In: Emre Bayamlıoğlu, Irina Baraluic, Liisa Janssens and Mireille Hildebrandt (eds), BEING PROFILED:COGITAS ERGO SUM. 10 Years of Profiling the European Citizen, 2018: Amsterdam University Press, 36-40. DOI 10.5117/9789463722124/CH5.

How is 'transparency' understood by legal scholars and the machine learning community?

Karen Yeung and Adrian Weller***

Abstract

Algorithmic decision-making systems are increasingly in use yet often lack transparency. The opacity of these 'black boxes' leads to decisions that can be hard to understand and contest, creating substantial risks of injustice. Unless this 'challenge of transparency' can be addressed appropriately, alongside concerns including reliability, 'fairness' and 'algorithmic accountability', the public is unlikely to trust these systems, despite their many benefits. Scholars from both the machine learning and legal communities are actively seeking to understand what types of transparency are required and why they matter, in order to respond to this challenge. Yet many terms are employed loosely in debates about transparency, which can lead to confusion. The purpose of this paper is to articulate and discuss various concepts and terms used in such discussions. It focuses on the legal and machine learning communities, with the aim of improving cross-disciplinary insight and dialogue between them.

Keywords: transparency, accountability, algorithmic decision-making, reasons, right to an explanation.

Introduction

Algorithmic decision-making systems are increasingly in use yet often lack transparency. The opacity of these 'black boxes' leads to decisions that can be hard to understand and contest, creating substantial risks of injustice. ML researchers are developing methods for improving the transparency of these systems ('explainable AI'). Unless this 'challenge of transparency' (Weller 2017) can be addressed appropriately, alongside concerns including reliability, 'fairness' and 'algorithmic accountability', the public is unlikely to trust these systems (RSA 2018), despite their many benefits.

Integrating legal scholarship with ML research

For lawyers, the challenge of transparency is familiar for *human* decision-makers, particularly for decisions by public authorities. Within contemporary constitutional democratic orders, governmental decision-makers must exercise their authority in accordance with law. Contemporary equality legislation is also concerned with preventing and remedying decision-making that is unfairly discriminatory in relation to 'protected' grounds (gender, race, etc.). The law imposes various constraints to address and prevent particular kinds of flaws in human decision-making. These constraints are ultimately grounded in recognition that decision-making authority is vulnerable to corruption and abuse. Transparency is critical for ensuring that decision-making is lawful and accountable. Since the advent of computerised decision-making systems, various jurisdictions have introduced legally enforceable duties, entitling those directly and significantly affected by certain kinds of fully automated decisions to receive an explanation for that decision, although the precise nature of this duty is uncertain.

Within debates about what transparency in machine decision-making requires, many terms are employed by different disciplines, leading to significant potential confusion. Accordingly, we seek to clarify various concepts and terms used in discussions about transparency in decision-making, focusing on the legal and ML communities. We consider why transparency matters to these two communities, aiming to improve cross-disciplinary insight. Because this entails sweeping generalisations, our reflections are offered as heuristics, seeking to capture the kinds of concerns that are frequently raised, thereby facilitating enhanced interdisciplinary understanding and dialogue.

Why transparency matters

For both the legal and ML communities, the needs for transparency are highly context-dependent. In ML, transparency is typically desirable for understanding both specific algorithmic behaviour and the broader socio-technical environment in order to consider how the system will be used in practice. For systems that rely upon data processing to generate decision outputs, transparency is also desirable for the datasets themselves: identifying which data is used, who decides this, and other questions about the data's provenance such as source, volume, quality and pre-processing (Gebru et al 2018). In relation to the computational component of the system, identifying what transparency requires is a function of its context and the character, capacities and motivations of the intended audience (Weller

2017). For example, developers typically want to understand how their overall system works, thereby enabling them to identify and rectify any problems and undertake improvements to system performance. In contrast, individuals directly affected by a machine decision may be concerned with how and why a particular decision was arrived at (a 'local explanation'), in order to evaluate its accuracy and fairness and to identify potential grounds to contest it. Different types of explanation might be appropriate for the affected individual, or for an expert or trusted fiduciary agent.

For human and organisational decision-making, lawyers also recognise the importance of context in identifying what transparency requires. Transparency concerns can be understood as grounded in the requirements of the contemporary concept of the rule of law, which captures a set of normative ideas about the nature and operation of law in society (Craig 1997). One of the rule of law's core requirements is that the laws themselves should be transparent: laws should be publicly promulgated (Fuller 1964, 49) so that all legal subjects can know the law's demands in advance and thus alter their behaviour accordingly. The existence of 'secret' laws of which legal subjects are unaware and could not reasonably have discovered is the antithesis of the rule of law ideal, its tyrannical consequences vividly depicted in Kafka's The Trial (Kafka 1998). The argument made by Schreurs et al. (2008), that data subjects should have access to the knowledge and potential secrets implied in the profiles that are applied to them when they match the criteria of a profile (including in private settings) 'in order to anticipate the actions and decisions that may impact our later life' is a specific application of this general principle applied to automated data-profiling.

The rule of law also requires that the exercise of power by public authorities has a lawful basis. Transparency is necessary to evaluate whether a decision is lawful, and therefore legally justified. Legal justifications typically require explanations. An explanation is typically comprised of the provision of reasons in response to the question: why did you decide that? These reasons, including the factors that were taken into account by the decision-maker, how much weight they were given, and how the totality of relevant factors was evaluated to arrive at a decision, would constitute such an explanation. Justification and explanation are different – an explanation may not, in itself, establish that a decision is legally justified. To justify a decision, the explanation must meet the criteria laid down by law, thereby establishing that the decision-maker had legal authority to make the decision, that no legally impermissible factors were taken into account, and, at least in relation to decisions made by public officials, that the legal conditions that constrain how the decision-making process is conducted were complied with, and whether the substantive decision itself falls within the bounds of legal acceptability (the terminological touchstone for which will vary between jurisdictions – in English administrative law, for example, this requires that the decision must not be 'so unreasonable that no reasonable decision-maker would have arrived at it' – the test established in the famous *Wednesbury* case). In short, an explanation is necessary but not sufficient for establishing that a decision is legally justified.

Even if a decision cannot be legally *justified*, it might nonetheless be lawfully *excused*. This distinction is significant: a justified decision entails no wrongdoing (Hart 1968); a decision or action that is not legally justified might still be lawfully excused, thereby reducing the seriousness of the wrong when considered in the law's response. For example, consider the case of 95-year old Denver Beddows. He repeatedly hammered his wife's head and struck her with a saucepan, despite his lifelong devotion to her, intending to respect her continual requests that he end her life following the deterioration of her health. He was convicted for attempted murder but, owing to the circumstances of the case, was given a suspended sentence in recognition of the moral context and significance of his actions (The Independent 2018). This example points to the crux of why explanations matter: as moral agents, we want not only to understand ourselves as rational actors who can explain our actions by reference to reasons (Gardner 2006) but we also want to understand *why we have been treated in a particular way by reference to reasons*, in terms that we can comprehend. Only then can we can evaluate, both legally and morally, whether that treatment was justified or otherwise excused. Accordingly, if computational systems make decisions that significantly affect us, we rightly expect – as a community of moral agents in a liberal democratic society – that those decisions can be explained by reference to reasons that are intelligible to us, thereby enabling us to evaluate whether the decisions were legally and morally justified.

Terminology

Transparency intersects with many related concepts, which are sometimes used interchangeably. To help avoid confusion within and across disciplines, we consider terms and their relationship to each other.

a) Interpretability, intelligibility and transparency: Within the ML community, a distinction is increasingly made between (i) 'transparency', understood as the ability to inspect the inner details of a system, for example by seeing the entire code, and (ii) 'interpretability', in the sense of intelligibility to an individual so she can understand why a particular output was generated, in terms that she can comprehend.

b) Information, reasons and explanations: Rendering any decision-making system intelligible to those directly affected by the decisions which it generates will typically require the provision of the underlying reasons why it was reached. For lawyers and legal scholars, providing reasons is distinct from providing information. As legal philosopher, Joseph Raz puts it:

> Whatever provides a (correct) answer to questions about the reasons why things are as they are, become what they become, or to any other reason-why question, is a Reason….What is important is the distinction between providing (or purporting to provide) information ('It is 4 pm', 'She is in Sydney') and providing (or purporting to provide) explanations. Reasons provide explanations (Raz 2011, 16).

In short, explanations require reasons. Raz explains that explanations may be relative to the person(s) for whom they are intended. For him, an explanation is a good one if it explains what it sets out to explain in a way that is accessible to its addressees, i.e. in a way that the addressees could understand were they minded to do so, given who they are and what they could reasonably be expected to do in order to understand it (Raz 2011, 16). Yet it is also necessary to specify what it explains in order to convey any useful information. But whether an explanation is a good one does not affect its character as an explanation. For Raz, an explanation of the nature of laser radiation suitable for university students is an explanation of laser radiation, even when addressed to primary school children (Raz 2011, 16).

c. Normative reasons and justifications: Explanations are the subject of a huge body of philosophical reflection, especially for philosophers interested in 'normative reasons'. Raz argues that normative reasons are those which count in favour of that for which they are reasons: they potentially justify and require what they favour (Raz 2011, 18) although they do not always do so. For both lawyers and philosophers, justifications are particularly important, because they serve to establish that a particular action was not morally wrongful and therefore not worthy of blame or punishment (Gardner 2006). Accordingly, if a decision generated by an algorithmic decision-making system can be regarded as justified, this means that that the decision entailed no wrongdoing. For the individual who is unhappy with the decision in question, then that individual would have no basis for challenging the outcome of the decision on the basis that the wrong outcome was arrived at.

Conclusion

Further work to clarify the needs for appropriate transparency is urgently needed for legitimate and effective deployment of algorithmic systems across society. For both communities, work to improve transparency may have a cost in terms of other values such as privacy. We shall explore these themes in a longer article to come.

* Karen Yeung is Interdisciplinary Professorial Fellow in Law, Ethics and Informatics at the University of Birmingham School of Law and the School of Computer Science.

**Adrian Weller is Programme Director for AI at The Alan Turing Institute, the UK national institute for data science and AI. He is a Senior Research Fellow in Machine Learning at the University of Cambridge, and at the Leverhulme Centre for the Future of Intelligence.

.

References

Craig, Paul. 1997. "Formal and Substantive Conceptions of the Rule of Law: An Analytical Framework." Public Law 33: 467–87.

Fuller, Lon L. 1964. The Morality of Law. New Haven: Yale University Press.

Gardner, John. 2006. "The Mark of Responsibility (With A Postscript on Accountability.)" In Public Accountability, Designs, Dilemmas and Experiences, edited by Michael W. Dowdle, 220-42. Cambridge: Cambridge University Press.

Gebru, Timmit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III and Kate Crawford. 2018. “Datasheets for Datasets.” Proceedings of the Accountability, and PMLR 80, 2018. arxiv.org/abs/1803.09010.

Hart, H. L. A. 1968. Punishment and Responsibility. Oxford: Clarendon Press.

Osborne, Samuel. 2018. “Elderly man who tried to kill wife with hammer so she could avoid care home spared jail after she forgives him.” The Independent, April 25, 2017. https://www.independent.co.uk/news/uk/crime/man-hammered-wife-to-kill-spared-jail-after-she-forgives-him-denver-beddows-olive-a7702076.html

Kafka, Franz, and Breon Mitchell. 1998. The Trial: A New Translation, Based on the Restored Text. New York: Schocken Books.

Raz, Joseph. 2011. From Normativity to Responsibility. Oxford: Oxford University Press.

RSA. 2018. “Artificial Intelligence: Real Public Engagement.” May 31, 2018. https://www.thersa.org/discover/publications-and-articles/reports/artificial-intelligence-real-public-engagement

Schreurs, Wim, Mireille Hildebrandt, Els Kindt, and Michaël Vanfleteren. 2008. *“Cogitas, Ergo Sum.* The Role of Data Protection Law and Non-discrimination Law in Group Profiling in the Private Sector.” In Profiling the European Citizen: Cross-disciplinary Perspectives, edited by Mireille Hildebrandt and Serge Gutwirth, 241-64. Dordrecht: Springer.

Weller, Adrian. 2017. “Challenges for Transparency.” International Conference on Machine Learning 2017 Workshop on Human Interpretability. arxiv.org/abs/1708.01870.