this practical identity, they need to understand the normative character of their work, and how the economy functions.\textsuperscript{13} A major reason for this is that without normative explanations of one’s workplace and society’s major economic institutions, one cannot work for the good of all, because one needs to fill out one’s role obligations. The economic sphere is one in which individuals act within socially circumscribed roles – teacher, supermarket clerk, police officer, working class, employed or not. Such roles are both institutionally and locally defined. What it is to be a good teacher, for example, is defined by society’s educational institutions, but it is also defined by the informal social norms and practices of one’s school or school district. Furthermore, social roles do not completely specify what one ought to do in all the circumstances that one will face \textit{qua} role occupier. Of course, individuals could respond to this under-determination of role obligations by acting randomly, or to their own private advantage. Doing so, however, goes against the reason they have to work for the good of all. Individuals thus need to inhabit their roles in such a way that they act well given their understanding of the role, especially their understanding of how it contributes to the good of all.

\textsuperscript{12} It is enough for our purposes that this is a \textit{prima facie} reason -- to be a sufficient reason for an individual, one’s society must be enough like a cooperative enterprise, i.e., structured such that people actually tend to work for the good of all. Also, note that this explanation of the reason we have to work for the good of all is not a purely altruistic one; for Hegel, for example, individuals’ “labor for universal ends, rather than being a “selfless” activity, is the very opposite, namely, a form of self-assertion; for it is through such activity that social members posit, or establish, their identities as particular selves” (Neuhouser 2010: 109).

\textsuperscript{13} One might think that the argument given above is much weaker for capitalist economies, since markets are -- roughly -- efficient. However, invisible hand arguments for the goodness of markets do not apply straightforwardly to the actual world, since they are based on models with highly idealized assumptions. Even if one accepted that working for the good of all meant producing goods and services that satisfied consumer preference with minimal social waste, the mere fact of working in a market economy does not guarantee that one is working for the good of all, nor that one’s economic institutions are designed for workers at large to work for the good of all. And, to compound the problem, the use of price signals to coordinate production and consumption makes economic institutions complex in a way that undermines non-experts’ ability to understand whether the economic system and their contribution to it are functioning for the good of all. (This point could also hold in a socialist society which used prices as signals to coordinate production and consumption, without defining features of capitalism such as private property, e.g., Carens 1981).
To complete both arguments, we need to add a final empirical premise. Not only do adults spend most of their time at work, but the structure and functioning of workplaces is increasingly determined by algorithms. Technological advances in AI have made new jobs available, made others defunct, and changed the skills needed for many others. AI has also heavily impacted the day-to-day functioning in the workplace. AI is used to schedule workers’ shifts, match workers in the gig economy to buyers of their services, nudge office workers to be more productive, or automate certain workplace decisions, such as whether a caller to a suicide hotline is prioritized as high risk. As AI is central to the structure and functioning of many modern workplaces and economies, it is necessary for individuals to have normative explanations of the end of the AI system, as well as how, in broad brushstrokes, it functions.

This section has defended the claim that normative explanations of the structure and functioning of one’s workplace and economic institutions are finally 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 towards their workplace and economic institutions.

4. Freedom in the AI-driven workplace

The previous section argued that explanations of the one’s workplace and economic institutions are of final value. However, those explanations are under threat – a threat that has been uniquely exacerbated by the increasing use of AI at work. While AI has also made work more objectionable, the focus of this section is on the systematic unavailability of normative explanations that could constitute a worker’s practical orientation due to features of AI. The phenomenology of this type of alienation is more like the phenomenology of the automaton, who carries out tasks at work without understanding the normative character of the workplace and economic institutions that require those tasks.\(^\text{14}\)

The first mechanism by which workplaces become opaque to workers is technical opacity (Section 4.1). Technical opacity has received the most attention in philosophy, computer science, and the law, and is the most distinctive AI-driven mechanism. But, the problem 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. These mechanisms are driven by core

\(^{14}\text{The distinction between the two types of phenomenology is from Cohen (1996-1997).}\)
properties of data gathering and AI, such as extensive surveillance in the workplace, learning, the use of proxies, scale, and matching, in virtue of which the AI-driven workplace is more vulnerable to undermining individuals’ ability to form an appropriate practical orientation.

We draw on three types of empirical evidence to argue for these claims about the institutional defects to which organizations and institutions that use AI are particularly vulnerable. The first type is evidence about the technical capabilities of AI systems. The second type is evidence from sociology about the deployment of AI in the workplace. The third type of evidence is anthropological evidence about the phenomenology of workers whose work is structured by opaque AI. We draw on this last source of evidence because we take worker phenomenology to have a rational structure: workers’ feeling of a loss of control, for example, is often a fitting response to having lost control.

4.1 Technical opacity

If individuals’ ability to develop a practical orientation towards their work is threatened, there seems to be a simple solution: just explain the relevant structure and functioning of the organization, and how their work fits in. However, the concerns about algorithmic opacity is a deeper concern, one about in principle explainability. Many in computer science and in the public sphere are concerned that some algorithmic outputs cannot in principle be explained in a way that would be understandable to a human being, given human cognitive limitations. Opaque algorithms thus seem to pose a devastating threat to the ability of individuals to understand their workplace.

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, non-linear 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.

---

15 Doshi-Velez and Kim (2017), p. 2 gloss explainability as “the ability to explain or to present in understandable terms to a human.”
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 are unavailable in the case of opaque algorithms. Scientists construct simplified models by idealizing — deliberating introducing false statements about a target system — and abstracting — omitting certain properties of a system. Idealizing and abstracting allow scientists to simplify models by reducing the number of variables. Scientists can thereby highlight the important explanatory relationships in a system, which are often causal.\(^\text{16}\) 

There are two important steps in the process of idealizing and abstracting that are made difficult by certain properties of machine learning algorithms. First of all, the complexity of these algorithms makes it difficult to isolate important variables and to construct simple equations that capture counterfactual dependencies between those variables. However, even if one could construct such equations, machine learning’s reliance on predictively useful proxies 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. This inability to pick out a smaller set of explanatorily relevant variables is a key source of opacity.

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,\(^\text{17}\) or by providing local counterfactual explanations, which show how perturbing some input will change the model’s prediction.\(^\text{18}\) Of course, the deep issue that machine learning algorithms produce classifications and predictions based on correlations, rather than causation, remains.

Is this a devastating blow for the use of AI in the workplace, if AI-driven decision-making is to be compatible with individuals understanding their work? We don’t think so -- at least, not for technical reasons. Often, a normative explanation in terms of the end for which

\(^{16}\) For different accounts of the nature and value of idealization and abstraction in the sciences, see Potochnik 2017, Strevens 2008, and Weisberg 2013, Chapter 6.

\(^{17}\) 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.

\(^{18}\) 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.
the AI system has been designed will be a sufficient normative explanation. 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 finally valuable (even if it is instrumentally valuable) to be able to explain why an algorithm set your bail at one particular sum. Likewise, taking the appropriate practical orientation towards 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 U.S. health insurance industry sets the price of coverage to know why it is wrong that healthcare is not widely available.¹⁹

Furthermore, in cases where individuals need to understand how the institution or AI system functions, so-called “functional transparency,” or knowledge of the high-level rules that convert inputs into outputs, is often what is required for individuals to understand their work.²⁰ That is because high-level rules usually best explain the structure and functioning of institutions and organizations, as long as individuals have good reason to trust that those rules are indeed implemented well. To form an appropriate practical orientation to a workplace that uses scheduling software, say, one needs to understand its end, such as minimizing workplace costs and maximizing the service of customers. One also needs to understand some of the high-level scheduling rules, such as the rule that anyone can be rescheduled with advance warning of less than twenty-four hours.

While technical opacity limits the availability of normative explanations, it does not seem to always undermine individual’s ability to form an appropriate practical orientation towards their workplace. And, as I will suggest in the rest of this section, is the mechanisms of control and isolation, which opaque and predictively powerful AI enable, that are more devastating to individuals’ ability to understand the modern workplace and their place within it.

¹⁹ Having a causal-historical explanation of why the 20th-century movement for universal healthcare in the U.S. did not succeed, or why the U.S. 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 Rahel Jaeggi calls a second-order problem, a problem with how a society handles problems. See Jaeggi (2018), Chapter 4.

²⁰ See Creel (2020) for the distinction between functional and other kinds of algorithmic transparency.
4.2 Control
A feeling of a loss of control is one of the key hallmarks of alienation that is produced by institutions that use AI for decision-making. Alienated relations generally involve a loss of power. Economic conceptions of alienation, for example, locate alienation in a loss of control over the means of production. Here we are interested in a more general loss of control over the social conditions that set out the possibilities in which you act. We will discuss two ways in which AI can reduce the availability of normative explanations by reducing worker control: by leading to workplaces whose rules “take on a life of their own,” and by enabling new forms of managerial control.

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, that they do not affirm. This external re-definition of what it is to be a good teacher induces feelings of a 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 two features of AI that make it a prized decision-making tool: learning and the use of observational data. Learning allows artificial agents to narrow down the space of hypotheses in response to experience. It is a particularly useful method for model-building 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 trained from company or sector data if it has a poor sense of what type of worker would be a productive employee. Learning thus can -- and often will -- produce models that change the decision rules, as its increased classificatory and predictive accuracy is due to learning more and new patterns in the data. And, changing these decision rules will often re-define the relevant role, rules, and values. This change can happen directly, if those rules are known. But, the change can happen indirectly even in cases where the rules are opaque. If a hiring algorithm uses a quality $q$ to predict worker retention, then

---

21 We will often have to read off those rules from the model, but we find it plausible that many modeling choices, such as choosing a threshold, implies or is rationalized by a normative standard (Hellman forthcoming).
hiring workers with that quality will change the types of people in the workplace, in a way that will impact how individuals understand and perform their roles.

Furthermore, the learning done by artificial agents relies on data that can be processed by the learning algorithm and is available. Both of these requirements often plague data science projects in the workplace, as data scientists 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. To satisfy these requirements, data scientists often define the target variable 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 at least five years without a leave of absence longer than a month.

While there are not in principle problems with using proxies for the relevant end in one’s decision-making, doing so tends to engender an instrumental mindset within institutions, where decision-makers come to value the proxy in itself, rather than the end it represents. Alongside teaching to the test, the US News and World Report college rankings are another 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 up in the rankings, without, it seems, reflecting on whether the resources could better improve student welfare if directed elsewhere.

When institutions take on a life of their own, the availability of normative explanations is undermined. One reason is that AI’s redefinition of institutional rules or roles will often be discontinuous with the rules and role-definitions it replaces. As discussed previously, in cases

---

22 This latter point leads to the somewhat comical situations where companies sometimes first learn a model to predict missing data, and then learn a model to accomplish the task of interest on the basis of those predicted data. One prominent example are models to predict the values of protected attribute variables when companies are legally or reputationally prevented from collecting those data.

23 One might object to this example that student welfare is not the universities’ end here – instead, their goal is to maximize student tuition fees, which they do through rising in the rankings. This reading of the example only strengthens the point made about. Assuming that student welfare was at some point an end of the university educational system, this reading of the example shows how an overly instrumentalist approach to a domain – say, education – can systematically and perniciously shift individuals’ valuing away from appropriate ends to means to those ends – money, one’s standing as determined by some a ranking, etc.
where decision-makers do not have a predictively accurate decision model, successful learning often produces models that achieve the defined task on the basis of surprising patterns in the data. These patterns are exploitable for purposes of prediction but may be 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 are not able to integrate the institutional roles within their broader practical orientation towards their social world, either because of a discontinuity or lack of intelligibility.

Managerial control can also reduce the systemic availability of normative explanations. Advances in data collection and AI-based model-building have dramatically re-organized the operations of firms and markets, and are often touted for increasing productivity and enabling learning and evidence-based decision-making. But, this re-organization 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.

The increased control of workers by managers undermines the availability of normative explanations. The first reason is the division of tasks that AI enables. The combination of real-time surveillance, algorithmically-mediated instantaneous, interactive feedback and nudging, and the use of automated programming interfaces (APIs) to divide tasks into micro-tasks enables the direction of workers at a grain previously only achieved in manufacturing. The division of work into micro-tasks limits the availability of the normative explanations required to develop a practical orientation towards one’s work. For example, because larger tasks are broken down into micro-tasks, individuals are prevented from knowing what larger task they are contributing to. 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 micro-tasks often does not give workers access to evidence about the relevant normative properties of the larger task, such as whether images from particular societies are inappropriately over-represented in the database. Both of

---

24 Kellogg, Valentine, and Christin 2020: 368-369.
25 E.g., Kellogg, Valentine, and Christin 2020.
26 These tools also raise concerns about domination and privacy, which are not our focus here.
27 This is a general organization problem – see Herzog 2018.
these features of the algorithmically mediated division of labor prevent individuals from understanding what, exactly, they are doing, which prevents them from forming an appropriate practical orientation towards it.

The second mechanism is the algorithmic recommendation and restriction that managers use to direct worker behavior. Uber, for example, exercises significant indirect control over driver behavior by not showing ride destination or fare before drivers accept a ride, and encouraging increased driver availability through slightly misleading messaging about predictions of increased demand. Algorithmic nudges reduce the availability of normative explanations. They, first of all, 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, some algorithmic direction is manipulative. Since platforms, such as Uber, claim not to employ those offering services on them, they thereby lose the authority to issue orders, except insofar as platform users fail to comply with the terms of service. Since they cannot order their drivers to continue driving, say, Uber might nudge a driver by suddenly making them aware of higher fees for the next hour. They thus use nudges to produce the desired behavior, in part by bypassing 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.

Learning, operationalization, and the division and direction of work, especially through nudges, all hinder individuals from developing an appropriate practical orientation by making normative explanations less available.

4.3 Isolation

The third mechanism that reduces the availability of normative explanations is that of worker isolation. Such isolation is most dramatic in the gig economy, where platforms such as Uber, Amazon’s Mechanical Turk, and TaskRabbit match individual laborers to tasks at certain price points. These platforms use an Application Programming Interface (or API) that defines a list of instructions that the program will accept, as well as how each instruction will be

---

28 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.

29 As Gray and Suri (2019: Chapter 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 20th century.
executed. Using API’s, businesses or private individuals can outsource projects for so-called “human computation,” as long as they can be broken down into discrete micro-tasks.

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, although he has a beard in his photo ID – and a task worker like Ayesha receives those photographs and is 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.

Technology has spurred the gradual dismantling of traditional employment in favor of platform-mediated contingent work. It has created work arrangements that isolate workers, and thereby has reduced the availability of normative explanations that are constitutive of a practical orientation towards one’s work and society’s economic arrangements. Platforms are generally designed to be opaque to workers. Such design choices serve the interests of platforms and companies, by, say, avoiding costly adjudication over judgments 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 neither co-located nor working on teams. They are thus prevented from sharing information about work conditions that would help them to develop an understanding of the structure and functioning of their workplace or the gig economy. And, they are prevented from sharing information that would allow them to understand themselves as part of a group of people in the same structural condition, the importance of which we will return to in Section 5.

4.4 Conclusion

This section analyzed the alienation that individuals often experience in workplaces and economic institutions driven by AI. This alienation is grounded in a lack of normative

---

30 Although, as Gray and Suri (2019: 124-129) 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. As Gray and Suri discuss, worker connection is not merely to facilitate higher earnings, as 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” (2016: 138).
explanations, produced by technical opacity, a loss of control, and isolation. One upshot of this section is that technical opacity is not the only -- and, indeed, not the most -- worrying contributor to worker alienation.

5. The value of understanding objectionable work

In this final section, we will deal with an objection that questions the significance of the argument. 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.\(^{31}\) 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 final value of explanation in AI-structured workplaces matters, despite the ubiquity of work not worth affirming, for three reasons. First, as we 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 artificial intelligence might play in a better world. 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.

Resistance to the value of an appropriate practical orientation may, we hazard, come from a misconception of a practical orientation as a theoretical attitude. A practical orientation is an attitude, one for which there can be the right kind of reasons, as we’ve stressed, but it is a \textit{practical} attitude. In other words, a practical orientation is not only valuable because it is valuable to have attitudes that are responsive to how things are, i.e., based on the right reasons. It is also relational, and guides how we act in our social worlds and relate to others and the norms, practices, and institutions that shape the conditions of our action. It allows one to

\(^{31}\) See Kellogg, Valentine, and Christin 2020: 371 for references.
more thoughtfully relate to one’s roles and social institutions, and perhaps, in small ways, to thereby improve them. A perceptive elementary school teacher in an unjust society, for example, is a better teacher of students whose parents do not have time or resources to spend much time reading to them at home, say, than is a teacher who assumes that busy parents do not value their children’s’ education.

Second, the value of understanding one’s social world bears on reflection about the role that artificial intelligence might play in a better world. Much of the legal and technical work on AI aims to make it less discriminatory, and, more broadly, more fair. Such work is intellectually and socially important. But, it is important to supplement this work with work on how to make AI and the institutions in which it is embedded more understandable. That is not only because of the value of understanding one’s social world. It is also because in less than fully just societies, a tension between making the social world more procedurally fair and making it more transparent often arises. Procedural fairness, which, very roughly, requires treating like cases alike, is often institutionally guaranteed by a set of standardized rules that remove individual discretion. And, procedural fairness may be enhanced by the use of AI, as, for example, bias is even more deeply hidden in human brains than algorithmic systems. The conceptual resources of this paper offer one diagnosis of the moral concerns created by a widespread loss of discretion, namely, that it comes at a cost to individuals’ understanding of their social world.

Finally, a practical orientation aims not only at understanding whether and why the social world hinders my freedom, but, more generally, whether and why the social world hinders the freedom of others. Understanding how social structures shape your work life is part of understanding yourself to share a structural position with others. Not only that, it is part of understanding the broader structure and functioning of one’s social world, which includes others’ social positions. Both can serve as bases for collective action, alongside shaping the day-to-day ways that individuals relate to each other in the workplace and in other domains.

---

32 Zheng 2018.
33 See Kleinberg, Ludwig, Mullainathan, and Sunstein 2019.
6. Conclusion

This paper used Hegel’s concepts of a practical orientation and its role in freedom to argue that explanations of the structure and functioning of one’s workplace are finally valuable. It also diagnosed three mechanisms by which AI tends to make normative explanations systematically unavailable: technical opacity, a loss of control, and isolation.

It also raises a number of important lines of further investigation. It raises questions in social epistemology, such as questions around the social epistemology of normative explanation. What practices, norms, and institutional structures facilitate individuals’ development of a practical orientation, and which hinder it? It also raises questions in political philosophy, such as questions around the power relations enabled by AI. Are AI-driven forms of control new relations of managerial power, or are they instead a continuation of forms of managerial control rooted in the Industrial Revolution? Finally, this paper raises questions about tensions between fairness and transparency. Are these tensions merely apparent? Is there only a tension in very unjust societies? And, do considerations of fairness always trump the value of transparency?
References

Anderson, E. 2017. Private Government. Princeton: Princeton University Press.

Bastani, O., C. Kim, and H. Bastani. 2017. “Interpretability via Model Extraction.” eprint arXiv:1706.09773.

Carens. 1981. Equality, Moral Incentives, and the Market. Chicago: University of Chicago Press.

Cohen. 1966-1967. “Beliefs and Roles.” Proceedings of the Aristotelian Society.

Creel, K. 2020. “Transparency in complex computational systems.” Philosophy of Science 87: 568-589.

Doshi-Velez, F. and B. Kim. 2017. “Towards A Rigorous Science of Interpretable Machine Learning.” eprint arXiv:1702.08608.

Gheaus and Herzog. 2018. “The Goods of Work (Other than Money!).” Journal of Social Philosophy.

Gray and Suri. 2019. Ghost Work. New York: Houghton Mifflin Harcourt Publisher.

Hardimon, M. 1994. Hegel’s Social Philosophy: the Project of Reconciliation. Cambridge, UK: Cambridge University Press.

Haslanger, S. 2020. “Failures of Methodological Individualism: The Materiality of Social Systems.” Journal of Social Philosophy: 1-23.

Hellman, D. forthcoming. “Measuring Algorithmic Fairness.” Virginia Law Review.

Herzog, L. 2018. Reclaiming the System. Oxford: Oxford University Press.

Jaeggi, R. 2014. Alienation. New York: Columbia University Press.

Jaeggi, R. 2018. Critique of Forms of Life. Cambridge: Harvard University Press.

Kellogg, K., M. Valentine, and A. Christin. 2020. “Algorithms at Work: The New Contested Terrain of Control.” Academy of Management Annals 14(1): 366–410.

Kleinberg, J., J. Ludwig, S. Mullainathan, and C. Sunstein. 2019. “Discrimination in the Age of Algorithms.” Available at SSRN: https://ssrn.com/abstract=3329669 or http://dx.doi.org/10.2139/ssrn.3329669

Kolodny, N. 2017. “Chapter 5 Help Wanted: Subordinates.” In, Anderson, Private Government. Princeton: Princeton University Press, 99—107.

Korsgaard, K. 1983. “Two Distinctions in Value.” The Philosophical Review 92(2): 169—195.

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

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

Rosenblat and Stark. 2016. “Algorithmic Labor and Information Asymmetries: A Case Study of Uber’s Drivers.” International Journal of Communication 10: 3758–3784.
Ross, A., M. Hughes, and F. Doshi-Velez. 2017. “Right for the Right Reasons: Training Differentiable Models by Constraining Their Explanations.” arXiv:1703.03717

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

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

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

Zheng, R. 2018. “What is My Role in Changing the System? A New Model of Responsibility for Structural Injustice.” Ethical Theory and Moral Practice.

Weisberg, M. 2013. Simulation and Similarity. Oxford: Oxford University Press.