Written evidence submitted by the Inter-Disciplinary Ethics Applied (IDEA) Centre, University of Leeds (ADM0025)

Task

1. The extent of current and future use of algorithms in decision-making in Government and public bodies, businesses and others, and the corresponding risks and opportunities;
2. Whether 'good practice' in algorithmic decision-making can be identified and spread, including in terms of:
    1. The scope for algorithmic decision-making to eliminate, introduce or amplify biases or discrimination, and how any such bias can be detected and overcome
    2. Whether and how algorithmic decision-making can be conducted in a 'transparent' or 'accountable' way, and the scope for decisions made by an algorithm to be fully understood and challenged;
    3. The implications of increased transparency in terms of copyright and commercial sensitivity, and protection of an individual's data;
3. Methods for providing regulatory oversight of algorithmic decision-making, such as the rights described in the EU General Data Protection Regulation 2016.

This page has links to the written evidence the committee has already received
http://www.parliament.uk/business/committees/committees-a-z/commons-select/science-and-technology-committee/inquiries/parliament-2017/algorithms-in-decision-making-17-19/publications/
It will give you some ideas on style and required brevity.
This page http://www.parliament.uk/get-involved/education-programmes/universities-programme/research-impact-at-the-uk-parliament/how-to-guides/ has some short guides on writing successful written evidence.

Written evidence submitted by the Inter-Disciplinary Ethics Applied (IDEA) Centre, University of Leeds (ADM0025)

Response

1. Extent, risks, opportunities

There are three main issues that we wish to raise in this submission. The first has to do with the extent of algorithmic decision-making, and in particular the question of which decisions are suitable for this approach. The second has to do with a risk, that it may lead to violations of respect for the dignity of human beings. Violations of that form of respect may also have knock-on consequences for the familiar political ideal of treating all citizens as equals. The third is that automated decision-making systems may not be driven by anything like ordinary-style reasons, and may therefore lack accountability. We identify two types of solution, but both have problems. The right solution will depend on context: the weightier the ethical considerations at stake, the more restrictions we will need to place on the way automated decision-making systems operate.

1.1 Making Decisions about Values

So, what kinds of decisions are inappropriate to be tackled by an algorithm, and which are appropriate? To claim that decisions about values, especially competing values can be decided correctly - i.e. morally correctly - by running an algorithm is a very, very controversial claim (even if the input data is very extensive). Clearly technical calculations involving measurable empirical data can be tackled well by algorithms, but most decisions in policy, both strategic and tactical, involve an alertness to, and an open-ended evaluation of the significance of, value considerations.

There is a significant literature on the relevant areas of ethics, e.g. around the codifiability of ethics and of virtuous decision-making (McDowell's "Virtue and Reason" is a classic paper in this area); around whether values are comparable at all, let alone commensurable (e.g. R. Chang, ed. *Incommensurability, incomparability, and practical reason*, Harvard University Press 1997); and around the extent to which generalisations in ethics hold at all (cf. Aristotle on precision in ethics, and contemporary 'particularists' e.g. Jonathan Dancy, Brad Hooker). There are discussions in the ethics of AI / Robotics, e.g. in relation to the military use of autonomous robots, about the extent to which even a very sophisticated algorithm-based device can appropriately weigh relevant ethical considerations (e.g. Simpson/Müller 2016).

What algorithms plausibly could do is continue a pattern of decision-making from a given (potentially large) sample. But this is itself fraught with further ethical questions about how one should select the relevant sample in ways that would not important significant ethical risk (e.g. of entrenching present injustices / prejudices, in the ways that have come to light in apps that

select faces for attractiveness -- in ways the designers did not predict), and exactly how the algorithm should extrapolate from the data set to new cases.

Could these concerns be managed in such a way as to render the ethical risks acceptable? That question might be answerable, but it is a very substantial matter to answer it, and it looks to be a vital question to answer for anyone defending the use of algorithms in decision-making about questions with evaluative content (i.e., let us recall, pretty much all policy decisions).

1.2. Algorithms and Opacity Respect

What does it mean to respect people as political equals, as individuals who all count for exactly the same when it comes to their rights and entitlements as citizens? Indeed, why should we respect people as equals at all when it is clear that we all differ in terms of our natural capacities? It is readily apparent that some people are physically stronger than others, or cleverer, or nicer, or prettier, and so on. Even when it comes to more abstract abilities, those abilities that mark us out as *rational agents*, such as reasoning, judging what is in our self-interest, and pursuing our particular conception of the good, we place on different points of a spectrum. The philosopher Ian Carter, in his 2011 paper 'Respect and the Basis of Equality', however, has argued that there is one crucial respect in which we are all the same, and that this explains why, and how, we are to be respected as equals. His thought is that we all have an interest in being able to conceal or cover up aspects of ourselves in order to maintain a level of outward dignity. If someone else was able to listen to all of my thoughts, or observe me at every moment, then they would know that I often do irrational things and behave in strange and eccentric ways. If that is right, then we have reason to be concerned about any use of algorithms that allows private individuals, companies, or the government to "look inside" of us.

Carter (2011, p.550) argues that our default position should thus be one of "evaluative abstinence" towards each other, in which we deliberately refrain from making judgments about how good or bad other people are at being agents and treat everyone simply as equals. This is what it means to show one another what he calls *"opacity respect"*. Of course, there will be contexts in which that is not possible. Doctors need to determine the capacity of their patients. Teachers need to measure the progress of their students. Friends and partners need to decide when their loved ones may need help or reassurance. However, we should distinguish our interactions with one another in these roles from the other ways in which we relate to each other, and one particular sphere in which a commitment to opacity respect is vital is in the relationship between citizens and their state.

So long as I cross the basic threshold for consideration as an independent moral agent, it would be deeply inappropriate for the state to make these kinds of judgments about me. It is especially stigmatising and insulting if the state deems an individual or a group to be less capable or less worthy of respect than anyone else. And it is also especially dangerous. The

awesome power of the state provides a temptation to interfere in the lives of citizens for their own good. If the state comes to view me, people like me, or the citizenry as a whole as malfunctioning or compromised, then the temptation to interfere may become too great.

The use of algorithms can *undermine opacity respect* from two directions. First, at the design stage, there may be pressure to discover and incorporate data about how good or bad certain types of people might be at performing key functions. Second, existing algorithms, or indeed datasets, may be turned from their original purposes, such as in making or facilitating very specific decisions, and used to “look inside” people as explained above.

1.3. Accountability Of Algorithmic Decision-Making Systems

A key issue with algorithmic decision-making systems is accountability: specifically, the difficulty of querying the reasoning behind a system's decisions. This is not merely the difficulty of interrogating the system. The system may not be driven by anything like ordinary-style reasons at all. Using artificial intelligence techniques, the system can work up whatever weird and wonderful principles fit the previous cases we fed it. So its decision-making model may not be driven by anything like reasons of the ordinary kind.

This is particularly problematic in policy-making. If we implement the recommendations of an algorithmic decision-making system, there is no guarantee that our policies could be justified in terms of reasons of the ordinary kind. So the policy-making process would lack transparency. This would also mean the policy-making process lacked accountability of the right kind. For example, in healthcare, if my fellow citizens deny me life-saving treatment, surely I am entitled to know why, and to insist that the reasons be good ones. With a policy derived from a "black box", policy-makers could not give me good reasons. It is for such reasons that the General Data Protection Regulation (GDPR) foresees a “right to explanation”.

A similar issue applies in the private sector, eg with respect to recruitment decisions and decisions about insurance premiums and loans. In recruitment, there are already commercial applications that sift CVs and use algorithms to aid human decision-making. One commentator said “I fear for diversity. If you plug AI into companies that already aren’t that diverse, it becomes a self-fulfilling prophecy." A recent study found that women were served Google ads for high paying jobs far less than men. So, if an algorithmic decision-making system rejects an ethnic minority candidate, but cannot state its reasoning understandably, how do we know whether or not the system reached its decision in a non-discriminatory way?

In the following, we consider various possible solutions ("Algorithm Accountability Tools"), and we then propose a process by which policy-makers should evaluate these possible solutions.

Written evidence submitted by the Inter-Disciplinary Ethics Applied (IDEA) Centre, University of Leeds (ADM0025)

ALGORITHM ACCOUNTABILITY TOOLS

ACCOUNTABILITY TOOL 1: Restrict algorithms to supporting human decision-making. - But this option will have few benefits on its own. Algorithms will still need to be transparent in this context, since otherwise, if the humans in question don't understand why an algorithm made its recommendations, they will have to either trust the algorithm unthinkingly, or go entirely on their own reasoning.

ACCOUNTABILITY TOOL 2: Help system users or data subjects understand how the system works in general terms without trying to explain how their particular decision came out as it did. - At the moment, it looks like GDPR only gives data subjects the right to explanations at this level. It's a poor form of transparency, but could have some value in certain contexts.

ACCOUNTABILITY TOOL 3: Restrict algorithms to inherently more understandable types, e.g. decision tree learning rather than “deep” learning in neural networks. Require systems to deploy understandable trade-off principles between conflicting considerations. The consequence should be that human decision-makers could reconstruct the reasoning behind each decision and check it for acceptability. This limits the benefit we can expect from algorithmic decision-making, but it might be appropriate in contexts where transparency is critical.

ACCOUNTABILITY TOOL 4: Restrict inputs that carry weight, ie prohibit certain types of variable. For example, in recruitment systems, we could exclude data about ethnicity but permit data about job-related abilities. - However, there are problems with this approach. One is that it would place severe limits on the value we get from algorithmic decision-making. The potential value of algorithms derives precisely from their ability to outrun us in terms of their selection of parameters and the way they combine them. Another issue is that this won't avoid the possibility of discrimination. For example, if you want to avoid discriminating against ethnic minorities and so you exclude variables based on ethnicity, there will still be a potential for discrimination based on the other information that correlates with ethnicity. Excluding all this other information will limit the usefulness of the system.

ACCOUNTABILITY TOOL 5: Demand Explainable AI. - Many academics are working on ways of getting AI systems to communicate their reasoning understandably. If we can understand the "reasoning" it will be easier to assess its acceptability. However, to the extent we demand explicability, we will limit the benefits we can get from algorithmic decision-making.

ACCOUNTABILITY TOOL 6: Demand ad hoc explanations. - Even if a system is impossible to explain in toto, the rationale for individual bits of output may be more understandable. If a system's output in a given domain raises concerns, data subjects or regulators can ask for an explanation of the output in that domain. As an illustration, when Google was accused of showing fewer ads for well-paid jobs to women, a study found that one possible reason was

that women are a more expensive demographic to advertise to. Google's algorithm is perceived as very complex, yet this explanation is quite simple.

ACCOUNTABILITY TOOL 7: Jenna Burrell suggests "Abandon answering the 'why' question and devise metrics that can, in other ways, evaluate discrimination (i.e. Datta et al., 2015). - For example, in 'Fairness Through Awareness' a discriminatory effect in classification algorithms can be detected without extracting the 'how' and 'why' of particular classification decisions (Dwork et al., 2011)." For example, we could evaluate the system's decisions in a large number of hypothetical scenarios, to determine whether the system has a tendency to disadvantage certain groups. This illustrates that accountability needn't always involve transparency. But an issue with this approach is that it requires us to know what the correct decision is in each scenario; if the motivation for deploying the system is that human decision-making is held to be fallible, then we will lack the requisite "Archimedean point" from which to evaluate the system's decisions.

HOW TO ASSESS THE ACCOUNTABILITY TOOLS

The question is how to choose between these tools. We may have to compromise between mitigating risks and exploiting the full value of algorithmic decision-making. The right balance will depend on the decision-making context. For example, if the decisions have serious consequences (eg life and death in healthcare), or if other important ethical considerations are identified (eg the possibility of discrimination in HR recruitment), then that will drive us towards stronger forms of accountability. Otherwise, perhaps we can accept weaker forms of accountability so that we can exploit the potential of algorithmic decision-making to its fullest.

We suggest that one way to address these issues is a research project with roughly the following stages (not necessarily in order):

1. Identify the accountability tools that are feasible, eg the types of transparency that might be achievable. For example, engage in dialogue with academic theorists of Explainable AI to clarify what might be possible in future.

2. Consider the value of these accountability tools in different decision-making contexts. Start by identifying the audiences to whom a system must be accountable in each context, eg internal system users vs external data subjects vs other stakeholders. Identify the risks those audiences are vulnerable to. In light of that, identify the most useful accountability tools in each context.

3. Consider the costs of each accountability tool in each context, eg in terms of the ways a tool might limit the benefits of algorithmic decision-making.

4. Develop a set of considerations and criteria to help policy-makers and regulators to decide how algorithmic decision-making can be regulated commensurately with the cost and benefits in each decision-making context, and to help business decision-makers plan their algorithm development projects with a view to incorporating a wider range of possible accountability tools.

We would expect these steps to be somewhat iterative. For example, consideration of the costs and benefits of a certain accountability tool may lead to a re-engagement with academic theorists to determine how far the boundaries of that tool might be pushed.

2. Recommendations for good practice

We would like to recommend the following principles to guide practice:

A. In designing algorithms to aid with decision-making, we should not attempt to "look inside" individuals for the sake of increasing predictive power or otherwise boosting their utility. The state in particular should neither commission nor use existing algorithms for the purpose of "looking inside" individual citizens, or groups of citizens.
B. Consider the costs and benefits of alternative accountability tools in different decision-making contexts. Identify the audiences to whom a system must be accountable in each context. Identify the risks those audiences are vulnerable to. In light of that, identify the most useful accountability tools in each context.
C. Conduct research along the lines proposed above to help policy-makers and regulators decide how algorithmic decision-making can be regulated commensurately with the cost and benefits in each decision-making context, and to help business decision-makers plan their algorithm development projects with a view to incorporating a wider range of accountability tools.

December 2017