Authenticity in algorithm-aided decision-making

Brett Karlan
Purdue University

Forthcoming in *Synthese.*

Abstract

I identify an undertheorized problem with decisions we make with the aid of algorithms: the problem of *inauthenticity*. When we make decisions with the aid of algorithms, we can make ones that go against our commitments and values in a normatively important way. In this paper, I present a framework for algorithm-aided decision-making that can lead to inauthenticity. I then construct a taxonomy of the features of the decision environment that make such outcomes likely, and I discuss three possible solutions to the problem.

1. Introduction

Imagine an overworked and underpaid graduate student undergoing a crisis of confidence. By the dim light of her laptop screen in a dark and empty library, she is trying to decide whether to continue her graduate education in philosophy, or instead to apply to law school. Feeling a genuine uncertainty, she utilizes a number of internet resources (such as Google's search engine and posts from current and former academics on algorithmically-curated social media sites) to weigh reasons for and against each decision. On the basis of the information these algorithms present to her, she decides to send off an apologetic email to her adviser and start looking into the next available LSAT examination.

Or imagine a hiring manager attempting to fill a position at her company. While she is committed to performing the responsibilities of her job to a high standard, she is also overwhelmed with the number of applications she receives. To manage the initial influx, she uses a resume filtering algorithm. The algorithm recommends a set of fifteen resumes (from an initial intake of over 250), from which she contacts six applicants for an initial interview (Deshpande et al. 2020). From this initial screening, she is able to hire an applicant that is well-suited for the role.

Finally, imagine a medical practitioner deciding how to treat a patient. The patient is incapacitated, and attempts to contact her family have met with no success. Among other

medically-relevant parameters for treatment, the practitioner wishes to abide by the patient's own wishes. Lacking any other recourse, the practitioner refers to a patient preference predictor (Rid & Wendler 2014), a (hypothetical, currently) statistical learning algorithm meant to predict whether a patient with certain demographic and medical features would prefer one or another treatment. The algorithm predicts that patients like her more often than not agree to invasive treatment in similar cases. Partially on the basis of this information, the practitioner decides to proceed with an invasive treatment in order to save the patient's life.

Decisions made with the substantive aid of machine learning and other algorithmic technologies (what I will call "algorithm-aided decision-making") have been the focus of considerable normative scrutiny. Some have wondered whether any use of an opaque algorithm that returns a judgment without an explanation is a system that a decision-maker should use *at all*, at least in making important decisions.[1] Trusting an algorithm that utilizes racial or other social categorizations[2] to make impactful decisions seems particularly worrisome, given the well-documented role of algorithms in perpetuating inequality and racial bias.[3] This might be true even if the decision-maker does not know that such a pattern is being exploited. We might also wonder about those on the receiving end of such decisions. If it turns out the patient preferred not to have an invasive treatment, in what way has she been wronged? How might the recipient of a biased algorithm-aided decision have been harmed?[4] Research done on human-algorithm interactions has often focused on these questions.

[1] This informs many of the calls for explainable AI; see Dosilovic et al. (2018), Das & Rad (2020), and Tjoa & Guan (2020), among others. XAI is discussed in detail in section 4.3 below.
[2] As many current technologies do, often indirectly (Johnson 2021).
[3] For instance: Obermeyer et al. (2019) collects evidence of extensive racial bias in popular algorithms used to assess population health. Cavazos et al. (2020) reports recent demonstrations of racial bias in facial recognition algorithms. See also Buolamwini & Gebru (2018).
[4] Like the literature on explainable AI, work on the downstream harms caused by algorithmic bias is both voluminous and growing. Some representative work includes Danks & London (2017) and the overview by Mehrabi et al. (2021).

In this paper, I focus on a related but distinct issue. I present some problems confronting *decision-makers* in algorithm-aided decision-making. The nature and function of the algorithms used in many important decision contexts renders our decisions vulnerable, I argue, to a particular normative failing. Our algorithm-aided decisions will often be *inauthentic*. Using algorithms in substantive ways in our decision-making threatens to render us untrue to our values and commitments, creating a gap between the way an agent wants to operate in the world and the tools she uses to do so. In this paper, I precisify this formulation. I discuss what kind of value authentic decision-making realizes, show how state-of-the-art deep neural networks can threaten authenticity, and discuss solutions for avoiding or alleviating the problem.[5]

The problem of authenticity in algorithmic decision-making is important for a number of reasons. As more and more decisions are made with the aid of proprietary and opaque algorithms, from whom to hire (Langenkamp & Costa 2020) to whom to offer a chance at parole (Brenan et al. 2009), having an account of their normative benefits and costs is crucial to understanding the proper use of algorithms in our lives. In some cases (e.g. fast trading decisions in a financial market; Culkin & Das 2017), a clear understanding of the decision environment might turn up little wrong with algorithmic reliance. In others (e.g. high-stakes medical contexts; Miotto et al. 2016), significant human oversight might be needed. At a more fundamental level, how we do (and should) function as practical agents in a world of widespread information technology use is a fundamental problem at the intersection of

[5] As far as I know, to date there has been little published work on the relationship between authenticity and algorithm-aided decision-making with a view to the decision-maker. The relationship between autonomy and algorithmic decision-making is explored at book length in Rubel, Castro, & Pham (2021). Their focus is on threats to autonomy of those on the recipient end of algorithm-aided decisions, however, not the decision-makers themselves (nor on the value of authenticity as distinct from autonomy).

normative philosophy and the philosophy of the computer sciences.[6] If the picture I sketch here is on the right track, there is a growing divide between the decision-making that many agents value and desire to participate in (as well as the kinds of values they bring to that decision-making) and the kinds of values that are encoded in opaque algorithms that are increasingly being used to aid those decisions. Having a normative theory of these decisions seems like a good first step towards alleviating, or at least understanding, the disconnect. I begin to construct that theory here.[7]

2. Algorithms and authenticity

2.1 A model of algorithm-aided decision-making

We do not make most of our decisions in a vacuum. We rely on a host of external aids to facilitate complex decisions. Whether we are justified in utilizing these aids in turn influences our normative evaluations of our decisions. One important kind of deference is our reliance on the aid of *agents* or information-providing *artifacts* to provide scaffolding for our decision-making. At one extreme, we believe on the basis of testimony and defer to experts.[8] At the other extreme, we rely on non-agential information sources to aid decision-making. (Think, for example, of consulting a map while planning a travel route or checking a reference book for the exact dates of an historical figure.) Whether we are justified in using external aid depends on the evidence we have about the quality of that aid. When agents are involved the evidence

[6] The work of C. Thi Nguyen is an exception to the general paucity of work in this area, and is a direct inspiration for the current project (Nguyen 2019; Nguyen 2022). See also work on digital minimalism as a requirement of agency by Aylsworth & Castro (2021).

[7] I focus throughout this paper only on algorithm-aided decision-making contexts. I do this not because I think these issues fail to arise in other contexts (they almost certainly do), but because the algorithmic context is particularly undertheorized. Future work connecting these cases with more commonplace decision-making is needed. I am grateful to an anonymous reviewer for pushing me to make the scope of this project more clear.

[8] There is a sophisticated and subtle literature on the epistemology of testimony and expert deference which is mostly orthogonal to issues of algorithmic reliance. See Kelly (2005) and Dorst et al. (2021), among many others.

might be particularly complicated, and it might matter (for instance) whether the subject matter we are interested in has an empirical or normative character.[9]

In the decision contexts that interest me in this paper, we occupy an interesting middle ground between agential and non-agential aid. It is controversial whether sufficiently complicated robots should be thought of as agents (Danaher 2019; Nyholm 2020). It is slightly less controversial that state-of-the-art deep artificial neural networks[10] are best not thought of as agents, in part because few of them are designed to act in the world. But consulting an algorithm to help one make complicated decisions is also not obviously analogous to checking a book or a map. Deep learning networks extract patterns from complicated input data that no human reasoner could reasonably expect to sort through. If we want to know whether to use the conclusion of an argument in a book in our deliberation, we can read the argument for ourselves and weigh the reasons the author provides. When consulting deep learning algorithms, no such independent checking is possible.[11]

This is the decision context I am interested in exploring. We are tasked with making an important decision, such as whether to continue pursuing graduate training in philosophy or instead start applying to law schools. We might have available a treasure trove of data, too much for us to sort ourselves; or, perhaps more likely, some company or research team has access to data that we do not. We also have available an algorithm that gives us an output recommendation, and we have no linguistic or other methods for validating that algorithm.

[9] See McGrath (2019) for a recent discussion of reliance on testimony and agents that doubles as an argument for the possibility of gaining moral knowledge across these modalities.

[10] In this paper, I will mostly be focused on our interactions with deep learning neural networks, which attempt to make classifications and other kinds of inferences on large datasets through training (e.g. unsupervised, supervised, and reinforcement learning). These networks are deep, in contrast to shallow connectionist networks from the 1980s, because they feature many layers of nodes that sequentially process information. Buckner (2019) offers an accessible philosophical overview of these networks. They are operant in many fields, but it should be noted that they are not the only kinds of networks available.

[11] Of course, our friends might lie, and our books might be inaccurate. The point is not that these other kinds of aids can never fail us, but that they are the kinds of things that can provide agents with justifiable reasons for acting. Outputs of opaque algorithms are less able to do this.

While these decision contexts are currently somewhat rare, they are both occurring with relative frequency in important contexts (e.g. hiring at large companies; Raghavan et al. 2020) and are likely to become more prevalent in everyday life in the future. It is no longer implausible to imagine that one could soon be making important decisions about how to act in the most important domains of one's life with the aid of algorithms. A catalog of the possible benefits and drawbacks of their use in practical agency is required.

2.2 The value of authenticity

The other main background claim important for this paper is that authentic decision-making is valuable. In very rough terms, to make an authentic decision is to make a decision that aligns with one's values. Slogans that one should "be true to oneself" express something of the value of authenticity. Many theorists have thought that making an authentic decision is uniquely valuable for agents. The notion of authenticity has been discussed in detail in neuroethics, for example, where authors worry that cognitive enhancements might reduce our capacity to make authentic decisions (Parens 2005; Levy 2011; Pugh et al. 2017). This applied notion, in turn, is indebted to a wide range of historical philosophers. Discussions of authenticity can be found in as disparate thinkers as Rosseau, Kant, Nietzsche, Heidegger, and some of the existentialists (Taylor 1992).

In general, there is little agreement on the nature of authenticity. Feldman and Hazlett (2013) catalog at least five distinct meanings of the term in applied ethics alone. Fortunately, we need not adjudicate between competing conceptions of authenticity for the purposes of understanding its role in algorithm-aided decisions. I will instead focus on one suitably broad conception of authenticity formulated by Brink (2003). I do not claim that this is the only way to

precisify authenticity. I only claim that this conception is useful in categorizing an aspect of our decision-making that seems to have value.

Brink writes that "authenticity requires acting on the ideals that the agent reflectively and sincerely accepts at the time of the action" (Brink 2003, p. 251). He compares authenticity to prudence, a value that "appears to require the agent to subordinate her current ideals to her future ones or at least to moderate pursuit of current ideals in light of future ones" (Brink 2003, ibid). Prudence requires us to consider both what we accept now and what we have good reason to think we will accept in the future. Authenticity, in contrast, only requires that we act on what we genuinely believe to be the right thing at the time we are acting. To make an authentic decision is thus to make a decision in line with one's commitments and ideals (in a domain where one has explicit commitments and ideals).

Suppose I am deciding whether to become a corporate lawyer or to continue writing and teaching philosophy. My commitments concerning the good life tell in favor of academic pursuits. I dislike the role of corporate lawyers in upholding the interests of the well-off, and I find the reflective and argumentative life of philosophy conducive to my own happiness. Yet I am tempted by the money and prestige of being a corporate lawyer. Suppose I decide to change course and send out applications to law school. One regrettable thing about my decision is its inauthenticity. It goes against the goals and ideals I reflectively endorse. Even if other negative aspects of being a corporate lawyer were not present (suppose corporate lawyers had the interests of the little guy at heart), we still might think it a shame that I made a decision that went against my own values. I wasn't being true to myself. My decision was inauthentic.

The proposal is in need of some refinement. What is it for someone to act in accordance with their commitments? As Brink formulates authenticity, they rule out cases where I act in alignment with my values for the simple reason that my decision doesn't meaningfully involve

my commitments. If I am deciding between chocolate and vanilla ice cream, there is some sense in which my opting for chocolate is in line with my values. I do not go against any hypothetical pro-vanilla commitments when I make my choice. Such a notion of authentic choice would be too broad, allowing only actions that explicitly contradict my own values to count as inauthentic. Authenticity requires a stricter match between commitments and actions. Other cases seem less straightforward. What about when what I am committed to speaks in favor of two mutually exclusive choices? Or cases where it is unclear what my commitments entail about a particular decision? A theory of authenticity should also make clear what it is for an agent to have a commitment or a set of values in the first place (c.f. Watson 1975). A full analysis of authenticity would require us to get clear on these (and other) questions. For our purposes, however, a rougher characterization is sufficient. We will focus on cases with a clear disconnect between a person's values, understood as their explicit beliefs and commitments in normatively important areas, and their (algorithm-aided) decisions.

It is important to distinguish the notion of authenticity, understood as an alignment between one's values and one's actions, from broader notions of *autonomy* that are also operant in much decision-making.[12] Whereas norms of authenticity recommend the agent be true to themselves, norms of autonomy recommend the agent make decisions in ways that are free and expressive of their rational agency. To make an autonomous choice is to make a choice that is free of coercion and properly respects the decision-maker as an agent. The value of autonomy in algorithmic decision-making has received some significant attention in Rubel, Castro, & Pham (2021). There are interesting questions to be asked about the relationship between autonomous

[12] The literature on autonomy is far too large to cite well here, but it plays a particularly important role in Kantian theories of normativity (see especially Korsgaard 1996) and in discussions of freedom of the will (see especially Hare 1965).

and authentic decision-making.[13] But at this stage, it will be sufficient to note that decisions can be autonomous without thereby being authentic. An agent can freely and rationally choose to go against a particular value or commitment, just as one could be compelled to make a decision that is (as a matter of fact) in line with one's commitments. While the notion of autonomy is certainly operant in many of the cases I discuss,[14] I will be setting those considerations aside going forward.

It is one thing to say that authenticity is valuable. It is quite another to give an account of the nature of that value. I will try to be neutral as to the exact value of authenticity in this paper, though I will defend some aspects of a view below. For Brink, the value of authenticity just is the value of following the dictates of prudence (Brink 2003, pp. 239-41). We should be authentic for the same reason that we should conform to the dictates of practical reason. One could, in contrast, imagine a position that stakes out the value of authenticity as separate from practical and moral value. Should the person who wrongly believes that forgoing her own personal gain to count blades of grass will improve her life do so for the sake of authenticity? Should the person who believes that her protest of an abortion clinic is morally required stick with her decision for the sake of authenticity, even if what she is doing is wrong?[15] If one has the intuition she should not, one does not believe there is a separate value to authentic decision-making *tout court*. If one has the intuition that she should, on the other hand, a distinct

[13] Some have, for instance, conceptualized authenticity as a more-demanding version of broader notions of autonomy (especially in biomedical ethics, e.g. Schwan 2022).

[14] In particular, one might worry that the cases discussed here are similar to nudges, raising concerns for the autonomy of both the nudged and algorithm-aided decision-maker (Schmidt 2017; Di Nuci 2013). While I do not doubt that some cases of algorithmic decision-making are structurally similar to nudging, I doubt both that nudges truly undermine our autonomy in all cases (Levy 2019) and that most of the important cases discussed in this paper are agency-undermining in the relevant way. The framework I develop in section 4.1 expands on this.

[15] This might, in turn, depend on whether one thinks there are more or less demanding epistemic constraints on authenticity. At one extreme, an agent must *know* that some action is in line with her values in order for the decision to be authentic. One worry about this extreme conception is that agents who are in suitably bad epistemic environments cannot make authentic decisions (compare with Ballarini 2022).

view of the value of authenticity is on the table. In what follows, I will not assume that authenticity is independently valuable over and above the general value of conforming to practical reason, since we need not take a stand on the issue in order to make progress in the algorithmic case.

When I claim that authentic decision-making is valuable, I aim to make a claim that isn't particularly controversial. Making decisions that align with one's values and commitments strikes me as a valuable enterprise, and one that we often treat as such. The threat to authenticity from opaque algorithms is, I think, a genuine threat to something we care about and want to uphold in our decisions. This is separate from different possible conceptions of authenticity's value.

3. Inauthentic algorithm-aided decisions

With the basic concepts of authenticity and algorithm-aided decision-making more clearly articulated, we can now return to the examples presented in Section 1. Recall the overworked hiring manager looking to make a first-round cut. Here is one possible precisification of her case:

> **Hiring 1 (H1).** Hallie, a hiring manager at a large technology company, is looking to hire for an entry-level role. She receives over 250 applications for the position, and must select 4-6 applications to schedule interviews. Hallie is committed to running a successful search that hires a qualified candidate. She is also committed to hiring in a way that is fair to all applicants. To aid her decision, she uses a resume filtering algorithm that looks for markers of success in education and previous employment. From a filtered pool of 15 applicants, Hallie selects six individuals to interview, and ultimately hires a candidate that performs well in the role.

Hallie runs a successful search, in line with her commitments. The candidate hired is competent, qualified, and performs well for the company. A question remains, however, concerning whether Hallie's commitment to running a *fair* search has been satisfied. With current resume screening technologies, there are good reasons to think it might not be. Resume

audit studies have shown that, for identical resumes, male, white-sounding names are significantly more likely to receive callbacks and interviews than those resumes associated with minority identities (see Gaddis 2018 for an introduction and overview). The widespread adoption of resume screening algorithms by large companies has not seen this trend abate. There are a number of reasons why statistical learning algorithms might exacerbate the hiring biases of humans (e.g. Desphande et al. 2020). It thus seems very likely that, in using a resume screening algorithm, Hallie has not met her own commitment to fairness. Though she hired someone qualified for the job, she might have left other, equally-qualified candidates on the table for no other reason than some feature of their resume was associated with poor performance by an algorithm.[16] Hallie's decision has not been fair. Given her sincere commitment to running a fair search, her decision was not authentic. It was not in line with her values. Were Hallie made aware of this disconnect, it would be rational for her to lament her decision-making and to avoid using such algorithmic aid in the future.[17]

So far, all this shows is that inauthentic algorithm-aided decision-making is possible. Many would have accepted this pretheoretically. Does algorithm-aided decision-making make inauthenticity *more likely*? While a full answer to this question requires the framework I develop in Section 4, we can already see one important element in the *opacity* (and ultimate *unanswerability*) of current machine-learning algorithms. Consider two further developments of the hiring case:

> **Hiring 2 (H2).** In this version, Hallie utilizes a proprietary resume filtering software to make her initial cut, just like in H1. She wonders, however, whether using the algorithm was ultimately the right call. She tries to query the algorithm itself about its processing, but there is no way to do so. She tries to contact the company who provides the

[16] The empirical picture is complicated, but seems to show this kind of bias in, for instance, resume search and ranking algorithms from job board websites (see Chen et al. 2018).
[17] Or, better: it *might* be rational for her to lament her decision-making. It remains a complicated question just what fairness requires in hiring, and how algorithms might undermine that fairness (Creel & Hellman 2022). The important point for us is merely that such a reaction is possible (and possibly justified).

algorithm, but they inform her that the internal functioning of their algorithm is a business secret. Lacking other options, and pressed for time, Hallie utilizes the algorithm in her decision-making.

Hiring 3 (H3). In this version of the case, Hallie does not utilize an algorithm at all. She hires an assistant to help her make the first cut of resumes. Looking over the assistant's work, she notices that the resumes the assistant has selected all have certain features (white-sounding names, degrees from elite private colleges, etc.). She queries the assistant about the features. The assistant informs her that they want to hire the "right kind of people" and have discarded resumes that do not fit this idea. Hallie, appalled by this behavior, refuses to use the assistant's recommendations in her hiring decisions.

H2 and H3 point to a fundamental disconnect between the kinds of contestability that exist when dealing with state-of-the-art machine learning algorithms and with more traditional decision aids (especially other agents). Procedures performed by human beings take place in a space of reasons-giving, where explanations and reasons can be asked for and received. Algorithms, on the other hand, need to be designed to be explainable, and often are not (Creel 2020; Rudin 2019). This is one way that Hallie's reliance on an algorithm is different than her reliance on her assistant. This difference is directly relevant to whether she can successfully act on her commitments.

Interestingly, reflections on hiring cases can also tell us something about the notion of authenticity. Compare H1-H3 with:

Hiring 4 (H4). This version is similar to H3, but Hallie's commitments are different. Hallie does not care about being fair in her hiring process. She only cares about picking a competent person for the job. When her assistant claims to only be looking for the "right kind of people," Hallie agrees with them. She uses this initial cut to aid her in hiring a candidate, who does in fact go on to do the job well.

There are many things we can criticize in H4. One thing we cannot criticize is any mismatch between Hallie's commitments and her unfair hiring practices. Hallie doesn't care about being fair in hiring. Since she hired a good person for the job, she is not worried about the candidates she has not given attention to. Cases like H4 push against the idea that there is something independently valuable about authenticity. Intuitively, there is nothing of value in Hallie's

decision-making in H4, despite the authentic match between discriminatory values and discriminatory information-processing.[18] While the decision may be authentic, it is unclear whether it has anything else to recommend it.

An initial lesson from reflecting on H1-H4: algorithm-aided decisions are more at risk of being inauthentic because they involve *offloading*. Faced with an overwhelming number of resumes and too little time to sort through them all, Hallie reasonably decides to utilize a resume filtering algorithm to undertake part of the decision-making process for her. In doing so, she cedes control of parts of her decision-making to the algorithm. Of course, we cede control of parts of our decision-making all the time. We ask experts and trusted friends for advice on the best course of action.[19] But we take our friends' advice precisely because they are accountable to us. They provide us with reasons for their advice. We can query them for clarifications or further explanation. If things go wrong, we can redress them for leading us astray. None of these forms of normative engagement are currently possible with machine learning and other algorithmic technologies. It is unclear whether algorithms are the right kinds of things that could be held accountable.[20] Hallie finds herself in a particular kind of normative bind. Either she attempts (perhaps *per impossible*) to sort through the resumes herself, or she is forced to use an algorithm that might undermine her commitment to running a fair search. This is the context in which the inauthenticity of algorithmically-aided decision-making looms large.

[18] This might be explained by a value externalism (along the lines of some kinds of objective list theories of welfare) that says one cannot successfully value what is not in fact valuable (c.f. Fletcher 2013).
[19] While the connection between advice-giving (e.g. Wiland 2021) and the possibility of trusting algorithms (e.g. Ferrario et al. 2020) is underexplored, it strikes me there is much interesting work to be done here.
[20] The kind of trust that Hallie exhibits is more akin to trust as an unquestioning attitude (Nguyen 2022). Hallie is open to what Nguyen calls "agential gullibility," trusting an algorithm to undertake a process that cannot reasonably guarantee an outcome that she cares about. The connection between Nguyen's account of trust and authentic decision-making is subtle and underexplored, though I just note it here due to space constraints.

Now return to the doctor who must make a decision about how to treat an incapacitated patient whose wishes for her own treatment are unknown. One possible way the case could develop:

> **Patient Preference Predictor 1 (PPP1).** Dahlia uses the patient preference predictor to learn what preferences her incapacitated patient might have, given facts about the patient's condition and demographics. The predictor predicts the patient would want an aggressive treatment with a high probability of saving her life. Dahlia has a deep commitment to respecting the rights of her patients, and she genuinely worries about making a decision with so little information. Dahlia nonetheless performs the procedure, which succeeds. The patient soon regains consciousness. Unfortunately, the patient informs Dahlia that the patient preference predictor was incorrect in this instance. She would rather have not faced a life of painful and expensive complications that will result from the procedure.

What might we say in our normative evaluation of PPP1? Though the patient preference predictor is controversial (Rid & Wendler 2014; Jardas et al. 2022; Mainz 2023), it seems plausible to me that Dahlia's patient has a reasonable claim that her autonomy has been violated. She had a wish (that she not be put through the difficulties of post-procedure life) that was not respected by Dahlia. This seems like a relatively straightforward case of violating her preferences. It might be a blameless autonomy violation. By stipulation, Dahlia had no other way to predict what her patient's preferences would be. A blameless autonomy violation is a violation nonetheless.

On the decision-maker's side, it is also plausible that Dahlia's decision is inauthentic. She has a deep commitment to respecting the wishes and autonomy of her patients, and this commitment has been violated. Of course, this is probably not the most worrisome feature of this situation. PPP1 demonstrates that, while authentic decision-making might be valuable, there will be cases where the fact that some decision is inauthentic is not the normative difference-maker in our evaluation of the case. But we also should not diminish the role of authenticity in this example. The disconnect between Dahlia's values and her actions might explain some of the remaining agent regret she might feel about the situation, even though she

might be blameless for her actions. Though it would have been unreasonable to expect her to do otherwise, her actions came apart radically from her commitments. It seems rational for her to feel some regret about this regardless of the difficulty of the decision. This is because, I claim, it is rational for Dahlia to care about her decision-making being authentic.

PPP1 might push us to take a stand on what kinds of matches between mind and world are sufficient to generate authentic decision-making. When can a certain fact, which the agent had no epistemic access to at the time of her decision, undermine the authenticity of that decision? A purely subjective account of authenticity would downplay this possibility. By Dahlia's own lights, she is doing as well as she can when she makes her decision. Given the opacity and complexity of the decision environment and the patient preference predictor, what else should she have done? Dahlia is not an engineer at the company that owns the preference predictor algorithm and its data. The generally positive media hype surrounding algorithmic technology might mean that Dahlia has no obvious way of knowing about possible problems with the predictor.[21] What else, this theorist might ask, could we want from an account of authenticity? This account seems to me to get something crucially wrong about the nature of authenticity, however. It is reasonable that Dahlia would feel lingering regret at her decision coming apart from her commitments and values, even if she had no way of knowing at the time about the mismatch. This suggests that epistemically unavailable facts about disconnects between commitments and actions can nonetheless influence whether a decision is seen as authentic by the deciding agent.

[21] Despite a growing skepticism of media techno-utopianism in the academic literature, much of the cycle of hype for technology marketing remains in place (Steinert & Leifer 2010). This is a plausible evidential situation for Dahlia to find herself in.

While these reflections seem on the right track to me, there are certain implications that deserve further scrutiny. Consider two versions of the case of the dejected graduate student. In one:

> **Career Selection 1 (CS1).** Georgia is trying to decide whether to hunker down and finish her dissertation chapter or start studying for the LSAT. Like most inquiring agents, Georgia has a commitment to gathering a wide and representative sample of information before making her decision.[22] She spends part of her study break Googling articles about the relative benefits of law school and philosophy academia. Unbeknownst to her, a cabal of law schools, interested in recruiting high-performing philosophy graduate students to their schools, have paid Alphabet to weigh more heavily in search results articles that argue philosophy graduate students should apply to law school. Georgia sees these articles and reads them. Partially on the basis of the information contained in them, she decides to start studying for the LSAT.

By the standards for authenticity we have established, Georgia's deception by the law school cabal undermines the authenticity of her decision. This can be true even if the information Georgia encounters during her search is completely accurate. Like many inquiring agents, Georgia's epistemic commitments include wanting to get a representative feel of the range of different positions on the issue. She does not merely want to be exposed to some accurate information during the course of her inquiry. Her commitments were not respected in this inquiry, and as a result, her decision is inauthentic. Moreover, that inauthenticity has an interesting feature. It is caused by the actions of other agents who use the opacity and complexity of Google's search results to present information in a way that might influence Georgia. This represents something like a nudge, though it is a nudge of a particular authenticity-denying variety.[23]

But opacity threatens to seep into cases where, intuitively, nothing is amiss:

> **Career Selection 2 (CS2).** Georgia is again deciding between law school and philosophy. She again uses Google to research some articles on the relative merits of the two career paths. This time, there is no cabal of law schools attempting to influence her decision.

[22] The rationality of inquiry is an area of epistemology that has grown considerably in recent years, and it is directly relevant to the evaluation of this case. See, e.g., Friedman (2020), Thorstad (2021), and Flores & Woodard (2023).

[23] This undermining can occur even if, as Levy (2019) has argued, nudges give agents reasons for acting.

> Instead, the recommender algorithm works as designed, predicting what information Georgia will find most useful in her search. It just so happens these articles are identical to those presented in CS1. After reading and considering them, Georgia starts to prepare for the LSAT.

Is there anything amiss about CS2? It is natural to draw a direct parallel with CS1. Georgia is presented with information that does not allow her to meet her commitment to collect a wide range of information before deciding on a course of action. This suggests that Georgia's decision is inauthentic in CS2. But is this the right result? It is plausible that, in many cases where we make mundane, (seemingly) normatively-innocent decisions with the aid of technologies like the Google search engine, we are in a situation like CS2.[24] Does this mean that something as innocent as Googling to look for information during an inquiry has the same normative standing as the shady nudges of CS1?

The answer, I want to say, is no. Answering this question will require an account of what features of the decision environment matter most for authentic algorithm-aided decisions. I will give such an account in the next section. To sum up the ground covered so far: there are a number of decision contexts where, when agents offload a portion of their decision-making to algorithms, they run the risk of making decisions that go against their stated values and commitments. Perhaps more disturbing from the agent's point of view, it will often be difficult (due to the nature of the technology) for the agent to know whether the algorithms they use are authenticity-undermining. This differs, in subtle but important ways, from the opacity that exists when one agent asks another for advice, or when the agent engages with traditional inquiry-related artifacts like texts. As these technologies become ever more present, the possibility of a loss of authenticity in decision-making becomes more pressing as well.

[24] This was, in fact, one of the main takeaways of the classic work on inquiry and rationality by Kelly (2003): if you have a train to catch, there is nothing irrational about only considering what you take to be the most relevant information for your inquiry.

4. Assessing and responding to the problem

4.1 Four dimensions of inauthenticity

Considering case types like Hiring, Patient Preference Predictor, and Career Selection allow us to get a sense of what features of a case make inauthentic decision-making more or less likely. In order to evaluate the solutions one might offer to the problem of inauthenticity, it will be helpful to have a taxonomy of these features. Here I consider four features (degree of algorithmic aid, opacity, problem space complexity, and end-user knowledge) that I think are important, though there are certainly others that might be relevant as well.

The first feature is a background condition on whether a decision counts as algorithmically-aided. The cases we have been discussing are ones where the algorithm occupies an intermediary position in helping us with our decision-making. Dahlia relies substantially on the patient preference predictor algorithm, but it is still her decision whether or not to give a treatment to her patient. She can use far more than just the algorithm in aiding that decision. Compare this with two more extreme cases. At the one extreme are the kinds of decisions discussed in section 2.1, where it is not plausible that the agent is using an algorithm of any kind to make a decision. The agent instead relies on run-of-the-mill aids like books or close friends to help her. At the other extreme are decisions that are so algorithm-dependent that they cease involving the agent in a substantive way. I have in mind versions of fast algorithmic trading programs (discussed more in section 4.2), where no human being could possibly make a decision in enough time to shape the algorithm's decision, except in a supervisory capacity. The decision is not obviously algorithmically *aided*, and concerns about authenticity are not as clear. If a trading algorithm makes a trade that an agent dislikes, it is not clear that the agent supervising the algorithm has done *anything*, let alone something authentic or inauthentic to her values and commitments. The cases we are interested in are middling ones.

Algorithmic aid is crucial to the decision-making procedure, but the decision is still up to the agent.

The second relevant feature has already been mentioned: the *opacity* of contemporary deep learning algorithms. Many decision algorithms are opaque in ways that directly impact how agents use them to make decisions (see Creel 2020 for a number of different distinctions). Many algorithms provide only a recommendation, delivering outputs to decision-makers without any information about how a decision was reached. There are also deeper computational opacities with how models with billions of parameters process and categorize information. It is often not clear to even the engineers of these models what the algorithm has done to produce an output. Further opacities in commercial algorithms stem from companies' desire to keep their intellectual property private, rendering opaque some algorithms that are not otherwise complicated in structure (Rudin 2019). Regarding opacity, it is clear that less is more from the point of view of authenticity. If a decision is made with an algorithm that is relatively opaque, it becomes more likely that a disconnect between the agent's values and the way the algorithm operates will arise. If Hallie were aware of the problematic functioning of resume filtering algorithms, she could adjust her decision-making accordingly. But because the algorithm is opaque to her inquiries, inauthenticity looms.

The third feature is another epistemic condition, though not necessarily one rooted in ignorance of the way an algorithm works. The *complexity* of the problem space is immediately relevant when assessing the authenticity of a resulting decision. Roughly speaking, the more complicated a decision is for an agent, the more likely she will need to rely on aid (algorithmic or otherwise) to make it. The more she must rely on the algorithm, the more likely it is that differences between her values and the value system encoded in the algorithm will be amplified. Complexity of a problem space is a feature, in and of itself, not particularly novel to algorithmic

decision-making. But complicated decisions based on large amounts of data are increasingly being required of agents who are not well-positioned to comb through the data themselves. In an epistemic environment filled with increasing access to data, but no corresponding increase in the cognitive capacities of agents, agents will rely on external aids (especially deep learning algorithms, whose abilities to sort through large amounts of data are some of their main selling points) to aid their decisions. When these aids pull against the values of the agents, inauthenticity looms.[25]

Finally, a number of epistemic conditions concerning the *background knowledge of the end user* are directly relevant to the possibility of an authentic decision. Roughly: the more an agent knows about the relevant technology, the more she can avoid cases of inauthenticity. There are different kinds of knowledge that might be relevant here. If Hallie were made aware of algorithmic audits (e.g. Raji et al. 2020), especially the poor audit fairness scores for some resume filtering algorithms, she could avoid or otherwise contextualize the output of the algorithm in her final decision. Second-order knowledge of algorithms and their effects is one of the main aims of the algorithmic literacy movement, discussed in section 4.4. But second-order knowledge of algorithms is only one kind of end-user knowledge. Given the complex sociotechnical nature of many of the biases and ethical issues with algorithms, it is just as likely

[25] This means that, in many actual cases, complexity of the problem space will feed directly into (or perhaps be reducible to) algorithmic reliance. While this is true practically, I think it is still helpful to separate out these conditions. Among other reasons, the complexity condition is needed to understand a prominent argument for interpretable AI (see section 4.3 below). By reducing the complexity space to one where linear variables can be interpreted and understood, interpretable AI does something preferable to merely explainable AI. I am grateful to an anonymous reviewer for pushing me on this point.

that a broad understanding of ethics and sociopolitical life will be important.[26] I return to these points in section 4.4.

Bringing these four conditions together allows us to see when inauthentic algorithm-aided decisions are likely to emerge: (a) complex decisions that are (b) moderately aided by (c) relatively opaque algorithms, where (d) little end-user knowledge can be expected. Having this framework in mind allows us to see the full problem of inauthenticity, as I see it. We are starting to be pushed, either by necessity (given the complexity of the problems we face) or by explicitly commercial interests, to use algorithms that subtly encode systems of valuing that they do not wear on their surface. How a resume filtering or search algorithm functions is not only unapparent to the vast majority of users. Its functioning is something that those who create the technology have good reason to want to obscure. To the extent that we continue to use these artifacts in complex and important decisions in our lives, we run the risk of undermining ourselves in an important way. We run the risk of making decisions that do not line up with our values and commitments because we do not, and could not, have known better. This is the problem of inauthenticity in algorithm-aided decision-making.

With this (admittedly informal) framework in hand, we are now ready to consider three proposed solutions for dealing with the problem of inauthenticity in algorithm-aided decisions-making.

[26] Just one example out of many: absent a complex understanding of a company's culture and the barriers to entry for high-status work in the first place, it might not be obvious to a decision-maker why relying on previous successes at a company as a model for hiring might be discriminatory. This points to another area where thinking about these cases might refine our theory of authenticity. What should we do in cases where (unbeknownst to the agent) two of their explicit values conflict? Suppose Hallie both values running a fair search and using track-record data at a previous company as a good proxy for future success. These values are clearly in tension with one another, and Hallie cannot act in ways that satisfy them both in one and the same action. Does this mean that neither action she takes is authentic? That either one is? Or that we need a theory of which explicit values *really* are fundamental to the agent and which are not? Future work on authenticity and algorithmic decision-making should focus on these questions (see section 5 below, especially 5.3).

4.2 Abstention

Perhaps the simplest response to the problem of inauthenticity, though one not always considered in a field hungry for technical solutions, recommends abstention. We should avoid using opaque algorithms in our decision-making when we can. If we do not use these technologies, we run no risk of the values they encode and propagate coming apart from our own values and commitments. Often arguments for abstention focus on the harm that algorithms cause, especially when easily-interpretable versions of the same algorithms are available (e.g. Du et al. 2019). My argument suggests another justification for abstention. We should avoid using opaque algorithms in our decision-making because they threaten something we value as agents, namely authenticity. The opacity and complexity of state-of-the-art deep learning algorithms makes their utilization too normatively costly.

What should we make of abstention? In some cases, abstention probably is required. Cases where we otherwise use machine learning algorithms frivolously and without obvious payoff, we should indeed abstain from using them. While this might sound like a trivial point about relative risk, some variations of Career Selection seem to recommend abstention. If Georgia is making a decision of fundamental importance that shapes all aspects of her future, why would she finalize that decision after only a cursory Google search? Why rate the results of her search so highly compared to the inputs of close friends and family? As other important aspects of the decision context are made more salient, the bar for how secure we have to be in our trust of algorithmic aid increases. It is not plausible that algorithmic aid is required in at least some of the areas we care about.

Abstention will be neither possible nor desirable in other decision contexts. Sometimes refraining from using some algorithmic aid is not realistic, given the complexity of the decision space and the limits of human thinkers. PPP1 has exactly this form, since (by stipulation of the

case) Dahlia lacks other ways of making a decision for her patient. Other, real-world cases come to mind. In fast-moving financial markets where hundreds of trades are made every second, the question is not whether to use an algorithm to make a decision, but which algorithm to use.[27] There are also harm reduction considerations for many algorithmic decisions. Given that many first-round cuts for hiring at large corporations use algorithmic aid of the kind sketched in H1, how best can we ensure a process that protects things we value? In these cases, abstention is practically impossible. We need mitigation strategies.

We also should not get too carried away thinking about the value of authenticity. To claim that authenticity is valuable is not to claim that authenticity should be the *overriding* value in our decisions. Modifying PPP1 slightly, consider a doctor who uses a diagnostic algorithm like Deep Patient (Miotto et al. 2016) to generate an initial hypothesis about a patient's ailment before doing extensive testing herself. The doctor will not know which factors, of the many in the patient's file, the algorithm used to come to a provisional diagnosis. Were she to merely accept the diagnosis, she runs the risk of making an inauthentic (and epistemically unjustified) judgment about the nature of the patient's ailment. But if she takes this output as a way of narrowing down possible initial hypotheses for diagnosis, she uses the algorithm in a way consistent with authentic choice. If a machine learning algorithm helps a semi-autonomous vehicle avoid colliding with a pedestrian, the ability to save the pedestrian's life outweighs any considerations of "authentic" decision-making on the part of the human driver. Abstention will neither be necessary nor desirable in these cases.

4.3 Explainable and Interpretable AI

[27] Whether such markets dominated by autonomous agents should exist in the first place is a separate issue (c.f. Wellman & Rajan 2017). Once they do exist, it makes no sense to avoid using machine learning algorithms to make trades.

A recent trend in the literature is directly relevant to the question of building authenticity-preserving algorithms. This research concerns methods for creating AI technologies that are more intelligible to human actors. Some advocate for *explainable AI*, artificial intelligence that delivers outputs explainable to, and thus intelligible to, end users and other interested human parties (Vilone & Longo 2020; Angelov et al. 2021). In a typical explainable AI system, a black box model is used as input data for a second model that learns how the first algorithm made a decision and explains that decision in an intelligible way. This work is often focused on the impacts an algorithmic decision has on members of the public. The European Union, for instance, has introduced a "right to an explanation" for those affected by algorithmic decisions, allowing them to ask for and receive explanations concerning why a decision was made (c.f. Kaminski 2019).

Another strand of research focuses on the related aim of *interpretable AI*, machine learning models that use simple and easy-to-interpret variables that competent agents can immediately read and interpret (Murdoch et al. 2019; Molnar et al. 2020). Advocates of interpretable AI are often skeptical of black box algorithms in important decision-making contexts. They point out that simpler models are often sufficient for making a particular decision. What unites these approaches is a commitment to the *intelligibility* of algorithmic outcomes as a key ethical and technical aim of AI research.

Work in explainable and interpretable AI can be extended to help decision-makers understand the nature of the algorithms that aid their decisions. As we saw, the opacity of deep learning algorithms is one of the main contributing factors to inauthentic decision-making. Reducing opacity will in turn reduce the likelihood of an agent unknowingly utilizing an algorithm in a way that conflicts with her commitments and values. Intelligible AI is good, not just because of how useful it might be for mitigating unjust outcomes, but also because it

protects the authenticity of algorithm-aided decisions. The argument for interpretable AI might actually be strongest when one thinks about authenticity. Merely being given an explanation for why some decision went against you is not itself sufficient to alleviate negative outcomes.[28] Being given information before making a decision is sufficient to ensure that the agent has the relevant knowledge to make a decision in line with her values. (Whether she ultimately does so is, of course, a separate issue.) Other things being equal, due to the extra uncertainty introduced by a second explainer model, it is better for authenticity for a model to be interpretable rather than explainable. Of course, other things are rarely equal. Interpretable models fail to perform nearly as well as many cutting-edge opaque deep learning algorithms. Many algorithms are also proprietary and thus not interpretable. In these cases, explainable AI will be the only game in town.

4.4 Second-order knowledge

As it stands, explainable and interpretable machine learning technologies are more of an ideal than a reality we can reliably exploit. For many algorithms, no explanation of their information processing is on offer. How can we make authentic decisions in these cases? There is a lot we can do to promote authentic decision-making with even the most opaque algorithms. The possibility of the current paper is a demonstration of this. Although we do not know the exact inner workings of many of the algorithms discussed here (and, indeed, although some of the algorithms are fictional and have no inner workings to know), we are able to discuss their inputs, their outputs, and their effects on our decisions. We have *second-order knowledge* about their antecedents and effects. In many cases, this will be enough to help us make authentic decisions.

[28] This is why some think we have, not a right to an explanation as such, but a "right to a better decision" in which explanations play, at best, an instrumental role (Edwards & Veale 2018).

In many cases, if decision-makers become aware of basic information about the nature and function of algorithms, it would be sufficient to allow for authentic decision-making. Providing Hallie with information about the fairness of resume filtering algorithms would have been sufficient to modify her decision-making in hiring for her company. Informing Georgia that the Google search algorithm is proprietary and open to unseen manipulation might have encouraged her to be more wide-ranging in her inquiries. None of these actors need to take courses in machine learning to understand that most algorithms are trained on biased data, can develop biases independently of their data, and are embedded in systems which can embed exploitative feedback loops. This knowledge is sufficient to temper reliance on the algorithm.

These considerations lead to an argument for algorithmic literacy to support authenticity.[29] As Aylsworth & Castro (2021) argue with regards to digital minimalism and social media use, how we use technology is not just an empirical, psychological question. It is intimately related to our capacity to act as practical agents. Learning about algorithms we use to make decisions is a requirement for acting rationally with them. To the extent that we have a duty to promote our own agency, we have a duty to learn about these algorithms. Of course, second-order knowledge can only go so far. Whether *this* algorithm utilizes information in a problematic way is hard to know from the general facts about functioning. A general understanding of the nature and function of algorithms that are similar to the one being used will allow the agent to approach with caution, or perhaps to abstain from use, but will not allow them to fully verify a match between algorithm and value.

Multiple approaches are thus necessary to ensure authentic decision-making in algorithmic contexts. A combination of (i) a willingness to use algorithms sparingly and cautiously, (ii) an expansion of algorithmic literacy, and (iii) a role for interpretable AI

[29] On algorithmic literacy and its relationship to digital literacy more broadly, see Oeldorf-Hirsch & Neubaum (2021).

(especially open-source algorithms that are not the proprietary software of a for-profit company) are all required to alleviate the threat of inauthenticity. Fortunately, research in these areas is ongoing. The purpose of this paper is to note that it does us a disservice to focus only on how this research might help make the *outcomes* of decisions more fair, just, and valuable. It can also help at the *input* level, making our algorithm-aided decisions more authentic.

5. Upshots and future directions

If everything I have argued so far is on the right track, we now have the basics of a framework for thinking about issues of authenticity in algorithm-aided decision-making. The problem with frameworks is that they are, by their nature, schematic. In this last substantive section, I will sketch a number of philosophical upshots and future directions for the study of authenticity in the age of deep learning. I make no claims that the avenues for future work I sketch here are anywhere near comprehensive. This section is instead meant to show how fruitful thinking about these issues in terms of authenticity might be.

5.1 Intersection with other theories of authenticity

In order to make the arguments presented in this paper manageable, I have focused on one particular account of authenticity: Brink (2003)'s conception of authenticity as a match between one's explicit values and one's practical actions. I focused on this account both because it is influential in the literature, and because it makes the basic problems of inauthenticity obvious in the cases discussed above. One might wonder, however, whether other accounts of authenticity will lead to similar results, or instead might complicate the judgments made above. While I cannot hope to do justice to the many different conceptions of authenticity in both the historical and analytic literature, I pursue one other theory here as proof of the flexibility of my approach.

Suppose one thinks that authenticity is tied in some important way to *self-fulfillment* (Vargas 2013; Rings 2017). This is another natural way to hear the commandment to be true to oneself. An agent is authentic, on this way of thinking, when she "develops her *life* on the basis of what is valuable to her" (Oshana 2007, p. 411, emphasis added). One can see the obvious parallels with Brink's definition of authenticity. But the upshots of self-fulfillment theory are quite different from Brink's. According to these theorists, authenticity is not about individual decisions and whether the dictates of practical rationality are in tension with one's explicit commitments. The target of authenticity is rather the agent's *life goals*, perhaps in the classic sense of Williams (1985). To understand whether an agent is making an authentic decision, we need much more information than what we get sketching an agent's mental state using Brink's method. For these thinkers, Brink also misses the fundamental normative character of authenticity. Authenticity is not (only) an ideal of practical rationality. It is also an *ethical* ideal, a kind of moral guidance that we can adopt in order to live better with one another (though how self-directed norms of authenticity are meant to do this is a matter of considerable disagreement).

This family of views is far away from Brink's account of authenticity. Yet they have interestingly similar things to say about cases we discussed above. The self-fulfillment theory is particularly salient in CS1 and CS2, where Georgia is making a decision about the overall shape of her professional life. Given our description of the case, it is plausible that an errant Google search is one way she can come to undermine a life plan that is in line with her values (as stipulated, a life of quiet contemplation is much more conducive to her self-fulfillment than one of high-priced lawyering). In this sense, Brink's theory and the self-fulfillment theory agree. Interestingly, however, the self-fulfillment view might push one to say there is *no real difference* between CS1 and CS2. In both cases, the Google search is leading Georgia away from fulfilling

her most basic life plans of quiet contemplation. A self-fulfillment theorist might not be able to say that there is a deep normative difference (rather than a difference of degree) between CS1 and CS2. A run-of-the-mill Google search might derail one's life plans just as much as one sponsored by a cabal of evil law school presidents.

The overall framework of authenticity in algorithm-aided decision-making allows us to make interesting choices about our theory of authenticity. If one is committed to the intuition that normal Google searching does not undermine authenticity, this is a reason to find the self-fulfillment theory of authenticity less plausible. If, on the other hand, even in CS2 one feels there's something fishy going on, the self-fulfillment theory allows us to see why CS2 is problematic. There are similar refinements that come from thinking about these views in parallel. For instance: it is plausible that the Patient Preference Predictor and Hiring cases less centrally involve the agent's overall life plans than the Career Selection cases.[30] The self-fulfillment theorist will likely say these cases are less central to our understanding of authenticity. If (as several readers and reviewers of this paper have testified) one finds Patient Preference Predictor and Hiring less authenticity-centered than Career Selection, this is again a reason to prefer the self-fulfillment view. If one thinks (as I do) that the cases form a tight core of related phenomena, this is another reason to doubt the self-fulfillment view.

There is much, much more one can say on these topics. But the basic point should be clear. Thinking of algorithmic decision-making in terms of authenticity allows us to both make judgments about specific kinds of decision-making, as well as refine (or perhaps jettison) different theories of authenticity in virtue of what they say about test cases. This is a significant upshot of the current framework, whatever theory of authenticity one has reason to adopt.

[30] Especially in Hiring, many people do not center their life plan on the particularities of some job they need in order to pay the bills. For these agents, whether the hiring they do is fair will have little bearing on the overall shape of their lives. The point is more complicated for Patient Preference Predictor, depending on whether performing well as a doctor is a central part of Dahlia's life plan.

5.2 Beyond autonomy

As discussed above, one of the core concerns in the literature on algorithm-aided decisions is the possibility that the recipient's or user's autonomy will be undermined (Rubel, Castro, & Pham 2021). If one thinks that algorithmic decisions are similar to nudges in important respects, then concerns about autonomy and nudging will be at the forefront of one's thinking on these issues as well. While debates about autonomy are interesting in their own right, the debate in that literature has reached something of an impasse. One benefit of the current framework, as I see it, is that it allows us to move beyond merely thinking about autonomy as the central value in debates about algorithm-aided decision-making.

This is true whether one thinks of authenticity as a species of autonomy (Schwan 2022) or as a value that might be in tension with autonomy (Rings 2017). There will be cases where the user's autonomy is respected while their authenticity will be violated. This is one way of interpreting the disconnect between CS1 and CS2, for instance. While controversial, it is plausible to me (and to Levy 2019) that nudges like CS1 or CS2 do not violate Georgia's autonomy in any obvious way. The pages she is reading give her genuine reasons for preferring law school to graduate school, and they do not inhibit her ability to develop her deliberative or rational capacities (Hausman & Welch 2010). From the perspective of basic autonomy, it is not clear that Georgia's autonomy has been violated in CS1 and CS2. And yet, at least in CS1, it seems like *something* has gone normatively astray. The value of framing Georgia's decision in terms of authenticity is that it allows us to explain what has gone wrong even if we do not think her autonomy has been violated. Georgia's autonomy has not been violated, but her ability to make authentic decisions has. Or, to put it in terms more amenable to Schwan (2022)'s theory of

authenticity: while autonomy in general has not been violated in CS2, the specific kind of autonomy that is authenticity has been.

This is a general virtue of the framework. It allows us to move past the stalemate of debates about autonomy, rationality, and nudges. The framework I argue for presents a wide-open theoretical space where we can investigate how thinking about authenticity changes our understanding of algorithm-aided decision-making more broadly. I have already argued for some results above. While one might think such an approach has shades of Luddism, I argued that abstention from algorithm use will only be necessary in certain specific cases. Answering other questions will allow us to see the panorama of value in this area. For instance: does violating someone's ability to make an authentic decision, in general, matter more or less than violating their autonomy? The answer to this question is not obvious. But the framework I have developed suggests a method for answering them, by considering a wider range of cases of algorithm-aided decision-making and consulting our intuitions about the authenticity of agents therein.

5.3 Beyond explicit values

One area ripe for further research concerns what happens to the authenticity framework when we relax the requirement (common to both Brink's practical action account and the self-fulfillment account) that authenticity involves a match between a subject's *explicit* preferences and values and their actions (or life plans). One might worry[31] that such a focus sets the bar significantly too high for authentic decision-making, given that the majority of our

[31] As several anonymous reviewers did worry. I am grateful to them for pushing me to think about this new direction for the framework, and for inspiring many of the avenues for future research sketched in this subsection.

values seem to be some interesting mishmash of explicit, implicit, and fully unconscious mental states. What would the framework look like if we relaxed this assumption?

Answering this question in detail is not possible here. But some areas for research can be noted. Some possible outcomes are cause for alarm. As users become more and more dependent on algorithms, it is possible that their conscious access to their own values might start to diminish. The worry is related to issues of deskilling in moral algorithm-aided decision-making (Vallor 2015), but with a distinctive agential flavor. If all I know about situation S is that I use algorithm A1 to come to a conclusion in S, is it right to say there are *any* values I have in that domain?[32] Perhaps the only thing I value in that case is A1 itself. Overreliance on algorithms might lead to an obliteration of my valuing in situations like S, at least in extreme cases. If, on the other hand, the whole sociotechnical system (including both myself and the algorithm) constitutes the values "I" have in this situation, then I start to take on the values of the algorithm the more I rely on it. To the extent that I do not (either explicitly or implicitly) value what the algorithm values, this makes it even *more* problematic to offload my deliberation to A1. In either case, an even more dire picture is painted of the relationship between (implicit) valuing and authentic algorithmic decision-making. Overreliance on algorithms might, in extreme cases, deskill agents to the point where it is not clear what their values in a situation even are (if they have them at all).

At the other extreme, one might think that a focus on explicit valuing unfairly stacks the deck against good cases of algorithm-aided decision-making. Of course, if we focus on my explicit and conscious values, it will be hard (perhaps even impossible) to see how I could legitimately offload my decision-making to algorithms without explicitly undermining my own authenticity (absent omniscience about the functioning of the algorithm). But what about cases

[32] While obviously extreme, this is the vision of future decision-making that many tech entrepreneurs and CEOs seem to think is desirable. See, for instance, Altman (2024).

where my implicit values are more in line with the way the algorithm functions? That is, what about cases where I might have implicit values that match the implicit values of the algorithm (Johnson 2021)? Perhaps then there are more opportunities for both harmony and disharmony. This is a related question to the authentic racist raised in H4, recast at the level of implicit valuing. Though much more work would need to be done on this point, I am inclined to import the authenticity externalism I defended in responding to H4 to this case as well. If the implicit values both I and the algorithm hold are good ones, then an authentic match between them might be possible (*modulo* concerns about whether my implicit values are properly mine; c.f. Zheng 2016). If, on the other hand, the implicit values I and the algorithm hold are bad ones, then even if the agent is being "implicitly authentic" in utilizing the algorithm to make their decision, that authenticity is not one that has any value. The general point remains: thinking this way significantly expands the playing field for authentic decision-making in ways that both enlarge and complicate our analysis of algorithmic aid in these contexts.

These three subsections have moved almost absurdly quickly. There are entire papers to be written about just some of the different moves sketched here. The point is not to defend any of them, nor even to defend particular ways of formulating the underlying issues. The point, rather, is significantly upstream. Thinking about authenticity in algorithm-aided decision-making is philosophically rich. It allows us to refine our theories of authenticity, and to notice ways those theories differ from related theories of (e.g.) autonomy. It gives us a way to sort through disparate cases and find value-theoretic currents that run between them. And it allows us to make sense of remaining intuitive worries about algorithm-aided decision-making, even when other normative concerns have been silenced. In short, thinking of algorithm-aided decision-making through the lens of authenticity is both theoretically powerful and philosophically useful.

6. Conclusion

In this paper, I have identified an undertheorized normative problem for decisions we make with the aid of algorithms: a threat to our authenticity as decision-makers. Algorithms often produce outputs with data and methods that clash with our values and commitments. Because of their opaque or proprietary nature, we are often not in a position to know this. We are practical agents whose agency intersects with advancing technologies. We need ways to make sure that our values and the values encoded in technological systems do not pull apart in objectionable ways. I have surveyed the possibilities, but much more work is needed. In the rush to make more ethical AI, we should also make AI that aids and amplifies our agency.[33]

References

Altman, S. (2024) The possibilities of AI. YouTube talk. https://www.youtube.com/watch?v=GLKoDkbS1Cg

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5), e1424.

Aylsworth, T., & Castro, C. (2021). Is There a Duty to Be a Digital Minimalist?. *Journal of Applied Philosophy*, 38(4), 662-673.

Ballarini, C. (2022). Epistemic Blame and the New Evil Demon Problem. *Philosophical Studies*, 1-31.

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and behavior*, 36(1), 21-40.

Brink, D. O. (2003). Prudence and authenticity: Intrapersonal conflicts of value. *The Philosophical Review*, 112(2), 215-245.

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy compass*, 14(10), e12625.

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

Cavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2020). Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?. *IEEE transactions on biometrics, behavior, and identity science*, 3(1), 101-111.

[33] My deepest thanks to Colin Allen, Anne Newman, Rob Reich, Michael Ball-Blakely, Anncy Thresher, Ting-an Lin, Henrik Kugelberg, Jon Vandenburgh, Valerie Soon, Diana Acosta-Navas, Benji Xie, Dan Kelly, Evan Westra, JP Messina, Javier Gomez-Lavin, and audiences at Stanford and Purdue for comments and conversation that greatly improved this paper.

Chen, L., Ma, R., Hannák, A., & Wilson, C. (2018). Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1-14).
Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science*, 87(4), 568-589.
Creel, K., & Hellman, D. (2022). The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy*, 52(1), 26-43.
Culkin, R., & Das, S. R. (2017). Machine learning in finance: the case of deep learning for option pricing. *Journal of Investment Management*, 15(4), 92-100.
Danaher, J. (2019). The rise of the robots and the crisis of moral patiency. *AI & Society*, 34(1), 129-136.
Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. In *IJCAI* (Vol. 17, pp. 4691-4697).
Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint* arXiv:2006.11371.
Deshpande, K. V., Pan, S., & Foulds, J. R. (2020, July). Mitigating demographic Bias in AI-based resume filtering. In *Adjunct publication of the 28th ACM conference on user modeling, adaptation and personalization* (pp. 268-275).
Di Nucci, E. (2013). Habits, nudges, and consent. *The American Journal of Bioethics,* 13(6), 27-29.
Dorst, K., Levinstein, B. A., Salow, B., Husic, B. E., & Fitelson, B. (2021). Deference Done Better. *Philosophical Perspectives*, 35(1), 99-150.
Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *41st International convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 0210-0215). IEEE.
Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, 63(1), 68-77.
Edwards, L., & Veale, M. (2018). Enslaving the algorithm: from a "right to an explanation" to a "right to better decisions"?. *IEEE Security & Privacy*, 16(3), 46-54.
Feldman, S. D., & Hazlett, A. (2013). Authenticity and Self-Knowledge. *Dialectica*, 67(2), 157-181.
Fletcher, G. (2013). A fresh start for the objective-list theory of well-being. *Utilitas*, 25(2), 206-220.
Flores, C., & Woodard, E. (2023). Epistemic norms on evidence-gathering. *Philosophical Studies*, 180(9), 2547-2571.
Friedman, J. (2020). The epistemic and the zetetic. *Philosophical review*, 129(4), 501-536.
Gaddis, S. M. (Ed.). (2018). *Audit studies: Behind the scenes with theory, method, and nuance* (Vol. 14). Springer.
Hare, R. M. (1965). *Freedom and reason*. OUP Oxford.
Hausman, D. M., & Welch, B. (2010). Debate: To nudge or not to nudge. *Journal of Political Philosophy*, 18(1), 123-136.
Jardas, E. J., Wasserman, D., & Wendler, D. (2022). Autonomy-based criticisms of the patient preference predictor. *Journal of Medical Ethics*, 48(5), 304-310.
Johnson, G. M. (2021). Algorithmic bias: on the implicit biases of social technology. *Synthese*, 198(10), 9941-9961.
Kaminski, M. E. (2019). The right to explanation, explained. *Berkeley Tech. LJ*, 34, 189.
Kelly, T. (2003). Epistemic rationality as instrumental rationality: A critique. *Philosophy and phenomenological research*, 66(3), 612-640.
Kelly, T. (2005). The epistemic significance of disagreement. *Oxford studies in epistemology*, 1(167-196).

Korsgaard, C. M. (1996). *The sources of normativity*. Cambridge University Press.
Langenkamp, M., Costa, A., & Cheung, C. (2020). Hiring fairly in the age of algorithms. *arXiv preprint* arXiv:2004.07132.
Levy, N. (2011). Enhancing authenticity. *Journal of Applied Philosophy*, 28(3), 308-318.
Levy, N. (2019). Nudge, nudge, wink, wink: Nudging is giving reasons. *Ergo* 6.
Mainz, J. T. (2023). The patient preference predictor and the objection from higher-order preferences. *Journal of Medical Ethics*, 49(3), 221-222.
McGrath, S. (2019). *Moral knowledge*. Oxford University Press.
Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys* (CSUR), 54(6), 1-35.
Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1), 1-10.
Molnar, C., Casalicchio, G., & Bischl, B. (2020, September). Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 417-431). Springer, Cham.
Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44), 22071-22080.
Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield Publishers.
Nguyen, C. T. (2019). Games and the art of agency. *Philosophical Review*, 128(4), 423-462.
Nguyen, C. T. (2022). Trust as an unquestioning attitude. In Gendler, Hawthorne, & Chung (Eds.). *Oxford studies in epistemology, volume 7*. Oxford University Press.
Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453.
Oeldorf-Hirsch, A., & Neubaum, G. (2021). What Do We Know about Algorithmic Literacy? The Status Quo and a Research Agenda for a Growing Field. *SoArXiv preprint.* https://doi.org/10.31235/osf.io/2fd4j
Oshana, M. (2007). Autonomy and the Question of Authenticity. *Social Theory and Practice*, 33(3), 411-429.
Parens, E. (2005). Authenticity and ambivalence: Toward understanding the enhancement debate. *Hastings Center Report*, 35(3), 34-41.
Pugh, J., Maslen, H., & Savulescu, J. (2017). Deep brain stimulation, authenticity and value. *Cambridge quarterly of healthcare ethics*, 26(4), 640-657.
Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33-44).
Rid, A., & Wendler, D. (2014). Use of a patient preference predictor to help make medical decisions for incapacitated patients. *Journal of Medicine and Philosophy*, 39(2), 104-129.
Rings, M. (2017). Authenticity, self-fulfillment, and self-acknowledgment. *The Journal of Value Inquiry*, 51, 475-489.
Rubel, A., Castro, C., & Pham, A. (2021). *Algorithms and Autonomy: The Ethics of Automated Decision Systems*. Cambridge University Press.
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215.
Schmidt, A. T. (2017). The power to nudge. *American Political Science Review,* 111(2), 404-417.

Schwan, B. (2022). Sovereignty, authenticity and the patient preference predictor. *Journal of Medical Ethics*, 48(5), 311-312.
Steinert, M., & Leifer, L. (2010). Scrutinizing Gartner's hype cycle approach. In *Picmet 2010 technology management for global economic growth* (pp. 1-13). IEEE.
Taylor, C. (1992). *The ethics of authenticity*. Harvard University Press.
Thorstad, D. (2021). Inquiry and the epistemic. *Philosophical Studies*, 178(9), 2913-2928.
Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11), 4793-4813.
Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy & Technology*, 28, 107-124.
Varga, S. (2013). *Authenticity as an ethical ideal*. Routledge.
Vilone, G., & Longo, L. (2020). Explainable artificial intelligence: a systematic review. *arXiv preprint* arXiv:2006.00093.
Watson, G. (1975). Free agency. *The Journal of Philosophy*, 72(8), 205-220.
Wellman, M. P., & Rajan, U. (2017). Ethical issues for autonomous trading agents. *Minds and Machines*, 27(4), 609-624.
Wiland, E. (2021). *Guided by voices: moral testimony, advice, and forging a 'we'*. Oxford University Press.
Williams, B. (1985). *Ethics and the Limits of Philosophy*. Routledge.
Zheng, R. (2016). Attributability, accountability, and implicit bias. In Brownstein & Saul (eds.), *Implicit Bias and Philosophy.* OUP.