

AN ASSESSEMENT OF ALGORITHMIC ACCOUNTABILITY METHODS

Cristina VOINEA^{a*}, *Radu USZKAI*^b

^a *University of Bucharest, Romania*

^b *The Bucharest University of Economic Studies, Romania*

ABSTRACT

The aim of this paper is to map and assess some of the solutions that have been put forth to problems like algorithmic opacity or algorithmic discrimination. We will focus mainly on algorithmic transparency – understood as the unconditional disclosure of the data algorithms are working on as well as of their inner workings – due to the prevalence of this solution in the literature (see, for example, (Brauneis & Goodman, 2017) (Diakopoulos & Koliska, 2017) (Zarsky, 2016) Transparency is the main type of accountability of algorithmic advanced by various researchers or digital rights activists. After critically assessing the practical consequences and some of the implications of making algorithms transparent, we claim that what is lacking in the online environment is trust in automated decision-making systems. We see trust as resting at the basis of accountability methods and claim that the main issue with transparency is that it cannot restore trust. As such, the edifices of accountability mechanisms will eventually collapse if they do not obtain their legitimacy from the trust that has been invested in them.

KEYWORDS: *algorithmic accountability, opacity, sharing economy, transparency, trust.*

1. INTRODUCTION

The issue of algorithmic accountability stems from the realization that although efficient, these artifacts present the same problems as human decision makers do. They discriminate, amplify structural injustices, produce errors and even mislead or produce errors. There have been many controversies connected to algorithmic failures recently. For example, the YouTube advertising algorithm placed ads from some of the biggest global brands on videos of hate speech (Plummer, 2017). Amazon has removed a recruiting engine which showed bias against women; more specifically, an algorithm which was supposed to sort through all the CVs that the company was receiving in order to select the top five candidates, was not ranking candidates in a gender-neutral way (Dastin, 2018). And, of course, Google had to deal with the fact that its algorithm directed people looking for information about the Holocaust to neo-Nazi and Holocaust denial websites (Lapowsky, 2018). As a result of this intensifying controversies, people have started to question the promise of objectivity and axiological neutrality algorithms were said to guarantee. The loss of trust can be accounted for by the fact that our lives are more and more mediated by highly complex automated services and platforms, whose effects we cannot control nor understand. This is one of the causes for which we start to see ourselves as subjects of hidden and incomprehensible sources of risk.

Even though mistrust is on the rise, the number of applications using decision-making algorithms and, implicitly, their number of users has been constantly expanding. How can we explain this apparently paradoxical situation? Nowadays, people in industrialized countries are locked-in within

* Corresponding author. E-mail address: cristina.a.voinea@gmail.com

digital ecosystems – as more and more activities are accomplished with the help of digital tools – and also, they cannot escape being subjects of algorithmic decision making (especially, when states use algorithmic decision-making systems to optimize public services). As such, the number of users and the spread of the use of apps and services based on algorithmic decision-making is not an indicator for users' trust, given that opting out of these systems is extremely difficult and sometimes even impossible.

But why is trust an important issue that must be considered when building accountability mechanisms? The answer is fairly easy and refers to an old problem in political philosophy, namely "who will watch the watchers?" As such, trust is needed especially in contexts in which robust guarantees are impossible. The proposal for algorithmic transparency, advocated by many, is said to restore trust in these artifacts. Our claim is that transparency is not the suitable method for this purpose. We will support this claim by a set of practical and philosophical concerns regarding transparency, following and building upon de Laat's account and analysis (2017). We will first give an overview of what algorithmic transparency actually means and on what assumptions it rests.

2. WHAT IS ALGORITHMIC TRANSPARENCY?

In order to mitigate the increasing power asymmetries between companies and users, privacy advocates and digital rights partisans have put forth the concept of algorithmic transparency, which is framed as the need to make available to users the model of the algorithm as well as its inputs, meaning the data the algorithm is working on (Diakopoulos & Koliska, 2017). The claim is that total transparency is the perfect recipe for restoring accountability for algorithmic decision-making systems which seem to reify old-age social and political problems.

Accountability is sometimes seen as a successor of trust, but more recent work on accountability proposes connecting the two of them: "an intelligent system of accountability would support intelligent placing of trust" (Morris & Vines, 2014, 20). As such, algorithmic transparency has been put forth as a method of accountability which would entail the consolidation of trust. The assumption is that when able to see inside a system, everybody can make their contribution by understanding what they are subject to and, in the same time, by amending or changing the rules or procedures whenever something goes wrong. The more facts revealed, the more people can understand by a logic of accumulation. Thus, algorithmic transparency rests on the claim that access to source code as well as to the data the algorithm is crunching is the most efficient way of granting people the opportunity to assess its fairness.

Our claim is that even if an algorithm is made explicit or transparent and can be inspected, the light it will shed on potential consequences may be minimal, particularly when the algorithm is complicated. In what follows we will briefly give an account of the main problems that could potentially arise when making algorithms and the data they are crunching transparent. This critical assessment will show that transparency is in no way a solution for restoring trust but, on the contrary, it could amplify the problem.

2.1 The loss of privacy when data sets become public

Defenders of algorithmic transparency claim that the data algorithms are processing must become openly available to everybody, especially because data can incorporate and reflect structural injustices. As such, one must first assess and make sure that the data is 'clean', meaning that it does not embed or reflect historical biases or other such issues. Access to data is seen as particularly important in the case of machine learning algorithms, which are so complex that even programmers cannot understand them – this the reason why they are sometimes called 'black box' algorithms (Domingos, 2015). In these cases, transparency is directly associated with access to data.

But evidently, when data sets become public, privacy could potentially be affected. Full transparency can do great harm. The Ashley Madison hack is a clear example that transparency of

data sets can have horrific consequences (Mansfield-Devine, 2015). After a group of hackers revealed the names, home addresses, search histories and credit card numbers of the users of the online dating site marketed to people who are married or in committed relationships (their motto was "Life is short. Have an affair!"), many users have seen their lives unravel. Their families have fallen apart, they lost their jobs, some of them have been subject to extortion and there were even some suicides reported in connection with this data breach. In this case, disclosure is not the same as transparency, but the consequences of the Ashley Madison hack can illuminate some potential adverse effects of transparency of data. It seems that without a rationale of why some part of a system should be revealed to the public, transparency can threaten privacy, harm individuals and even destroy trust. Moreover, transparency of data may expose individuals or groups to intimidation by powerful and potentially malevolent authorities or organizations. By combining multiple datasets, people can infer who the subjects are, and this is why transparency of data sets must only be made possible for authorities or organizations which are clear about their intentions and interests (Laat, 2017): "In particular, prudence suggests that the datasets involved are only to be made available upon request to intermediate parties that have oversight authority of a kind. The public at large is not to be entrusted with the power to scrutinize datasets as used in machine learning (although the entries of individuals should of course be open for personal inspection)."

This restriction on the transparency of data sets does not logically extend to situations where individuals ask access to the data companies hold on them (a possibility which was actualized when GDPR was enforced). Privacy harms can emerge when unauthorized actors have access to multiple datasets that could be merged so as to combine multiple datapoints about particular individuals which can then be used to reveal their identity.

2. 2 Gaming the system

If the aim is to retain the efficiency of algorithms and to keep using them for optimizing various decision-making processes, while in the same time correcting some of the errors they produce, then making the algorithms available to the wide public is not a very wise solution. The disclosure of the inner working of algorithms might have unintended consequences, like 'gaming the system'.

A paper called "Manipulating Google Scholar Citations and Google Scholar Metrics: simple, easy and tempting" (Lopez-Cozar, Robinson-Garcia, and Torres-Salinas 2012) has showed how easy it is to game the Google scholar algorithm in order to increase the number of citation to a paper. A couple of researchers have discovered a way to trick the algorithms into counting more citations than a paper had. The authors only had to upload six papers authored by a fictitious researcher to an institutional web domain. These papers referenced all the publications for which the authors of the study wanted to increase the number of citations. In a month, they observed an explosion in their Google Citations profiles. Of course, this is a case of gaming the system through reverse engineering. The authors did not have direct access to the algorithm, but we can only imagine how much easier it would have been if they could directly inspect the algorithm.

Another example might offer even more evident indications that sometimes, disclosing the algorithms might have unintended effects. It is well known that IRS uses special detection software in order to predict possible tax evasion. The algorithm behind that software looks for certain metrics in tax returns that are correlated with tax evasion, based on returns previously analyzed. But if the public knew exactly the criteria which the algorithm interprets as signs of tax cheating, they might easily use that knowledge in order to adjust their behavior so as to avoid leaving such traces in their tax returns (Laat, 2017). The consequence would be that some very useful such systems might lose their predictive value. Gaming potentially undermines the algorithm's accuracy and its effectiveness, which could potentially be one downside of making algorithms transparent.

2.3 The potential loss of competitive edge

The companies that develop and create decision-making algorithms are usually opting for protecting their innovations either by using patents (although not so often, given that patents presuppose exposing the mechanism for the public) or trade secrets. Most of the companies choose trade secrets because they can be extended indefinitely in time, which is one of the reasons for considering them better mechanisms for guaranteeing competitive edge over competitors operating in the same field.

Intellectual property is not, of course, the only mechanism for protecting a company's prosperity and financial gains. Recently, there have been more and more organizations that opt for alternative means of protecting intellectual property, like Creative Commons. But many companies, especially the 'big tech' – Google, Facebook, Twitter, Amazon or Microsoft – have built their business on different organizational models. And many start-ups tend to follow their way.

One could reasonably build a case for showing that today's intellectual property regime is unjustifiably harsh, especially for consumers and small artists. Despite this, because of the massive lobbying led by creative and technologies industries, the situation will not probably change too soon.

Moreover, even if a company would disclose its proprietary algorithm to the public, the consequences could be minimal in terms of restoring trust. Users, policymakers or other regulatory bodies are unlikely to understand what an algorithm does or how it works, and this would be especially the case when the algorithm undergoes continuous change over time in reaction to new data inputs. And this leads to the forth practical argument against transparency. While the first three arguments refer directly to the possible harms resulting from making algorithms transparent, the following one emphasizes the fact that the benefits resulting from disclosure might be minimal.

2.4 Sophisticated algorithms are inherently opaque

The fourth argument against transparency, which has already become a common place in the literature, is that machine learning algorithms are essentially 'black boxes' for both experts and users alike (Pasquale, 2015; Domingos, 2015). What do we mean by opacity in regard to algorithms? Algorithms and data are co-evolutionary mechanisms: by crunching datasets, algorithms produce an output, meaning a classification (for example, if a citizen has high chances of being a terrorist, if she receives a loan or an insurance and so on). Opacity refers to the fact that the subject of the output of the algorithms does not know how or why that algorithm has arrived at that particular result. Very often, neither the authorities, nor the individuals understand why someone has been denied a loan or has been targeted as a potential criminal or terrorist, because some types of algorithms (machine learning and neural nets) are inherently opaque (O'Neil, 2016).

In a very interesting article, Burrell (2016) draws a distinction between three forms of opacity. The first one, refers to opacity as a form of proprietary protection or as 'corporate secrecy' (which we discussed in the previous section). The second type is called "opacity as technical illiteracy" and has in view the difficulty of reading, interpreting or understanding code. Programming is a highly specialized technical skill not widely available or easily understood by the large public. Finally, the third type of opacity refers to the complexity and scale of machine learning algorithms and the possibility of understanding the algorithms in action, operating on data. Algorithmic decision-making is a highly complex process which sometimes means that even technical knowledge is not enough for understanding it. In this case, making the algorithm available for public scrutiny would simply be inefficient and ineffective.

3. SOME PHILOSOPHICAL CONCERNS

The issue of trust in online environments is one of utmost importance if we take into account the fact that we are contemporaries with an ongoing economic revolution. Just like during the Agricultural and Industrial revolution, Michael Munger (2018) argues that we are currently

experiencing a transformation of similar proportions towards something that people are already labelling the "sharing economy". While it might seem that a company like Uber is in the business of selling something similar to taxi rides, Munger argues that both Uber and other similar market players like Amazon and Airbnb sell something different: a reduction in transaction costs. The first two elements of the transaction costs involved are Triangulation (providing key information regarding the price of a transaction and the identity of sellers and buyers) and Transfer (effectively transferring payment). Reducing the costs associated with Triangulation and Transfer would be in vain without generating trust between buyers and sellers. Uber and similar companies use outsourcing techniques in order to generate honesty and the trust that a contract will be respected. Reducing the costs associated with generating trust will be, according to recent trends, one of the main goals that businesses will try to achieve.

Following Onora O'Neill (2002) in her deep and profound analysis of the link between transparency and trust, we claim that transparency is not the unconditional good it is sometimes assumed to be. Ideals of transparency and openness are today taken for granted. The main assumption is that transparency destroys secrecy which is oftentimes taken as an indicator of hidden motives or intentions and implicitly, deception. As such, transparency is seen as a universal panacea for resolving multiple issues connected to corruption, fraud, discrimination and so on. But O'Neill cautions us that: "while transparency might destroy secrecy, it may not limit the deception and deliberate misinformation that undermine relations of trust. If we want to restore trust we need to reduce deception rather than secrecy" (O'Neill, 2002, 68).

Transparency means disclosing the inner workings of a system, making it available for public scrutiny. Analogously, making algorithms transparent would entail making their inner workings (implicitly, the databases they are crunching) available for assessment and critique by everybody. Although in theory this might sound like an ideal solution for the problems of algorithmic decision-making that we had to face recently, such an attempt might have unforeseen circumstances. For example, it might produce „a flood of unsorted information that provides little but confusion unless it can be assessed" (2002, 68), a particularly difficult task when the large public lacks the necessary technical skills.

As a consequence, transparency could be highly ineffective in a world that experiences the so-called "digital divide". With the advancement and widespread adoption of information and communication technologies, economists and sociologists have written extensively on the gaps and inequalities both between individuals and nations when it comes to access to computers but also, more recently, to the internet (Pick & Sarkar, 2015). The unequal possession of digital skills is also a recurring topic of interest in the scholarly debate surrounding the digital divide. For example, van Deursen and van Dijk (2010, 895) distinguish between four types of skills associated with using ICT and the internet: (a) operation skills (basic skills associated with using the internet); (b) formal skills (being able to navigate within the hypermedia structure of the internet); (c) information internet skills (the skills required in order for a user to fulfill her information needs) and (d) strategic internet skills (they refer to the "capacity to use the internet as a means of reaching particular goals and for the general goal of improving one's position in society. The emphasis lies on the procedure through which decision-makers can reach an optimal solution as efficiently as possible").

If individuals possess this skills in an unequal way, and recurring empirical work has shown this time and time again, then we are faced both with an ethical problem (fairness in skill distribution) but also a pragmatic one (trust cannot be achieved through transparency while the digital divide still lingers). Some proposals to mitigate this problem could be advanced, but it would involve a discussion far beyond our purpose for this paper. For example, radical forms of cognitive enhancement as a path for achieving superintelligence and cognitive superpowers are available on the market of ideas (Bostrom, 2014), but the debate surrounding whether or not such radical transformations should be mandatory is still open.

Unconditional access to information is not a necessary prerequisite for building relations of trust; what is needed for this purpose is to be able to cross check or verify the information received. In other words, "trust grows out of active inquiry, rather than blind acceptance" (O'Neill, 2002, 70).

REFERENCES

- Bostrom, N. (2014). *Superintelligence. Paths, Dangers, Strategies*. Oxford: Oxford University Press.
- Brauneis, R., Goodman, E. P. (2017). Algorithmic Transparency for the Smart City. *The Yale Journal of Law and Technology*, 20, 103-176.
- Burrell, J. (2016). How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data & Society* 3 (1): 2053951715622512. <https://doi.org/10.1177/2053951715622512>.
- Dastin, J. (2018). Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women. *Reuters*. Retrieved October 22, 2018, from <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women idUSKCN1MK08G>.
- Diakopoulos, N., Koliska, M. (2017). Algorithmic Transparency in the News Media. *Digital Journalism*, 5(7), 809–28.
- Domingos, P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. 1 edition. New York: Basic Books.
- Laat, Paul B. de. (2017). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?. *Philosophy & Technology*, November, 1–17. <https://doi.org/10.1007/s13347-017-0293-z>.
- Lapowsky, I. (2018). Google Autocomplete Still Has a Hitler Problem. *Wired*. Retrieved February 12, 2018, from <https://www.wired.com/story/google-autocomplete-vile-suggestions/>.
- Lopez-Cozar, E. D., Robinson-Garcia, N., Torres-Salinas, D. et al. (2012). Manipulating Google Scholar Citations and Google Scholar Metrics: Simple, Easy and Tempting. *ArXiv:1212.0638 [Cs]*, December. Retrieved from <http://arxiv.org/abs/1212.0638>.
- Mansfield-Devine, S. (2015). The Ashley Madison Affair. *Network Security*, (9): 8–16.
- Morris, N., Vines, D. (2014). *Capital Failure: Rebuilding Trust in Financial Services*. Oxford University Press.
- Munger, M. (2018). *Tomorrow 3.0. Transaction Costs and the Sharing Economy*. Cambridge: Cambridge University Press.
- O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.
- O'Neill, O. (2002). *A Question of Trust: The BBC Reith Lectures 2002*. Cambridge University Press. Retrieved from https://books.google.ro/books?id=h_rTsfy4srQC.
- Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press. Retrieved from <https://books.google.ro/books?id=ll3rBQAAQBAJ>.
- Pick, J. B., Sarkar, A. (2015). *The Global Digital Divides. Explaining Change*. Berlin: Springer.
- Plummer, L. (2017). YouTube Cracks down on Ads Placed on 'Hate Speech' Videos. *Wired UK*, Retrieved June 2, 2017, from <https://www.wired.co.uk/article/youtube-hate-speech-advert-crackdown>.
- van Deursen, A., van Dijk, J. (2010). Internet skills and the digital divide. *New Media & Society*, 13(6), 893–911.
- Zarsky, T. (2016). The Trouble with Algorithmic Decisions An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *Science, Technology & Human Values* 41 (1): 118–32. <https://doi.org/10.1177/0162243915605575>.