

Reviewable Automated Decision-Making

Jennifer Cobbe, Jatinder Singh
Compliant and Accountable Systems Research Group
Dept. of Computer Science and Technology
University of Cambridge, UK
firstname.lastname@cst.cam.ac.uk

ABSTRACT

In this paper we introduce the concept of ‘reviewability’ as an alternative approach to improving the accountability of automated decision-making using machine learning systems. In doing so, we draw on an understanding of automated decision-making as a socio-technical process, involving both human (organisational) and technical components, beginning before a decision is made and extending beyond the decision itself. Although explanations for automated decisions may be useful in some contexts, they focus more narrowly on the model itself and therefore do not provide the information about that process as a whole that is necessary for many aspects of accountability, regulatory oversight, and assessments for legal compliance. Drawing on previous work on the application of administrative law and judicial review mechanisms to automated decision-making in the public sector, we argue that breaking down the automated decision-making process into its technical and organisational components allows us to consider how appropriate record-keeping and logging mechanisms implemented at each stage of that process would allow for the process as a whole to be reviewed. Although significant research is needed to explore how it can be implemented, we argue that a reviewability framework potentially offers for a more useful and more holistic form of accountability for automated decision-making than approaches focused more narrowly on explanations.

KEYWORDS

automated decision-making; accountable systems; reviewability

1 INTRODUCTION

Our work in this area concerns the accountability of machine learning systems used in automated decision-making (ADM). We consider accountability here to involve technical and organisational mechanisms allowing those responsible for a system to understand how and why it is functioning and to themselves be accountable to external actors and oversight bodies for that functioning and their role in it. We propose that future ADM systems should be engineered to be *reviewable*, a targeted form of transparency that supports accountability of the decision-making process as a whole. *Reviewability* considers ADM not as consisting of a machine learning model and its inputs and outputs, but as involving a technical system and the broader human processes, structures, and systems around it. Although work in progress, we argue that a reviewability approach to accountable ADM can be more holistic and more useful than explanations or other approaches focused more narrowly on models.

2 LIMITS OF EXPLANATIONS

ADM is a complex socio-technical process, with the algorithm itself forming only one part of a broader *algorithmic system* [5] that includes both human (organisational) and technical components. ‘Accountable automated decision-making’, therefore, does not simply mean making the model itself accountable in some way. It involves a view of the whole process from selection of training data and construction of the model; through training, testing, and verification; to inputting data for individual decisions; and on to the effects of those decisions [3]. This requires logging, record-keeping and transparency – not necessarily for those subject to decisions (the danger of falling into the ‘transparency fallacy’ around informing the subjects of decisions has been well argued elsewhere [2]), but as a means of facilitating accountability to and oversight more generally by designers, developers, deployers, users, and overseers.

As such, explanations focused on how the model itself has arrived at a particular output may miss much of what is important. Moreover, from a legal point of view, explanations may not provide the information necessary to determine whether a decision was arrived at lawfully, whether it is discriminatory, to facilitate regulatory oversight, and so on. For that, a more holistic view of the sociotechnical process is needed [1]. Determining lawfulness could include, for example, information on data used to train the model or make the decision in question (including any proxies); information on testing and auditing procedures; information on inferences drawn by the model in decision-making; or information on the effects of decisions and aggregated data on treatment of protected characteristics. Moreover, over-reliance on explanations to subjects of decisions places the burden of challenging that decision on them. Given that subjects of automated decisions are often from an already vulnerable group, it is unrealistic to assume that they have the resources and knowledge to effectively disagree with the decision and advocate for better decisions.

3 REVIEWABILITY: AN ALTERNATIVE APPROACH

Reviewability involves exposing the information required to assess the algorithmic system, its context, and its decisions for legal compliance, for whether it is operating within expected or desired parameters, and so on. This approach is derived from administrative law, the body of law concerned primarily with holding human decision-making in the public sector to account. Administrative law developed over centuries to contend with the opacity of human decision-making and to maintain standards for even the most consequential decisions of life and death. It is therefore particularly relevant to attempts to improve accountability and oversight of automated decisions. In administrative law, there is no general duty

to give reasons (or explanations), but decision-makers are required by law to act in line with long-established principles of good administration throughout the decision-making process. Judicial review of public sector decision-making does not simply therefore assess the decision itself, but the decision-making process as a whole.

In previous work, we considered how administrative law as a framework and judicial review as a form of oversight can apply to public sector ADM [1]. In doing so, we drew on administrative law’s understanding of human decision-making as a process that begins before the decision and that has consequences that resonate afterwards. In administrative law, various aspects of that process are considered both discretely and as part of the whole, enabling the development of principles applying to those various aspects so as to ensure good decision-making. This approach allows courts and other oversight bodies to effectively review automated decisions made by public bodies without, for the most part, requiring explanations of consideration of the model itself.

Step	Stage
Procurement	Commission
Problem definition	
Data collection	Model building
Data cleaning	
Training	
Testing	
Deployment	Decision-making
Use	
Consequences	
Audit	Investigation
Disclosure	

Table 1. Stages of the ADM process, each consisting of multiple steps.

Reviewability of ADM does not therefore necessarily involve *explanations* (although in some cases these may be appropriate). Rather, it is about exposing the decision-making process, understood broadly, including: evaluations by those wishing to deploy systems; decisions by engineers in developing systems; data used to train and test systems; training and testing processes themselves; data used to make automated decisions; inferences drawn by the system in the process of making automated decisions; and the fairness, effects, and lawfulness of those automated decisions in practice. While ‘reviewability’ as a high-level concept has applications in various areas [4], reviewable ADM thus takes a holistic approach to transparency and accountability of algorithmic systems, beyond the narrower focus of explanations, facilitating effective review of the entire decision-making process.

Thinking of ADM as a socio-technical process allows us to break it down into several components – *steps* in producing an automated decision from conception of the system through to the consequences of that decision (Table 1). These broadly group into *stages* of the process. As in administrative law, these steps and stages can

be considered discretely and as part of a whole, and they provide the foundations for setting out the general principles of a framework for developing reviewable ADM systems. At each step there is an opportunity to place limits on the ADM process (established in law, regulation, policy, or otherwise), implement appropriate technical and organisational logging and record-keeping mechanisms to enable review for compliance with those limits and for functioning more generally, and to feedback into the functioning of the decision-making process more generally. This enables a non-linear, cyclical process of review, feedback, and revision in line with the understanding of accountability discussed above and with the view of ADM as a socio-technical process.

4 CONCLUSION AND FURTHER RESEARCH

Reviewability potentially offers a way of thinking about accountable automated decision-making that allows for the socio-technical process to be considered holistically. This is achieved by breaking down that process into its constituent technical and organisational components and using appropriate record-keeping and logging mechanisms to provide targeted transparency that supports meaningful accountability and assessment of the functioning of the algorithmic system.

There is clearly much research needed on implementing reviewability. Although we have undertaken relevant work on using provenance techniques to expose decision pipelines and supply chains [3], the specifics of what kind of record-keeping and logging might be appropriate at each step of the process, of what kind of information would be useful to retain, and of how this information can best be presented to overseers so as to facilitate effective review of the algorithmic system’s operation will depend on the system in question, the domain in which it is deployed, and its purpose. However, the contribution of this piece is the approach to reviewability that we set out herein, rather than in details of implementation. Acknowledging that there is significant work to do on developing reviewability as a framework that can be put into practice, we argue here only at a high level for reviewability as a potentially useful approach to improving the accountability of ADM.

ACKNOWLEDGMENTS

We acknowledge the financial support of the University of Cambridge, through the Cambridge Trust & Technology Initiative, and the UK Engineering and Physical Sciences Research Council (EPSRC) through awards ‘Towards a legally-compliant Internet of Things’ (EP/P024394/1) and ‘Realising Accountable Intelligent Systems’ (EP/R033501/1).

REFERENCES

- [1] Jennifer Cobbe. 2019. Administrative Law and the Machines of Government: Judicial Review of Automated Public Sector Decision-Making. *Legal Studies* 39, 4 (2019). <https://doi.org/10.1017/lst.2019.9>
- [2] Lilian Edwards and Michael Veale. 2017. Slave to the Algorithm? Why a ‘Right to an Explanation’ Is Probably Not the Remedy You Are Looking For. *Duke Law and Technology Review* 18 (2017). <https://ssrn.com/abstract=2972855>
- [3] Chris Norval Jatinder Singh, Jennifer Cobbe. 2018. Decision Provenance: Harnessing Data Flow for Accountable Systems. *IEEE Access* 7 (December 2018). <https://doi.org/10.1109/ACCESS.2018.2887201>
- [4] Chris Norval, Jennifer Cobbe, and Jatinder Singh. 2020 (To appear). *Towards an accountable Internet of Things: A call for ‘reviewability’*. IET.
- [5] Nick Seaver. 2013. Knowing Algorithms. *Media in Transition* 8 (2013).