Forthcoming in *Journal of Applied Philosophy*

Title:

Inscrutable processes: algorithms, agency, and divisions of deliberative labour

Author:

Marinus Ferreira

University College Dublin, School of Philosophy

Correspondence:

marinus.ferreira@ucd.ie

Acknowledgements:

Thank you to Rowland Stout, Molly Wade, and the autumn 2019 class of ‘PHIL 41510 - Ethics: Theory and Practice’ at UCD for their discussions, the editor and two anonymous reviewers for their help in improving this paper, and Fred Kroon, Glen Pettigrove, Lina Eriksson, and Bruno Verbeek for their extensive feedback on earlier work that turned into this paper.

Funding:

This work was done with the financial support of the Centre for Ethics in Public Life at University College Dublin.

Abstract:

As the use of algorithmic decision-making becomes more commonplace, so too does the worry that these algorithms are often inscrutable and our use of them is a threat to our agency. Since we do not understand why an inscrutable process recommends one option another, we lose our ability to judge whether the guidance is appropriate and are vulnerable to being led astray. In response, I claim that guidance being inscrutable does not automatically make its guidance inappropriate. This phenomenon is not restricted to algorithms, and there are many social processes which we should conform to but are similarly unable to judge for ourselves. I provide a framework for how we can depend on inscrutable processes by introducing a distinction between knowing conformity (where I understand what justifies the guidance) from mere conformity (where I merely do what I am told), and showing how mere conformity is often positively valuable, because it allows for extended processes that in turn allow us to accomplish more than we could on our own. This is in effect a division of deliberative labour, which I argue is something commonplace but not often recognised, of which algorithmic guidance can be an example.

I. The challenge of algorithmic decision-making to agency

We need to come to a responsible understanding about the use of big data methods in our lives, and there is a lot to come to grips with. There are concerns about privacy,[1] about people having their data harvested exploitatively,[2] about being subject to increasingly onerous surveillance regimes,[3] and so on. What I focus on is the worry that the widespread introduction of algorithmic decision-making into important domains threaten people's agency.[4] The threat is that these algorithms are inscrutable, meaning that you cannot explain why they come to one decision rather than another. This means we cannot depend on our understanding of the process involved to judge when it is not working as it should, or how to rectify it if it starts to fail, or how to assign accountability for its results.

Many complaints about the inscrutability of algorithms arise from the fact that their workings are often deliberately hidden, and thus we cannot have a say in their use. In this case, our inability to judge the algorithm is the result of a political choice to hide it from us, and is inscrutable to us in the familiar way of propriety information, state secrets, backroom dealings, and conspiracies.[5] These are serious problems, but not my focus here.

I want to give special attention to a different way in which algorithmic systems can be inscrutable. Sometimes their workings are of a kind that individuals are unlikely to be able to track or reconstruct, like deep-learning algorithms with dozens, thousands, or billions of factors.[6] Because of the sheer number of factors, and because the algorithm depends on extremely fine-grained manipulations of them, there is no real prospect of human oversight, since humans simply do not have the information-processing capacity to track that many parameters and judge their significance weighted against each other. This inscrutability is not a political choice but results from the differences between human reasoning and algorithmic processes.[7] In this case, while we can know important things about the algorithm in broad outline (operational factors like the learning process it implements, the number of parameters it tracks, the number of layers it implements, etc.) we cannot in any real sense explain in particular cases how it comes to one decision rather than another. As such, we need to seriously consider how to respond to guidance from sources there is no realistic prospect of us fully understanding.

Here I argue that guidance from inscrutable sources is not automatically inappropriate. For my purposes, I will take it as read that the process is successful at guiding us to the ends we use the process for, though of course many processes will fail to do so. While it is not the focus of this piece, my view is that we need to judge inscrutable processes directly on their outcomes, by seeing whether it does reach the ends we are hoping to achieve with it. I do not think any other measure is available.

My argument hinges on the fact that the guidance may be the result of deliberation of someone or something within a larger system of which I am a part. By following guidance I receive from the process I can receive the benefits of the process, and be involved in its maintenance, without being able to myself reproduce the deliberation that results in that guidance. I call this a *division of deliberative labour*. Like all distributions of labour, it can be equitable or not, responsive to my changing needs or not, and so on, but it allows those who engage in it to accomplish more than they could as individuals. As such, I set out to show how these processes can act as scaffolds for individual agency even if we do not understand how they work. I use algorithmic guidance as a prominent example of this framework in action, but I also highlight how the point generalises to any division of deliberative labour.

I introduce divisions of deliberative labour and the comparison between algorithmic guidance and reliance on socially distributed deliberation in §II. I describe how divisions of deliberative labour are possible in §III. and §IV. , with the fullest example of the view in §IV. ii. After placing the view in context in §V. , I canvas the shortcomings of relying on processes we do not understand in §VI. , and in §VII. describe how nonetheless relying on inscrutable guidance does not undermine agency.

II. How divisions of deliberative labour address the challenge

An important and easy to miss point is that our vulnerability to having our agency reduced by algorithmic decision-making is similar to our vulnerability to having our agency reduced by social decision-making processes. Humans being political animals, it is ubiquitous to have important parts of our lives take place within a social context such that many operative decisions concerning me are not made solely by me, but through some social decision-making process. It can be as simple as some other person having the authority to issue orders concerning me (such as a bank manager deciding whether to accept my application for a loan), or be the effect of delegated authority (such as my having a say in representative government through my local member of parliament), or of a voting procedure (like a ballot on my union starting industrial action), or be the effect of concerted social action (such as standards of care set by health professionals), and so on. In each of these cases there is a division of deliberative labour.[8]

Having a division of deliberative labour is often unavoidable, because the number of factors and moving parts involved means that not everyone can effectively intervene in the process at every point. Like all distributions of labour, the capacities of groups which use them are likely to far outstrip those of groups in which each individual engages only in self-contained tasks. This process can be equitable or not, responsive to my changing needs or not, and so on. But by the same token, distributions of labour are liable to alienate participants by not giving them opportunity to engage with it as a whole. I address this in §VII.

There is important existing work about how algorithmic decision-making slots into our existing deliberative mechanisms and structures of accountability. Luciano Floridi introduces ‘distributed morality’, where responsibility for a process is spread across many different agents through links typical in information systems.[9] Kirsten Martin suggests that we incorporate algorithms into our decision-making systems and treat them the way we would other people in a committee, including the fact that people do not always share their reasons and they sometimes make mistakes.[10] Their suggestions are of obvious import to mine, but I am tackling a different part of the same problem. What I am doing here is not tracking attributions of responsibility (as Floridi does) or describing the relationship that algorithms have towards human reasoners (as Martin does), but instead describing how people slot into overarching processes which they are not in a position to understand themselves. I engage here in the philosophy of action of how individuals act within divisions of deliberative labour. As such Floridi, Martin, and I, can make useful and usefully different contributions to the same subject matter.

To give substance to the comparison between depending on algorithms and depending on social processes, consider the literature in legal theory about how one court may defer to the decisions of another on an issue it is considering.[11] In that case, the deferrer makes the positive decision not to relitigate (literally) an issue, because it is happy to take that issue as settled by the deliberation of another court. Usually, the deferrer will explicitly note the relevant decision by the deferree, meaning that it is not inscrutable. But alongside this usual case of deference, Yoav Dotan identifies what he calls ‘avoidance deference’, where the deferrer specifically avoids engaging with the reasoning of the deferree and simply takes the result of the deferree’s deliberation as read.[12] Dotan specifically notes that one reason for this may be that the deferrer may not feel it has the expertise to judge the matter themselves. This is much like the situation we find ourselves in with inscrutable algorithms. Furthermore, when it comes to the question of where and how a court may decide to stop deferring, pragmatic issues like there being too large a distance between the deferrer’s judgement and the views of the deferree; in this respect the to-and-fro between courts looks like that between human and algorithmic deliberations as described by Martin.

One important respect in which reliance on algorithms is different from reliance on the judgements of other people is that (at least in principle) we can enter a discussion with the other people, or have their reasoning be presented to us in some process of judging its appropriateness. This prospect of reason-giving comes in degrees, of course, but inscrutable algorithms are an example of the limiting case where entering into a process of reason-giving is not at all feasible.

It may be objected that placing algorithms outside of our practices of reason-giving is too stark, because the authors of the algorithms have at least some insight in their workings, and they can enter into these practices of reason-giving. But that is too optimistic, because for very many of these algorithms (prominently including deep-learning algorithms) they are designed to change how they work in response to ongoing interaction with data, and that training process causes the working of the algorithm to be inscrutable. While the training process has been authored, the resulting algorithm has not. This is in addition to the challenge to understandability offered by the sheer amount and variety of factors which these algorithms use as inputs.[13]

In turn, authors of algorithmic decision-making sometimes claim that the application of the algorithm is blameless, exactly because its workings are not the product of its authors but instead is guided by the data it is trained on. But this obscures the design decisions which are made by the authors, decisions like which data sets to train it on, what criteria to make it pursue, and what role in the division of deliberative labour the algorithm is to play. The outputs of the algorithm may not be authored, but those features are, and both the authors and the clients of the algorithms can be held to account for them.[14] For instance, online platforms ask us to tag or categorise images, but do not volunteer that this is to provide high-quality training data for their proprietary information systems.[15] Here not explaining the task is a conscious decision on the part of those implementing the information system, and is open to criticism as a political source of inscrutability, as discussed in §I.

We should not immediately reject inscrutable guidance as inappropriate, because there are many cases where people who depend on guidance from a division of deliberative labour never gain the expertise needed to scrutinise the deliberation involved. Consider the guidance I receive from medical experts about my health. I could in principle get to know why the guidance they give me is appropriate by undergoing the training required and engaging with the body of expertise myself. But in a very real sense this is an option only in principle because the time and effort it would require is prohibitive. This is a large part of why healthcare involves a distribution of labour (deliberative and otherwise). Nonetheless, the appropriateness of the guidance is unaffected.

The fact that in the social case there often are expert bodies who act as oversight for divisions of deliberative labour is insufficient grounds for insisting that there should always be someone put in place to scrutinise such deliberation. The prospect for scrutiny comes in degrees: medical bodies have higher openness to scrutiny often as a matter of law and because of the large amount of existing experts who can intelligently comment on them; the uppermost management of private firms have lower openness, and even less if they are not publicly traded and need not report to shareholders; deliberation behind closed doors is noted for its especially low openness to scrutiny. Algorithms are on the same spectrum, and at least some of them are at the extreme end of least openness to scrutiny. Nonetheless, the standing of a division of deliberative labour is unaffected by its degree of openness to scrutiny. When a clique of high-ranking party officials uses established methods to depose the head of government, they are playing their role in society's distribution of political deliberation to no lesser extent and with no lesser import than any other political decision.

And so too for algorithmic guidance.

III. What we learn from action-guidance

This section concerns action-guidance in general, that is, when we are directed to perform some particular action, be it by a command, request, recommendation, or anything similar. Much of the unease around the use of algorithmic decision-making is because we become disconnected from the reasoning that produces the action-guidance, and this in turn means that we are likely to lose track of what is and is not justified. Here I discuss how commonplace this disconnect is, and that it is not limited to algorithmic decision-making.

I want to distinguish between two different standards of knowledge we can have: *first-order knowledge* and *higher-order knowledge*. These are knowledge of, respectively, the first-order and the higher-order features of some piece of action-guidance. The first-order features are those that occur directly as the objects of that action-guidance. For instance, for the action-guidance ‘you should close the door’, the first-order features are the door that should be closed, that you close it by swinging it on its hinges, that it counts as closed when the latch has engaged with the door frame, and so on. More precisely, consider what is sometimes called the ‘satisfaction conditions’ of an imperative, the state of affairs that would count as having followed the imperative.[16] For instance, for the action-guidance ‘close the door’ the satisfaction conditions are captured in the proposition ‘the door is closed’. Let us call it *first-order success* when the satisfaction conditions are met; thus, first-order knowledge is what you need to know to garner first-order success. In turn, first-order ignorance is when you do not know what these features are, meaning you do not know the satisfaction conditions of the action-guidance. An example would be if you do not know which door to close, or what counts as closing the door due to its unfamiliar latch system.

In contrast, the higher-order features are those that are not themselves objects of the action-guidance, but instead have the action-guidance as its object. It will include things like the reason you want the door closed, why you expect the other party to do as you ask, the reasoning which resulted in you issuing the action-guidance, and so on. For instance, the fact that closing the door will lead to the room being less draughty would be a higher-order feature. It is not a first-order feature because someone can follow the action-guidance successfully and it nonetheless not make the room less draughty, like if the door is not effective at stopping the draft or you mistook what causes the draft. Higher-order knowledge is to have cognitive access to these higher-order features. Not knowing this would be an example of higher-order ignorance, as in the comical situation where I close the door and then open a window in order for there to be a breeze.

IV. How to succeed when you only know a little

Our focus here is on how to succeed at implementing a process even if we know little about it, or in my terms, of how it is possible to have first-order success despite higher-order ignorance. I will introduce three more terms, two that distinguish two different ways of reaching first-order success, and one more to explain how that is possible. There are two different states we could be in when we succeed at doing what we are told to: *knowing conformity*, when you have both the relevant first- and higher-order knowledge; and *mere conformity,* when you only have the relevant first-order knowledge. This is meant to map on to the difference between following guidance you understand, and merely following guidance without knowing why.[17] My claim is that when we receive inscrutable guidance, like that we receive from the most complex algorithms, we are mere conformers to that guidance. I offer a qualified defence of why this is a positive good for us in §VII. ii. VII. i.

i. The Alternative Method Model for success in inscrutable processes

Divisions of deliberative labour require mere conformity because many people in such a distribution are removed from the deliberation and as such often not equipped to knowingly conform to the guidance given. To explain how mere conformity it possible, I offer what I call the *alternative method model* (AMM for short): in some cases there are multiple procedures that would suffice for attaining first-order success, and in those cases knowledge of one method allows ignorance about a different one. The point of this is that someone can attain first-order success through one method, but lack any grasp on why that end is the one they should be aiming for, because that justification is tracked by a different method. In the social case, the alternative method is where we depend on the judgement of others in a division of deliberative labour, without trying to recreate their reasoning ourselves. In the algorithmic case, it is following the guidance of algorithmic decision-making even when the workings of the algorithm are inscrutable.

Consider how a child may have two different ways to know how to avoid being burnt by the stove: by one method, the child appreciates how hot the stove can get and that the stove will damage their skin and flesh when heated; by the other, the child knows that their parent has told them to avoid the stove. First-order success in this case is not getting burnt: avoiding the stove when it is hot, being careful not to touch it, and so on. But what the child has in mind in these two cases is very different, since the child who understands the dangers of the stove will respond directly to those dangers, whereas the child who responds only to the parents' warning will be responding to features of the parents, not to the dangers of the stove. The first way would be knowing conforming, understanding the danger that is being avoided by avoiding the stove, and the second way is mere conforming, where the action-guidance is followed without this understanding.

The AMM has four components. First, there is some given practice that we want to succeed. Because this paper looks at whether the inscrutability of a process makes it automatically inappropriate, we will take as given that the processes in question are worthwhile and their ends justifiable, and consider whether ignorance about how the process is working threatens that justification. In algorithmic cases the practices are things like medical diagnoses or insurance provision, but something as workaday as ‘be safe in the kitchen’ is a perfectly good example.

Second, there is the end that is recommended by that framework within some given situation, like not getting burnt on the hot stove. This is the standard of first-order success we have used thus far. The first and second components stand in a process-product relationship: the process of maintaining kitchen safety (first component) has the product of you avoiding getting burnt (second component).

Third, there is the action-guidance that tells us how to attain success in that practice in this situation in a way that explains why it is likely to be successful. In our example, this is ‘avoid a burn by not touching the hot stove’. Call this the *privileged method*. It is privileged because it is the method that captures why first-order success at the given task is worth having. It is what you know if you are a knowing conformer.

Fourth and finally, there may very well be an *alternative method*, such that conforming to that method also reaches first-order success, but the action-guidance highlights different features than those picked out by the privileged method. In our example, the alternative method would be for the child to obey the parent in this case, following the guidance ‘do not touch the stove because mommy and daddy says so’. Notice that this action-guidance makes no reference to heat or burns. Nonetheless, given what the world is like, avoiding touching the stove is also to avoid injury—first-order success.

The AMM allows for first-order success despite higher-order ignorance because we can learn to do something in way different from the way that highlights why the action in question is appropriate. The claim is not merely that there are multiple good reasons to do something. The claim is that we can follow methods which do not come along with justifications, as long as following it reaches the same end as the privileged method. This means that the alternative method may not on its own count as sufficient explanation for why to do something; it may even appear frivolous. And the alternative method is not sufficient on its own, but acts as an enabler for the process reaching the target end.

The suitability of the alternative method rises and falls with the extent that it guides us to the target end, even if it obscures or outright misrepresents why we act as we should. Many children grow to rebel against the advice of their parents, and they are right insofar as ‘my parents told me to’ is not a sufficient explanation for why they should do many of the things they are told. But, when parenting works, it is not the mere fact that the parents say so which makes their advice appropriate, but because their advice responds to the needs of the child and helps their long-term welfare.[18] I endorse mere conformity to inscrutable guidance on the same grounds, and with the same caveat: we should conform to it if it enables a worthwhile end, even when the outcomes produced by adhering to the guidance is all we have to judge it by. The claim is that guidance having an inscrutable source does not automatically make it inappropriate.

The above does not mean that the AMM will endorse any method which can arrive at the target end, including ones that are bad for other reasons. What matters is whether the alternative method systematically reaches the end required for the process to continue, so merely accidental success is not good enough, nor is a method which undermines the end in some other way.

There may be contexts where even highly reliable processes are not enough, but we need the participants to have genuine knowledge, as we aim for in education and it has been seriously argued we need for legal judgements.[19] In those cases, mere conformity will not do, because the end of the process is not just to keep the process ticking over, but that the participants gain understanding of the process. So, mere conformity would not suffice for reaching the end in question, because the end is knowing conformity among participants. The AMM cannot operate here. Nonetheless, the AMM will work in most contexts. In

Section VII I discuss why it is counterproductive to expect knowing conformity as a general requirement.

ii. Algorithmic guidance and the Alternative Method Model

Let us now look at the algorithmic case through the lens of the AMM. Consider the algorithmic guidance that a bank manager may receive about what risk category a business loan application belongs to. One important use of algorithmic lending is evaluating loan applications from individuals who do not have enough of a financial record to allow for a traditional credit score. These algorithms typically harness big data methods to evaluate a wider range of factors than traditional methods can manage.[20]. In this case the process is the institution of bank loans to support business ventures. The privileged method would be establishing that the applicant has sufficient prospects for being able to pay back the loan to make supporting the business profitable for the bank and thus mutually beneficial. The action-guidance that is required to maintain this process is the recommendation to approve loans from businesses sufficiently likely to succeed and deny the others, usually done through placing the applications in the appropriate risk category. The alternative method for doing so is referring to an algorithm which categorises applications based on their similarity to examples in the training data who were known either to succeed or fail to repay their loan. For instance, it may pick up that one factor in whether a small business is likely to have enough cashflow to repay a loan is the customer feedback it receives.[21] This and many other such features are typically too cumbersome and indistinct for humans to do much with, but are grist for the algorithmic mill. If the algorithm manages to make the right categorisations often enough, then the process can continue, and the participants (banks and businesses both) can use it as a scaffold for their agency. On the AMM, this will work even if neither the bank manager nor anyone else can explain why the algorithm makes the categorisations that it does.

Note that this view is neutral about whether the algorithm itself implements the privileged method. What matters for the AMM is whether the method allows the process to continue through enabling its participants to succeed at the target end. It may be that it is good at judging similarity to other cases, and this is itself only an alternative method for judging the prospects of an applicant to pay back a loan. This kind of worry is frequently voiced around algorithmic methods in natural language processing.[22] But in my terms that is a debate about whether natural language processing algorithms would count as knowing conformers to the rules of a language. When the proponents of these algorithms appeal to what has been called the 'unreasonable effectiveness of data', they are highlighting that these models succeed at and maintain linguistic practices such as translation.[23] This is enough for these algorithms to count as at least mere conformers to the rules of a language.

So, if there is a worthwhile process with multiple efficacious methods for engaging with it, an agent can play their part through following an alternative method, even if that means they do not understand what it is that makes the process they are engaging in worth pursuing. This is how divisions of deliberative labour are kept in place.

V. Placing the view in context

The discussion of having and gaining knowledge throughout this paper may give the mistaken impression that I am endorsing reliabilism in epistemology, because both reliabilists and I endorse processes based on whether they succeed at reaching the target end to some sufficient extent, even if we do not understand that process. But my view is distinct from reliabilism and rises or falls independently from it. This is because knowledge of the process is the starting point and not the conclusion for my account, since the end of action-guidance is action, not knowledge.

If we were to adopt reliabilism, that would not get us to my view. Consider again the bank manager example from §IV. ii. Let us grant that the process of relying on algorithmic guidance about bank loans meets a reliabilist standard, and thus that the bank manager is justified in believing, say, that the applicant in question is a sub-prime candidate likely to be able to repay the loan. This knowledge does not obviate the threat to agency we are considering. Even if we go from this directly to the action-guidance to approve the loan, this guidance is vulnerable to exactly the same problems that inscrutable processes are meant to have: we cannot tell in which circumstances the process will lead us astray, nor how to fix it when it does, nor do we have avenues to hold people accountable for failures in the processes. These simply do not feature in the goods that reliabilism tries to secure.[24]

Similarly, my approach does not appeal to consequentialism for justification. Consequentialism is not the only normative theory which allows us to judge processes directly on the appropriateness of its outcomes. A prominent example of this is the anti-consequentialist 'difference principle' of Rawls. In that case, a social order is judged on whether the people most benefitted from it are those who are worst off, which is the judgement that a process is appropriate or not because of who it targets. Rawls's reasons for adopting the difference principle does not depend on evaluating the discursive reasons that actually lead to the establishment of that order; seeing who it targets and how is enough. So, consequentialism has no special privilege for judging processes by their outcomes, and accordingly has no special relevance here.

Another avenue we may have looked to but which does not help us is the recent explosion of interest in group agency and collective intentions. This is because there is a lacuna in this literature around how to handle social phenomena that do not feature in the intentions of any individual. To illustrate this point, consider how this lacuna occurs on both sides of the most prominent divide in that literature, between collectivists who hold social actions to involve intentions that are shared among groups, and individualists who believe that these can be reduced to combinations of individual intentions. Cases of mere conformity to a division of deliberative labour is not a we-intention (in the terms of Tuomela) [25] nor a joint commitment (as Gilbert has it)[26] shared by humans and algorithms in the process: on the machine's side because it does not have the relevant type of psychology at all; and on the human side because the cases being considered here are exactly those where the individual does not understand enough to form this kind of attitude about it. 'I will go along with what the machine recommends' is not a collective action, no more than 'I will go gardening if the sun is out' is a joint commitment made with the weather. And on the individualist side, we cannot have the so-called 'interdependence of individual plans' that is highlighted by Miller and Bratman,[27] because the machine does not have plans at all in the relevant sense, and because in many cases the humans do not know enough about what the machine is doing to make the necessary predictions and allowances.

While there is an increasing recognition in this literature that we need to consider inanimate objects as part of the constitution of groups,[28] we still need to come to grips with what happens when these objects are not just resources to be harnessed, but themselves play a role in deliberation. That is part of what I am doing here.

I mention algorithms above, because that is my central example, but the lacuna exists even when it is only humans involved in a division of deliberative labour. That is because on the human side of the equation the problem is that the participants do not know enough about the processes to intelligibly engage with it any further than to conform to its first-order action-guidance. This applies both in the algorithmic and the social deliberation cases.[29]

VI. Acknowledging the shortcomings of mere conformity

In this section I lay out an important problem with depending on processes we do not understand for guidance, and then argue in §VI. VII. that despite being a second-best way to engage in a process, guidance from an inscrutable source is not automatically inappropriate and it would be counterproductive to insist on knowing conformity.

One problem is that higher-order ignorance undermines your ability to extrapolate from cases where you do have guidance to ones where you do not. If you lack a certain piece of pertinent higher-order knowledge, then you cannot use it to inform your other decisions. In general, as the extent of your higher-order ignorance grows, your ability to make good extrapolations diminishes. For instance, a bank manager may notice that good customer reviews is one factor that the algorithm considers when judging whether to approve a business loan, but without being able to track how this feature interacts with the others the algorithms considers, would not be equipped to use this partial insight, because they do not know how much weight to assign to it. The less the manager understands about the algorithmic process as a whole, the less good their extrapolations are likely to be.

Someone who suffers from higher-order error, and not just ignorance, suffers from a worse version of this inability to extrapolate correctly, because insofar as they assent to false beliefs about higher-order features of some instance of action-guidance, they will assent to extrapolations based on these false beliefs. For instance, if the bank manager believes that customer reviews are the determining factor (whereas it is likely to be one of dozens of factors, all weighed against each other dynamically in a deep-learning algorithm), then the manager is likely to come to over-hasty conclusions about how profitable it would be to lend to the applicant, based on what he can see about their customer feedback, not sufficiently weighing the import of the many other factors in play. Someone who is just ignorant struggles to make extrapolations, faulty and correct alike; someone who suffers from higher-order error will be liable to make faulty extrapolations.

Many of the worries about the use of algorithmic decision-making are worries about whether depending on these algorithms is to fall victim to faulty extrapolations. A simple example is how some algorithms for diagnosing skin cancer have been found to spuriously correlate the cancerousness of a lesion with whether a photo of the lesion includes a ruler, because rulers are more often placed next to lesions that turn out to be cancerous.[30] A more intricate example is that one of the worries about predictive policing systems is that they often are extrapolating from 'dirty data' which reflects unjust police practices. A much-discussed example is where the excessive stop-and-search practices targeted at minority populations means there is a preponderance of cases where members of these populations are found with contraband (though the rate they found to carry contraband is proportionally no higher than for populations not targeted in this way). Since there is a disproportionate focus on stopping-and-searching members of minority populations, there will be a disproportionate number of cases contraband found in minority populations in the training data. Since the algorithm is asked to predict where to find contraband, and it often sees it being find in minority populations, its recommendations turn *de facto* official bias against minority populations as a *de jure* recommendation, thereby magnify existing biases.[31]

VII. Why mere conformity does not undermine agency

The worries surveyed above make clear that mere conformity is second-best. Nonetheless, allowing for mere conformity is a positive good, and not allowing for it is a positive harm. The reason why we should stick with mere conformity is that it does not create higher-order ignorance, but instead allows practices to persist in the face of it. We must not confuse something forestalling the harms that arise with some threat as inviting this threat—this would be blaming the remedy rather than the malady.

It is tempting to think about processes involving mere conformity as involving a trade-off, but that is not quite right. If it were a trade off, I would give something up when I move from merely conforming to something to knowingly conforming to it, as all of us do as we come to understand more of the world around us. What instead happens is that there is one range of practices that knowingly conform to, and another range of practices we merely conformity to. Rather than being the product of trade-offs, divisions of deliberative labour are scaffolds that allow for individuals to accomplish things they could not accomplish on their own. That is what we turn to now.

i. Mere conformity is a positive good

Here I argue that mere conformity is a positive good, because when the existence of mere conformity makes a difference, there is no alternative scenario where the AMM is not in effect and the community is better off.

One set of scenarios is where there is no mere conformity, and thus no alternative method in place. In the absence of the AMM, the only ends that the individuals can achieve are ones that they are able to reason towards on their own power. Thus, everybody is a knowing conformer, and everybody makes use of the privileged method. So, these possibilities say nothing about the effects of the AMM.

The other relevant scenarios are where mere conforming makes a difference in what processes people engage in. This difference can be in one of two ways: people conform to a good process which has a target end as its first-order success, or conform to some malformed process which does not have such an end. If mere conforming leads them to engage in malformed processes, then that is undoubtedly a harm. But people in this scenario are not equipped to do any better without mere conformity. Their behaviour may be different than those of people who fail to reach their target end without engaging in a malformed process, but the result is the same: failure.

The third scenario is where the AMM makes a difference by empowering mere conformers to it to achieve the target end. The scenario without the AMM has the individuals achieve the ends that their higher-order knowledge equips them to; in the scenario with the AMM they achieve those ends and then the further ones that the AMM allows first-order success despite higher-order ignorance. This means that in a situation where the AMM is available, to insist on only your own higher-order knowledge is to choose failure over success.

Someone may object to equating the outcomes of mere conformity to some malformed process and the outcomes of not acting due to ignorance. It is easy to imagine situations where conformity to something like a faulty extrapolation as in §VI. leads to a worse outcome than would result from a process breaking down. But in response, it is equally easy to imagine the reverse. For instance, the fact that there is an existing process means that we already have an avenue for improvement available, by reforming the malformed process. We have nothing but speculation for deciding whether a specific bad outcome is worse or better for having a (malformed) process in place. The possible outcomes of acting without a way of securing target ends are largely unconstrained, and there just is not much to be said about them as a class.

Furthermore, we should resist the temptation to think that not acting does not have an outcome: for people to not conform to anything in particular is a course of action with its own outcomes just like any other. When you lack a reliable method of attaining a target end, there is no telling where you will end up, and it is unlikely to go well. Accordingly, while I cannot claim that any failure to reach a target end is alike, I can claim that there is not enough to go on to find systematic differences between various ways of being adrift.

Someone may object that the issue is not how the AMM influences actions when it secures first-order success, but the general effect of a community depending on processes they do not understand. This is an intelligible concern, but it makes no difference. The alternative to accepting mere conformity is to not allow it, which means that individuals are then restricted to only being able to attain target ends that they can secure through their own higher-order knowledge. And now we get to the same situation as in scenario one above: refusals to make use of the AMM stops members of a community from attaining ends that the AMM would enable them to secure.

ii. Allowing for mere conformity makes action-guidance more robust

Mere conformity makes the processes involved more robust. It does so for two distinct reasons, one pertaining to individuals, the second regarding the social dimension of action within a community.

The first reason is that there are fewer requirements for an individual playing their part in a process, because they are not required to have the higher-order knowledge that the privileged method requires. Consider a spectrum of higher-order ignorance, which ranges from someone suffering from total higher-order ignorance at one end to someone who is a moral sage and has perfect knowledge of what is right at the other, and most of us occupying some spot in the middle of these extremes. The AMM allows a process to persist among individuals who fall anywhere on this spectrum. Accordingly, an individual can make their way along this continuum without endangering the process.

One reason the AMM is important is to facilitate moral education. If we required someone to have a full understanding of a process before they are involved with it, there would be no pathway along which they can develop their competence, because there are no steps between having no competence and having a full understanding of the process. This means that we have no ladder for reaching higher levels of understanding.

Similarly, in the algorithmic case, allowing for differing degrees of knowing conformity means that as algorithmic processes become more complex and more inscrutable, their benefits do not diminish and the threat they pose to agency does not increase.

Another important reason is that if the bar for conformity is too high, then the survival of the practice will be in danger because too many people will fail to do their part in it. In the frequent cases where the process requires our conformity to persist, our failing to conform not only makes you not garner success, it also denies others the chance to succeed.

The desire to not be vulnerable to changes in circumstances is one important reason people want to be knowing conformers or why they admire moral sages. But in social contexts where individuals are dependent on each other deciding to play their part in a process, an insistence on knowing conformity makes the process less secure, which in turn makes changes in circumstances more likely as these processes fall away. In these cases, the desire to only engage in knowing conformity is counterproductive and doomed to failure.

Let us return to the notion of a division of deliberative labour. Instead of seeing the parent and the child or the bank manager and the loan categorising algorithm as different individuals interacting with each other,[32] we can instead look at groups where the members follow the lead of some individual who is able to garner first-order success (maybe the same individual every time, maybe not). In this case we are aggregating deliberative capacities of the group together, which allows these to be better mobilised by the group.[33]

Conclusion

This paper has dealt with the threat of algorithmic decision-making to agency by explaining how individuals can play their part in some process without fully understanding what that involves. When we distinguish first- and higher-order knowledge of an instance of action-guidance we see that success at a given process strictly requires only first-order knowledge. This means that they can do their part and reach the end that justifies the process even if they are ignorant about the deeper reasons for why it is appropriate, as when we follow guidance from algorithms and other inscrutable sources. I presented the alternative method model, the mechanism that allows for these divisions of deliberative labour: if the process we are engaging in is robust enough we can replace the reasoning of a fully-informed individual knowingly conforming to what the process requires with something less comprehensive, including merely conforming to the guidance from an algorithm or other inscrutable source. I surveyed various problems that can arise if individuals suffer from higher-order ignorance, of which the most prominent is that they are liable to faulty extrapolations from mistaken judgements about what the pertinent higher-order features are. But there is ample reason to allow for mere conforming, because not doing so would diminish the range of target ends that are reached if we do not allow mere conforming, with no compensating benefit. The fact that we can be justified in participating in processes we do not understand is what makes divisions of deliberative labour possible.

Notes

[1] Priyank Jain, Manasi Gyanchandani, and Nilay Khare, "Big Data Privacy: A Technological Perspective and Review," *Journal of Big Data* 3, 1 (2016).

[2] Lambèr Royakkers et al., "Societal and Ethical Issues of Digitization," *Ethics and Information Technology* 20, 2 (2018).

[3] E.g. Sarah Brayne, "Big Data Surveillance: The Case of Policing," *American Sociological Review* 82, 5 (2017).

[4] E.g. Peter Grindrod, "Beyond Privacy and Exposure: Ethical Issues within Citizen-Facing Analytics," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, 2083 (2016); Brent Daniel Mittelstadt et al., "The Ethics of Algorithms: Mapping the Debate," *Big Data & Society* 3, 2 (2016); Matteo Turilli and Luciano Floridi, "The Ethics of Information Transparency," *Ethics and Information Technology* 11, 2 (2009).

[5] Examples in the popular media include Charles Duhigg, "How Companies Learn Your Secrets," *The New York Times Magazine*, 16 February 2012; Harry Davies, "Ted Cruz Using Firm That Harvested Data on Millions of Unwitting Facebook Users," *The Guardian* 11 (2015); Julia Angwin et al., "Machine Bias," *ProPublica*, May 23 2016; Peter Waldman, Lizette Chapman, and Jordan Robertson, "Palantir Knows Everything About You," *Bloomberg Businessweek*, 19 April 2018.

[6] The GPT-3 language parser, which is on the cutting edge of this technology, tracks 175 billion parameters. Tom B Brown et al., "Language Models Are Few-Shot Learners," *arXiv preprint arXiv:2005.14165* (2020).

[7] I concur with the view of Martin, discussed below in §II. and n14, that sometimes the choice to use an algorithm that is inscrutable in this sense is a political choice and is open to criticism on those grounds. The point is that not all the problems of inscrutability are because of political choices, and even the most transparent algorithm in the political sense can be inscrutable in the sense that humans cannot follow the reasoning.

[8] The notion of a division of deliberative labour links with extensive literatures regarding delegated responsibility, paternalism, and ‘nudging’, too large to cite here.

[9] Luciano Floridi, "Distributed Morality in an Information Society," *Science and Engineering Ethics* 19, 3 (2013); "Faultless Responsibility: On the Nature and Allocation of Moral

Responsibility for Distributed Moral Actions," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, 2083 (2016).

[10] Kirsten Martin, "Ethical Implications and Accountability of Algorithms," *Journal of Business Ethics* (2018); "Designing Ethical Algorithms," *MIS Quarterly Executive* 18, 2 (2019).

[11] See Philip Soper, *The Ethics of Deference: Learning from Law's Morals* (Cambridge University Press, 2002); James A. Grant, "Reason and Authority in Administrative Law," *The Cambridge Law Journal* 76, 3 (2017); Yoav Dotan, "Deference and Disagreement in Administrative Law." I thank the editor for this suggestion.

[12] "Deference and Disagreement in Administrative Law," 773-74.

[13] For a survey, see Mittelstadt et al., "The Ethics of Algorithms: Mapping the Debate," 10-12.

[14] Martin, "Designing Ethical Algorithms," 138-39; "Ethical Implications and Accountability of Algorithms," 840-41.

[15] Paola Tubaro, Antonio A. Casilli, and Marion Coville, "The Trainer, the Verifier, the Imitator: Three Ways in Which Human Platform Workers Support Artificial Intelligence," *Big Data & Society* 7, 1 (2020).

[16] For an overview, and a critical discussion of whether satisfaction conditions are reducible to propositions, see Chris Fox, "Imperatives: A Judgemental Analysis," *Studia Logica: An International Journal for Symbolic Logic* 100, 4 (2012).

[17] This has a historical analogue in the debate between the Stoics and their ancient skeptic opponents about whether it is possible to have knowledge of the world secure from error. Stoics held that a fool assents to uncertain opinions, and a sage only assents to opinions that are certain to be true. Mere conformity would be foolish assent, and knowing conformity sagely assent. See e.g. Meinwald Constance, "Ignorance and Opinion in Stoic Epistemology," *Phronesis* 50, 3 (2005).

[18] See Diana Baumrind, "Authoritative Parenting Revisited: History and Current Status," (2013).

[19] See e.g. Georgi Gardiner, *Legal Burdens of Proof and Statistical Evidence* (Routledge London, 2018). I thank an anonymous reviewer for raising this issue.

[20] Matthew Adam Bruckner, "The Promise and Perils of Algorithmic Lenders' Use of Big Data Fintech's Promises and Perils," *Chicago-Kent Law Review* 93, 1 (2018).

[21] Ibid., 13-14, 22.

[22] For a survey, see Joe Pater, "Generative Linguistics and Neural Networks at 60: Foundation, Friction, and Fusion," *Language* (2019).

[23] A. Halevy, P. Norvig, and F. Pereira, "The Unreasonable Effectiveness of Data," *IEEE Intelligent Systems* 24, 2 (2009).

[24] C.f. for instance the debate about the so-called 'epistemic consequentialism' of reliabilism, which discusses various epistemic goods to do with truth and justification but nothing that could help here: Selim Berker, "Epistemic Teleology and the Separateness of Propositions," *The Philosophical Review* 122, 3 (2013); Kristoffer Ahlstrom-Vij and Jeffrey Dunn, "A

Defence of Epistemic Consequentialism," *The Philosophical Quarterly* 64, 257 (2014); Alvin I. Goldman, "Reliabilism, Veritism, and Epistemic Consequentialism," *Episteme* 12, 2 (2015).

[25] Raimo Tuomela, *Social Ontology: Collective Intentionality and Group Agents* (Oxford University Press, 2013).

[26] Margaret Gilbert, *Joint Commitment: How We Make the Social World* (Oxford University Press, 2013).

[27] Seumas Miller, *Social Action: A Teleological Account* (Cambridge: Cambridge University Press, 2001); Michael E Bratman, *Shared Agency: A Planning Theory of Acting Together* (Oxford University Press, 2013).

[28] Brian Epstein, *The Ant Trap: Rebuilding the Foundations of the Social Sciences* (Oxford University Press, USA, 2015).

[29] One way to avoid the lacuna is to describe group agency as the result of aggregating individual attitudes and actions using social-choice-theoretic tools, a foundational feature of which is that the aggregations have properties which occur in no individual. See Christian List and Philip Pettit, *Group Agency: The Possibility, Design, and Status of Corporate Agents* (Oxford: Oxford University Press, 2011).

[30] Akhila Narla et al., "Automated Classification of Skin Lesions: From Pixels to Practice," *Journal of Investigative Dermatology* 138, 10 (2018).

[31] E.g. Rashida Richardson, Jason Schultz, and Kate Crawford, "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice," *New York University Law Review Online* (2019).

[32] *Pace* Martin.

[33] As in List and Pettit, *Group Agency*.