Philos. Technol.
DOI 10.1007/s13347-017-0279-x

RESEARCH ARTICLE

Fair, Transparent, and Accountable Algorithmic Decision-making Processes

The Premise, the Proposed Solutions, and the Open Challenges

Bruno Lepri[1] · Nuria Oliver[2,3] · Emmanuel Letouzé[3,4] · Alex Pentland[3,4] · Patrick Vinck[3,5]

Received: 18 February 2017 / Accepted: 6 August 2017

Abstract The combination of increased availability of large amounts of fine-grained human behavioral data and advances in machine learning is presiding over a growing reliance on algorithms to address complex societal problems. Algorithmic decision-making processes might lead to more objective and thus potentially fairer decisions than those made by humans who may be influenced by greed, prejudice, fatigue, or hunger. However, algorithmic decision-making has been criticized for its potential to enhance discrimination, information and power asymmetry, and opacity. In this paper, we provide an overview of available technical solutions to enhance fairness, accountability, and transparency in algorithmic decision-making. We also highlight

✉ Bruno Lepri
lepri@fbk.eu

Nuria Oliver
nuria.oliver@vodafone.com

Emmanuel Letouzé
eletouze@datapopalliance.org

Alex Pentland
pentland@mit.edu

Patrick Vinck
pvinck@hsph.harvard.edu

[1] Fondazione Bruno Kessler, via Sommarive 18, Trento, Italy

[2] Vodafone Research, London, UK

[3] Data-Pop Alliance, New York, NY, USA

[4] MIT Media Lab, Cambridge, MA, USA

[5] Harvard Humanitarian Initiative, Cambridge, MA, USA

Published online: 15 August 2017

the criticality and urgency to engage multi-disciplinary teams of researchers, practitioners, policy-makers, and citizens to co-develop, deploy, and evaluate in the real-world algorithmic decision-making processes designed to maximize fairness and transparency. In doing so, we describe the Open Algortihms (OPAL) project as a step towards realizing the vision of a world where data and algorithms are used as lenses and levers in support of democracy and development.

Keywords Algorithmic decision-making · Algorithmic transparency · Fairness · Accountability · Social good

1 Introduction

Today's vast and unprecedented availability of large-scale human behavioral data is profoundly changing the world we live in. Massive streams of data are available to train algorithms which, combined with increased analytical and technical capabilities, are enabling researchers, companies, governments, and other public sector actors to resort to data-driven machine learning-based algorithms to tackle complex problems (Gillespie 2014; Willson 2016). Many decisions with significant individual and societal implications previously made by humans alone—often by experts—are now made or assisted by algorithms, including hiring (Chalfin et al. 2016), lending (Khandani et al. 2010), policing (Wang et al. 2013), criminal sentencing (Barry-Jester et al. 2015), and stock trading (Kearns and Nevmyvaka 2013). Data-driven algorithmic decision making may enhance overall government efficiency and public service delivery, by optimizing bureaucucratic processes, providing real-time feedback and predicting outcomes (Sunstein 2012). In a recent book with the evocative and provocative title "Technocracy in America," international relations expert Parag Khanna argued that a data-driven direct technocracy is a superior alternative to today's (alleged) representative democracy, because it may dynamically capture the specific needs of the people while avoiding the distortions of elected representatives and corrupt middlemen (Khanna 2017). Human decision making has often shown significant limitations and extreme bias in public policy, resulting in inefficient and/or unjust processes and outcomes (Akerlof and Shiller 2009; Fiske 1998; Samuelson and Zeckhauser 1988; Tverksy and Kahnemann 1974). The turn towards data-driven algorithms can be seen as a reflection of a demand for greater objectivity, evidence-based decision-making, and a better understanding of our individual and collective behaviors and needs.

At the same time, scholars and activists have pointed to a range of social, ethical and legal issues associated with algorithmic decision-making, including bias and discrimination (Barocas and Selbst 2016; Sweeney 2013), and lack of transparency and accountability (Citron and Pasquale 2014; Pasquale 2015; Zarsky 1989, 2016). For example, Barocas and Selbst (2016) showed that the use of algorithmic decision making processes could result in disproportionate adverse outcomes for disadvantaged groups, in ways suggestive of *discrimination*. Algorithmic decisions can reproduce and magnify patterns of discrimination, due to decision makers' prejudices (Pager and Shepherd 2008), or reflect the biases present in the society (Pager and Shepherd 2008). A recent study by ProPublica of the COMPAS Recidivism Algorithm (an algorithm used to inform criminal sentencing decisions by predicting recidivism) found that the algorithm was significantly more likely to label black defendants than white defendants, despite similar overall rates of prediction accuracy between the two groups (Angwin et al. 2016). Along this line, a nominee for the National Book Award, Cathy O'Neil's book, "Weapons of Math Destruction," details several case studies on harms and risks to public accountability associated with big data-driven algorithmic decision-making, particularly in the areas of criminal justice and education (O'Neil 2016).

In 2014, the White House released a report titled "Big Data: Seizing opportunities, preserving values" (Podesta et al. 2014) highlighting the discriminatory potential of Big Data, including how it could undermine longstanding civil rights protections governing the use of personal information for credit, education, health, safety, employment, etc. For example, data-driven algorithmic decisions about applicants for jobs, schools or credit may be affected by hidden biases that tend to flag individuals from particular demographic groups as unfavorable for such opportunities. Such outcomes can be self-reinforcing, since systematically reducing individuals' access to credit, employment and education will worsen their situation, and play against them in future applications. For this reason, a subsequent White House report called for "equal opportunity by design" as a guiding principle in those domains (Munoz et al. 2016). Furthermore, the White House Office of Science and Technology Policy, in partnership with Microsoft Research and others, has co-hosted several public symposiums on the impacts and challenges of algorithms and Artificial Intelligence, specifically relating to social inequality, labor, healthcare and ethics.[1]

At the heart of the matter is the fact that technology outpaces policy in most cases; here, governance mechanisms of algorithms have not kept pace with technological development. Several researchers have recently argued that current control frameworks are not adequate for situations in which a potentially unfair or incorrect decision is made by a computer (Barocas and Selbst 2016).

Fortunately, there is increasing awareness of the detrimental effects of discriminatory biases and opacity of some data-driven algorithmic decision-making systems, and of the need to reduce or eliminate them. A number of research and advocacy initiatives are worth noting, including the Data Transparency Lab,[2] a "community of technologists, researchers, policymakers, and industry representatives working to advance online personal data transparency through research and design," and the DARPA Explainable Artificial Intelligence (XAI) project.[3] A tutorial on the subject was held at the 2016 ACM Knowledge and Data Discovery conference (Hajian et al. 2016). Researchers from New York University's Information Law Institute—such as Helen Nissenbaum and Solon Barocas—and Microsoft Research—such as Kate Crawford and Tarleton Gillespie—have held several workshops and conferences

[1] https://www.whitehouse.gov/blog/2016/05/03/preparing-future-artificial-intelligence

[2] http://www.datatransparencylab.org/

[3] http://www.darpa.mil/program/explainable-artificial-intelligence

these past few years on the ethical and legal challenges related to algorithmic governance and decision-making.[4] Lepri et al. (2017) recently discussed the need for *social good decision-making algorithms* (i.e., algorithms strongly influencing decision-making and resource optimization of public goods, such as public health, safety, access to finance and fair employment) to provide transparency and accountability, to only use personal information—created, owned and controlled by individuals—with explicit consent, to ensure that privacy is preserved when data is analyzed in aggregated and anonymized form, and to be tested and evaluated *in context* by means of living lab approaches involving citizens.

In this paper, we focus on two of the main risks (namely, discrimination and lack of transparency) posed by data-driven predictive models leading to decisions that impact the daily lives of millions of people. There are additional challenges that we do not discuss in this paper. For example, issues relating to data ownership, privacy, informed consent and limited understanding (literacy) about algorithms' abilities and resulting risks among the general public are not discussed here. Instead, focusing on discrimination and lack of transparency, we provide the readers with a review of recent attempts at making algorithmic decision-making more fair and accountable, highlighting the merits and the limitations of these approaches. Finally, we turn to the description of a recent project, called Open Algorithms (OPAL), whose goal is to enable the design, implementation and monitoring of development policies and programs, accountability of government actions, and citizen engagement while leveraging the availability of large-scale human behavioral data in a privacy-preserving and predictable manner.

2 Discriminatory Effects of Algorithmic Decision-making

From a legal perspective, (Tobler 2008) described discrimination as "the application of different rules or practices to comparable situations, or of the same rule or practice to different situations". Barocas and Selbst (2016) argued that discrimination may be an artifact of the data collection and analysis process itself. More specifically, even with the best intentions, data-driven algorithmic decision-making can lead to discriminatory practices and outcomes: algorithmic decision procedures can reproduce existing patterns of discrimination, inherit the prejudice of prior decision makers, or simply reflect the widespread biases that persist in society (Crawford and Schultz 2014). Some have argued it could exacerbate prevailing inequalities by suggesting that historically disadvantaged groups actually deserve less favorable treatment (O'Neil 2016).

Algorithmic discrimination may arise from different sources. First, input data into algorithmic decisions may be poorly weighted, leading to *disparate impact*. For example, as a form of *indirect discrimination*, overemphasis of zip code within predictive policing algorithms can lead to the association of low-income African-American neighborhoods with areas of crime and as a result, the application of

[4] http://www.law.nyu.edu/centers/ili/algorithmsconference

specific targeting based on group membership (Christin et al. 2015). Second, discrimination can occur from the decision to use an algorithm itself. Categorization can be considered as a form of *direct discrimination*, whereby algorithms are used for disparate treatment (Diakopoulos 2015). Third, algorithms can lead to discrimination as a result of the misuse of certain models in different contexts (Calders and Zliobaite 2013). Fourth, in a form of feedback loop, biased training data can be used both as evidence for the use of algorithms and as proof of their effectiveness (Calders and Zliobaite 2013).

The use of algorithmic data-driven decision processes may also result in individuals being denied opportunities based not on their own action but on the actions of others with whom they share some characteristics. For example, some credit card companies have lowered a customer's credit limit, not based on the customer's payment history, but rather based on analysis of other customers with a poor repayment history that had shopped at the same establishments where the customer had shopped (Ramirez et al. 2016).

These are both old and new risks. There is ample evidence of detrimental impacts of current non-algorithmic approaches to access to finance, employment, and housing. Backgrounds checks for example are widely used in those procedures, with people's knowledge and consent. But hundreds of thousands of people have been treated unfairly as a result of mistakes (for instance, misidentification) in the procedures used by external companies to perform background checks.[5] On the one hand, the occurrence of such trivial procedural mistakes may be bound to decrease once performed through data-driven methodologies. But the effort required to identify the causes of unfair and discriminative outcomes can be expected to be exponentially larger, as exponentially more complex will be the black-box models employed to assist in the decision-making process. But it also means that should such methodologies not be transparent in their inner workings, the effects are likely to stay though with different roots.

This scenario thus highlights particularly well the need for machine learning models featuring transparency (understood as openness and communication of both the data being analyzed and the mechanisms underlying the models), accountability (understood as the assumption of accepting the responsibility for actions and decisions), and fairness (understood as the lack of discrimination or bias in the decisions).

3 Techniques to Prevent Algorithmic Discrimination and Maximize Fairness

A simple way to try to maximize *fairness*—understood as the lack of bias and discrimination—in machine learning is precluding the use of sensitive attributes (Calders and Verwer 2010; Kamiran et al. 2010; Schermer 2011; Barocas and Selbst

[5] http://www.chicagotribune.com/business/ct-background-check-penalties-1030-biz-20151029-story.html

2016). For example, if we want a race-blind or a gender-blind decision-making process, we may exclude these attributes (i.e., race, gender, etc.) from the process. However, this solution has several technical problems. First, the excluded attributes can often be implicit in non-excluded ones (Pedreschi et al. 2008; Romei and Ruggieri 2014; Zarsky 2016). For example, when race is excluded as a criterion for granting or not a loan, some implicit information can be present in the individual's zip code, given that zip code may be a good proxy for race (Schermer 2011; Macnish 2012). An additional problem with the *blindness approach* was identified by Dwork et al. (2012): a learning decision rule can select the opposite of what is intended. Consider the example provided by Dwork et al. (2012): in a certain culture *S* the most talented students tend to study engineering and science whereas the less talented study finance. However, in another culture *C,* the trend is reversed such that the most talented students are encouraged to study finance and the less talented are guided towards engineering and science. An organization from culture *C* ignorant of these cultural differences may select candidates for "economics," potentially selecting the wrong candidates from culture *S* even while maintaining parity. This is an example of a suboptimal outcome in a "fairness through blindness" approach as the errors are due to ignoring cultural membership.

In the last few years, several researchers have proposed different technical definitions of *fairness* in machine learning, most of which formalize some notion of *group fairness* (Calders and Verwer 2010; Kamishima et al. 2011; Zemel et al. 2012; Feldman et al. 2015). One of the most used notions is *statistical parity*, which requires that an equal fraction of each group should receive each possible outcome (Calders and Verwer 2010; Kamishima et al. 2011; Zemel et al. 2012; Feldman et al. 2015). Recent papers have also considered approximate relaxations of statistical parity, motivated by the formulation of disparate impact in the US legal code (Feldman et al. 2015; Zafar et al. 2015). Work in these directions has also developed learning algorithms that penalize violations of statistical parity (Calders and Verwer 2010; Kamishima et al. 2011).

However, as pointed out by Dwork et al. (2012), the *group fairness* often fails at both accurate learning and actual fairness. The case of lending can be used as an example to make the point: if two groups have different proportions of individuals who are able to pay back their loans, the algorithm's accuracy will suffer when constrained to predict an equal proportions of paybacks for the two groups. Moreover, *group fairness* definitions do not guarantee that a creditworthy individual from one group has an equal probability of receiving a loan as a similarly creditworthy individual from the other group.

To overcome these limitations, Dwork et al. (2012) argued that technical definitions of fairness should focus on *individual fairness*. More precisely, they proposed a framework based on a task-specific externally defined similarity metric between individuals. The goal of this metric is to achieve fairness through the principle that similar people should be treated in a similar way: thus, any two individuals who are similar with respect to a given task should be classified in a similar way (Dwork et al. 2012). The technical definition of similarity between individuals, proposed by Dwork et al. (2012), resembles partly the notion of *strict equality of opportunity* proposed by the political scientist John Roemer (Roemer 1996, 1998). For Roemer, *strict equality* *of opportunity* is achieved when people, irrespectively of circumstances beyond their control, have the same ability to achieve advantage through their free choices.

Following Dwork et al. (2012) and Joseph et al. (2016) have recently proposed a specific definition of individual fairness that can be considered as a mathematical formalization of the Rawlsian principle of “fair equality of opportunity” (Rawls 1971). This principle affirms that those individuals, “who are at the same level of talent and have the same willingness of using it, should have the same perspectives of success regardless their initial place in the social system” (e.g., income, race, etc.) (Rawls 1971). Thus, this principle is stronger than a “formal equality of opportunity”: Rawls, indeed, argued that an individual should not only have the right to opportunities, but also should have an effective equal chance as another individual of similar natural abilities. In their proposed approach, Joseph et al. (2016) include a notion of fairness in a sequential decision-making framework called contextual bandits in the machine learning literature. Their notion of fairness requires that at every step the learning algorithm never favors applicants whose attributes are lower than the ones of another applicant. Hence, their aim is to design a machine learning algorithm that would (provably) converge to an optimal decision while being (provably) fair at every step. They show that learning algorithms can be proven to be fair in such a way that the cost (from the perspective of rate of convergence to an optimal decision) of adding fairness to the algorithm is small.

In a recent work, Hardt et al. (2016) proposed a fairness measure, based on a similar notion of *equality of opportunity*, that tries to achieve two important objectives. First, to overcome the main conceptual shortcomings of statistical parity as a fairness notion. Second, to build higher accuracy classifiers, in line with the central goal of supervised machine learning. To this end, they proposed a criterion for discrimination against a specified sensitive attribute in supervised learning, where the goal is to predict some target based on available features. Assuming the availability of data about the predictor, the target, and the membership in the protected group, they showed how to optimally adjust any learned predictor so as to remove discrimination according to their definition. The proposed framework also changes incentives by shifting the cost of poor classification from disadvantaged groups to the decision maker, who can respond by improving the classification accuracy. They illustrate their approach and compare different fairness measures in the case of FICO scores with the protected attribute of race, i.e., they build machine learning models that aim to predict a credit risk from a number of attributes with race being a protected attribute. In their conclusions, they highlight that with their framework it is possible to measure unfairness rather than to prove fairness and emphasize the importance of having access to reliable target variables, which is not always the case in practical scenarios.

Another interesting result is the one discussed by Kleinberg et al. (2017). In their paper, they formalized three fairness conditions that constitute the heart of the debates about discrimination in machine learning. Moreover, they proved that, except in highly constrained special cases, there is no method that can satisfy these three conditions simultaneously. Specifically, a first condition—known as *calibration within groups* in the literature—is that the probability estimates provided by the decision-making algorithms should be well-calibrated: for example, if the algorithm identifies a set of people as having probability z of constituting positive instances, then approximately a z fraction of this set should be positive instances (Foster and Vohra 1998). In addition, this condition should be valid when applied separately in each group: if we think of potential differences between an outcome z for Afro-Americans and Asians, this means that a z fraction of men and z fraction of women assigned a probability z should possess the property in question. A second condition focuses on the people who constitute positive instances: the average score received by those people should be the same in each group. This represents a *balance for the positive class*: indeed a violation of this condition would mean that people who are positive instances in one group receive consistently lower probability estimates than people constituting positive instances in another group. Let resort to the case study investigated by ProPublica where one of the concerns raised was that white defendants who went on to commit future crimes were assigned risk scores corresponding to lower probability estimates in aggregate. This is an example of a violation of the *balance for the positive class* condition. A similar condition holds with respect to negative instances, which is called *balance for the negative class*. In short, these balance conditions can be considered as generalizations of the notions that both groups should have equal false negative and false positive rates. The authors outline a few lines of future research, including the fact that there might be use cases where the cost of false positives differs greatly from the cost of false negatives and thus it should be taken into account. In line with this approach, Chouldechova (2016) and Corbett-Davies et al. (2017) consider conditions close to the balance for negative and positive classes together with a form of calibration adapted to binary predictions. The calibration requires that for all people given a positive label, the same fraction of people in each group should truly be part of the positive class. Interestingly, they show that no classification rule can satisfy the required constraints. Finally, a paper by Friedler et al. (2016) defines two axiomatic properties of feature generation and shows that no mechanism can be fair under these two properties.

The results obtained by Kleinberg et al. (2017), Chouldechova (2016), and Corbett-Davies et al. (2017) highlight that it is not enough to simply demand *algorithmic fairness*. We may need to investigate deeply and critically each problem and determine which notion of fairness is considered to be the most relevant and meaningful for that particular problem. Critically, what constitutes fairness changes according to different worldviews: for example, the philosopher Robert Nozick in his book "Anarchy, State, and Utopia" (Nozick 1974) proposed a libertarian alternative view where he is concerned that eliminating discrimination biases, present in society, may create new harms to new groups of people. Instead, Dworkin's egalitarian view of fairness (Dworkin 2000) is based on the principle of *equality of resources*.

For this reason, we feel the urgency to extablish a call for action putting together researchers from different fields—including law, ethics, political philosophy and machine learning—to devise, evaluate and validate in the real-world alternative fairness metrics for different tasks. In addition to this empirical research, we believe it will be necessary to propose a modeling framework—supported by empirical evidence—that would assist practitioners and policy makers in making decisions aided by algorithms that are maximally fair.

4 Information Asymmetry and Lack of Transparency

The mandate for accountable algorithms in government and corporations' decision-making tools is fundamental in both validating their utility toward the public interest as well as redressing potential harms generated by these algorithms.

Transparency, which refers to the understandability of a specific model, can be a mechanism that facilitates accountability. More specifically, transparency can be considered at the level of the entire model, at the level of individual components (e.g., parameters), and at the level of a particular training algorithm. In the strictest sense, a model is transparent if a person can contemplate the entire model at once. Thus, models should be characterized by low computational complexity. A second and less strict notion of transparency might be that each part of the model (e.g., each input, parameter, and computation) admits an intuitive explanation (Lou et al. 2012). A final notion of transparency might apply at the level of the algorithm, even without the ability to simulate an entire model or to intuit the meaning of its components.

However, the ability to access and analyze behavioral data about customers and citizens on an unprecedented scale gives corporations and governments powerful means to reach and influence segments of the population through targeted marketing campaigns and social control strategies. In particular, we are witnessing an *information asymmetry* situation where a powerful few have access and use resources and tools that the majority do not have access to, thus leading to an—or exacerbating the existing—asymmetry of power between the state and big companies on one side and the people on the other side (Akerlof 1970), conceptualized as a "new digital divide" (Boyd and Crawford 2012). In addition, the nature and use of various data-driven algorithms for social good, as well as the lack of computational or data literacy among citizens (Bhargava et al. 2015), makes algorithmic transparency difficult to generalize and accountability difficult to assess (Pasquale 2015).

Burrell (2016) has provided a useful framework to characterize three different types of opacity in algorithmic decision-making: (1) *intentional opacity*, whose objective is the protection of the intellectual property of the inventors of the algorithms. This type of opacity could be mitigated with legislation that would force decision-makers towards the use of open source systems. The new General Data Protection Regulations (GDPR) in the EU with a "right to an explanation" starting in 2018 is an example of such legislation.[6] But powerful commercial and governmental interests will make it difficult to eliminate intentional opacity; (2) *illiterate opacity*, due to the fact that the vast majority of people lack the technical skills to understand the underpinnings of algorithms and machine learning models built from data. This kind of opacity might be attenuated with stronger education programs in computational thinking and "algorithmic literacy" and by enabling independent experts to advise those affected by algorithmic decision-making; and (3) *intrinsic opacity*, which arises by the nature of certain machine learning methods that are difficult to

[6] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

interpret (e.g., deep learning models). This opacity is well known in the machine learning community (usually referred to as the *interpretability problem*). The main approach to combat this type of opacity requires using alternative machine learning models that are easy to interpret by humans, despite the fact that they might yield lower accuracy than black-box non-interpretable models.

5 Techniques to Improve Transparency and Accountability

As previously described, algorithmic decision-making might lack transparency. A simple solution to this limitation would consist of asking for transparency and openness of the algorithm's source code as well as inputs and ouputs that are used to make relevant algorithmic decisions. However, transparency alone is not sufficient to provide accountability in all cases. First of all, it is often necessary to keep secret certain elements of an algorithimic decision policy, the way how the policy is implemented, the key inputs, or the outcome. This is a way to help prevent strategic *gaming* of the system. Furthemore, when the decision being regulated is a commercial one—such as a bank decision to give a loan–, a legitimate business interest in protecting proprietary information or algorithms may be incompatible with full transparency. Again, an algorithmic decision-making system may use as input or may create as output sensitive data that should be not shared to protect business interests, privacy, etc. In some domains, such as finance and healthcare, disclosure may be limited by regulations.

A strategy for verifying and making transparent the behavior of an algorithmic decision-making process is *auditing*. An auditing strategy deals with the decision process as a black box, whose inputs and outputs are visible, while inner workings are not (Sandvig et al. 2014). However, several researchers have shown that a black-box evaluation of decision processes and systems is the least powerful of a set of available methods for understanding their behaviors (Datta et al. 2015). Datta et al. (2015) study the *opacity* (i.e., lack of transparency) of web-based ads by means of their AdFisher tool. They ran several experiments to investigate the transparency provided by Google's Ad Settings. In particular, they analyze if visiting webpages related to a certain interest would lead to a change in the ads shown that is not captured in the settings. They found instances of opacity as they encountered cases where there were significant differences in the ads shown to different profiles while their tool failed to show any type of profiling. They attribute the instances of opacity to remarketing or to other causes related to the complexity of Google's massive, automated advertising system. The authors advocate—as we do—for additional research to create machine learning algorithms that automatically provide transparency to the users.

Furthermore, some algorithmic decision-making processes aim to determine variables that are not measurable in a direct way, for example the risk of credit default. For this reason, these values are computed by means of proxy variables such as the consumer's credit history, the consumer's income, some consumer's personal characteristics or even their mobile phone usage patterns (San Pedro et al. 2015). Consumers able to understand these processes would be tempted to control the proxy variables (O'Neil 2016). Thus, secrecy discourages consumers' strategic behaviors and prevents violations of legal restrictions on disclosure of data. In a recent paper, Hardt et al. (2016) have proposed adversarial methods for designing algorithmic decision-making processes that remain robust in the face of gaming. Moreover, for algorithmic decision processes involving some element of randomness, the full transparency of the source code, inputs, operating environment, and results does not exclude the possibility that the process may produce unpredictable results. Finally, systems that change over time cannot be fully understood through transparency alone. For example, online machine learning algorithms can update their model for predictions after each decision, which increases the complexity of a strategy to ensure transparency in the decision-making process.

Ultimately, we need accountability in decision-making algorithms such that there is clarity regarding who holds the responsibility of the decisions made by them or with algorithmic support. Transparency is generally thought as a key enabler of accountability. However, transparency and auditing do not necessarily suffice for accountability. In fact, in a recent paper, Kroll et al. (2017) have introduced computational methods able to provide accountability even when some information is kept hidden. The authors used advanced techniques to allow the governance of secret algorithmic decision-making processes: specifically, software verification (i.e., a set of techniques for proving mathematically that a piece of software has certain properties), cryptographic commitments (i.e., equivalents of sealed documents held by a third party or in a safe place), zero-knowledge proofs (i.e., cryptographic tools that allow a decision maker, as part of a cryptographic commitment, to prove that the decision policy that was actually used has a certain property, but without revealing either how the property is known or what the decision policy is), and fair random choices (i.e., a technique allowing software that makes random choices to be fully reproducible). These methods can guarantee that both the input data and the software that analyzes such data satisfy the requirements for procedural regularity, even when they are kept secret.

Another approach to provide transparency in algorithmic decision-making entails providing explanations regarding the processes that lead to the decisions such that they are interpretable by humans. In a recent work, Ribeiro et al. (2016) proposed to (i) provide explanations for individual predictions as a solution to the so-called "trusting a prediction" problem, and (ii) select multiple such predictions and explanations as a solution to the so-called "trusting the model" problem. Specifically, they proposed LIME, a novel technique that explains the predictions of any classifier by learning an interpretable model locally around the prediction. They also proposed a method to explain models by showing representative individual predictions and their explanations in a non-redundant way. In the paper, they showed the value of these explanations by means of experiments, both simulated and with human subjects, on several scenarios, such as (i) deciding if one should trust a prediction, (ii) identifying a classifier that should not be trusted, and (iii) choosing between different classification models, etc.

A recent challenging contribution by Lipton (2016) examined the motivations underlying the raising interest in *interpretability*, finding them to be diverse and sometimes discordant. In general, the desire for an interpretation suggests that predictions alone are not sufficient, thus implying a discrepancy between the real-world objectives of machine learning researchers and practictioners and the simple objectives optimized by most machine learning models.

However, the concept of *interpretability* is often vaguely defined. The learning model's properties that enable or that compromise interpretability broadly fall into two categories. The first one relates to *transparency*, that is *how does the model work*. The second one consists of *post-hoc interpretations*, that is *what else can the model tell*.

As distinct notion of interpretability, post-hoc interpretations consist of explanations that need not elucidate the exact process by which models work. These interpretations include natural language explanations (McAuley and Leskovec 2013), visualizations of learned representations or models (e.g., saliency maps in deep neural nets (Simonyan et al. 2013)), and explanations by example (e.g., a tumor is classified as malignant because to the model it looks like these other tumors) (Caruana et al. 1999). One advantage of this concept of interpretability is that we can interpret opaque models after-the-fact, without sacrificing predictive performance.

Algorithmic decision-making processes have the potential to lead to fairer and more objective decisions, grounded in data that are representative of the community where the decisions apply. However, as explained in the previous sections, algorithmic decision-making might lead to discrimination, information asymmetry and lack of transparency. Hence, we believe that it is not only important but also urgent to engage multi-disciplinary teams of researchers, practitioners and policy makers to propose, implement and evaluate in the real-world algorithmic decision-making processes that are designed to maximize their fairness and transparency.

In the next section, we describe one of such proposals, the OPAL project.

6 The OPAL Project

The Open Algorithms (OPAL) project,[7] is a multi-partner socio-technological platform led by Data-Pop Alliance, Imperial College London, the MIT Media Lab, Orange S.A., and the World Economic Forum, that aims to leverage private sector data for public good purposes by "sending the code to the data" in a privacy preserving, predictable, participatory, scalable and sustainable manner. The project came out of the recognition that accessing data held by private companies –including Call Detail Records (CDRs) collected by telecom operators for billing purposes, and banking data– for research and policy purposes has been a conundrum. To date for example, CDRs have been accessed and analyzed either internally, or externally through ad-hoc data challenges or through bilateral arrangements with a limited number of groups under Non-Disclosure Agreements. These types of engagements have offered ample evidence of the promise and demand, but they do not scale nor address some of the most critical challenges discussed above.

[7] http://opalproject.org/

Building on the lessons of the past, OPAL is a key milestone towards a vision where data is at the heart of societal development around the globe, by providing a far better picture of human conditions to official statisticians, policy makers, planners, businesses leaders, and citizens, while enabling greater inclusion and inputs of all members of societies on the kinds and uses of analyses performed on data about themselves. As such, OPAL will reflect and foster the double objective to turn Big Data on its head and "save it from itself" (Pentland 2014).

OPAL's first core feature is technological: the platform allows sending queries (i.e., running algorithms) on the partner companies' servers, behind their firewalls, and not the other way around, so that raw data are never exposed to theft and misuse. The second one is socio-political: it involves co-designing the algorithms so that they are not only open but also serve local needs and respect local standards. For example, a key idea of OPAL is that algorithms should be verified by experts, policy makers, citizens to be as free as possible from biases and unintended side effects such as discrimination. To this end, OPAL makes use of vetting the algorithms that are permitted to run on a given data-set within a specific data repository. Once an algorithm has been vetted, it becomes a template that is digitally signed by the issuers (e.g., expert themselves, public institutions, representatives of the communities that will be affected by algorithmic decisions, etc.). This template algorithm can be shared among a group of entities (e.g., within a consortium) or even be published on a public site. Note that this vetting does not guarantee the quality of the output, which is a function of the quality of the input data.

The OPAL model of moving the algorithm to the data and of using vetting allows a data repository to choose whether or not it is willing to accept a submitted OPAL algorithm (query). If the data repository accepts a given vetted algorithm, it also has the option to impose additional filtering on the resulting data prior to being returned as answer to the querier (e.g., defining the degree of personal information within a given answer). Moreover, each repository can introduce machine learning algorithms as additional mechanisms to protect privacy. Such algorithms allow a repository to detect if multiple accesses from the same entity may result in compromising Personal Identifiable Information (PII).

Finally, blockchain technology can be used to capture and log both vetted-queries and safe-answers, thus providing a mechanism to support post-event audit and accountability. One easy way would be for the querier to compute a cryptographic hash of the query sent, and for the data repository to compute the hash of the response. In addition, the technology may be used by data owners when they want to monetize their data, including the ability to link money and data flows, or micro-payments with small transaction costs.

OPAL is currently being deployed through pilots in Senegal and Colombia, where it has been endorsed by and benefits from the support of their National Statistical Offices and major local telecom operators. Local engagement and empowerment will be central to the development of OPAL: needs, feedback, and priorities have been collected and identified through local workshops and discussions, and their results will feed into the design of future algorithms. These algorithms will be fully open, therefore subject to public scrutiny and redress. A local advisory board is being set up to provide guidance and oversight to the project. In addition, trainings and dialogues will be organized around the project to foster its use and diffusion as well as local capacities and awareness more broadly. OPAL aims to be deployed in 2 more countries by the end of 2018.

Initiatives such as OPAL have the potential to enable more human-centric accountable and transparent data-driven decision-making and governance, by involving a wide range of stakeholders in and through their design and implementation. Note, however, that OPAL does not fully address the issues of algorithmic fairness and transparency. It does not address internal uses of data and potentially discriminatory behavior by corporations. It does not advance an open data agenda, which may be unrealistic when working with corporate-owned data carrying high competititve value and posing severe privacy risks. It also does not in itself enable control for bias in the data itself.

Despite these limitations, we believe that OPAL will provide an avenue for an array of users, from official statisticians to community organizers, to openly query data and have those queries and results examined through a fairness and anti-discrimination lens. It is an important concrete step towards a world where data and algorithms can be leveraged through participatory processes for societal development and democracy around the globe.

7 Conclusions

We live in an unprecedented historic moment where the availability of vast amounts of human behavioral data, combined with advances in machine learning, are enabling us to tackle complex problems through algorithmic decision-making. The opportunities to have positive social impact through fairer and more transparent decisions are paramount. However, algorithmic decision-making processes might lead to discrimination, information asymmetry and lack of transparency.

In this paper, we have provided an overview of both existing limitations and proposed solutions regarding fairness, accountability and transparency in algorithmic decision-making. We have highlighted open challenges that would still need to be addressed and have described the OPAL project as an exemplary effort that aims to maximize algorithmic fairness and transparency to support decision-making for social good. We would like to emphasize the importance and the urgency to engage multi-disciplinary teams of researchers, practitioners, and policy makers to propose, implement, and evaluate in the real-world algorithmic decision-making processes that are designed to maximize their fairness and transparency.

The opportunity to significantly improve the processes leading to decisions that affect millions of lives is huge. As researchers and citizens, we believe that we should not miss on this opportunity. Hence, we would like to encourage the larger community—e.g., researchers, practitioners, policy makers—in a variety of fields—e.g., computer science, sociology, economics, ethics, law—to join forces so we can address today’s limitations in data-driven decision-making and contribute to fairer and more transparent decisions with clear accountability and significant positive impact.

References

Akerlof, G. (1970). The market for lemons: quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, *84*(3), 488–500.

Akerlof, G., & Shiller, R. (2009). *Animal spirits: how human psychology drives the economy, and why it matters for global capitalism*. Princeton: Princeton University Press.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barocas, S., & Selbst, A. (2016). Big data's disparate impact. *California Law Review*, *104*, 671–732.

Barry-Jester, A.M., Casselman, B., & Goldstein, D. (2015). *The new science of sentencing*. The Marshall Project. https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing.

Bhargava, R., Deahl, E., Letouzé, E., Noonan, A., Sangokoya, D., & Shoup, N. (2015). *Beyond data literacy: reinventing community engagement and empowerment in the age of data*. Data-Pop Alliance White Paper Series. http://datapopalliance.org/wp-content/uploads/2015/11/Beyond-Data-Literacy-2015.pdf.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication, & Society*, *15*(5), 662–679.

Burrell, J. (2016). How the machine thinks: understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 1–12.

Calders, T., & Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, *21*(2), 277–292.

Calders, T., & Zliobaite, I. (2013). Why unbiased computational processes can lead to discriminative decision procedures. In Custers, B., Calders, T., Schermer, B., & Zarsky, T. (Eds.) *Discrimination and privacy in the information society* (pp. 43–57).

Caruana, R., Kangarloo, H., David, J., Dionisio, N., Sinha, U., & Johnson, D. (1999). Case-based explanation of non-case-based learning methods. In *Proceedings of the 1999 american medical informatics association (AMIA) symposium* (pp. 212–215).

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review*, *106*(5), 124–127.

Chouldechova, S. (2016). Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. arXiv:1610.07524.

Christin, A., Rosenblatt, A., & Boyd, D. (2015). *Courts and predictive algorithms*. Data & Civil Rights Primer.

Citron, D., & Pasquale, F. (2014). The scored society. *Washington Law Review*, *89*(1), 1–33.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Fair algorithms and the equal treatment principle. Working Paper.

Crawford, K., & Schultz, J. (2014). Big data and due process: toward a framework to redress predictive privacy harms. *Boston College Law Review*, *55*(1), 93–128.

Datta, A., Tschantz, M.C., & Datta, A. (2015). Automated experiments on ad privacy settings. In *Proceedings on privacy enhancing technologies* (pp. 92–112).

Diakopoulos, N. (2015). *Algorithmic accountability: journalistic investigation of computational power structures*. Digital Journalism.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness throug awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214–226). New York: ACM.

Dworkin, R. (2000). *Sovereign virtue: the theory and the practice of equality*. Cambridge: Harvard University Press.

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 259–268).

Fiske, S. (1998). Stereotyping, prejudice, and discrimination. In Gilbert, D., Fiske, S., & Lindzey, G. (Eds.) *Handbook of social psychology* (pp. 357–411). Boston: McGraw-Hill.

Foster, D., & Vohra, R.V. (1998). Asymptotic calibration. *Biometrika*, *85*(2), 379–390.

Friedler, S.A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. arXiv:1609.07236.

Gillespie, T. (2014). The relevance of algorithms. In Gillespie, T., Boczkowski, P., & Foot, K. (Eds.) *Media technologies: essays on communication, materiality, and society* (pp. 167–193). Cambridge: MIT Press.

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: from discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2125–2126). New York: ACM.

Hardt, M., Megiddo, N., Papadimitriou, C., & Wootters, M. (2016). Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science* (pp. 111–122). New York: ACM.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Proceedings of the international on advances in neural information processing systems (NIPS)* (pp. 3315–3323).

Joseph, M., Kearns, M., Morgenstern, J., Neel, S., & Roth, A. (2016). Rawlsian fairness for machine learning. arXiv:1610.09559.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. In *Proceedings of 2010 IEEE international conference on data mining* (pp. 869–874). Washington, DC: IEEE.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2011). Fairness-aware classifier with prejudice remover regularizer. In: *Proceedings of the european conference on machine learning and principles of knowledge discovery in databases (ECMLPKDD), Part II* (pp. 35–50).

Kearns, M., & Nevmyvaka, Y. (2013). Machine learning for market microstructure and high frequency trading. In O'hara, M., Lopez de prado, M., & Easley, D. (Eds.) *High frequency trading*. London: Risk books.

Khandani, A.E., Kim, A.J., & Lo, A.W. (2010). Consumer credit risk models via machine-learning algorithms. *Journal of Banking and Finance*, *34*, 2767–2787.

Khanna, P. (2017). *Technocracy in America: rise of the info-state*. CreateSpace Independent Publishing Platform.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th innovations in theoretical computer science conference*. New York: ACM.

Kroll, J.A., Huey, J., Barocas, S., Felten, E.W., Reidenberg, J.R., Robinson, D.G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, *165*, 633–707.

Lepri, B., Staiano, J., Sangokoya, D., Letouzé, E., & Oliver, N. (2017). The tyranny of data? The bright and dark sides of data-driven decision-making for social good. arXiv:1612.00323.

Lipton, Z.C. (2016). The mythos of model interpretability. In *2016 ICML workshop on human interpretability in machine learning*.

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2012). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 623–631). New York: ACM.

Macnish, K. (2012). Unblinking eyes: the ethics of automating surveillance. *Ethics and Information Technology*, *14*(2), 151–167.

McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on recommender systems*.

Munoz, C., Smith, M., & Patil, D. (2016). *Big data: a report on algorithmic systems, opportunity, and civil rights*. Tech. rep., Executive Office of the President.

Nozick, R. (1974). *Anarchy, state, and utopia*. New York: Basic Books.

O'Neil, C. (2016). *Weapons of math destruction: how big data increases inequality and threatens democracy*. Crown.

Pager, D., & Shepherd, H. (2008). The sociology of discrimination: racial discrimination in employment, housing, credit and consumer market. *Annual Review of Sociology*, *34*, 181–209.

Pasquale, F. (2015). *The Black Blox Society: the secret algorithms that control money and information*. Cambridge: Harvard University Press.

Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 560–568).

Pentland, A. (2014). Saving big data from itself. *Scientific American*, *311*(2), 64–67.

Podesta, J., Pritzker, P., Moniz, E., Holdren, J., & Zients, J. (2014). *Big data: seizing opportunities, preserving values*. Tech. rep., Executive Office of the President.

Ramirez, E., Brill, J., Ohlhausen, M., & McSweeny, T. (2016). *Big data: a tool for inclusion or exclusion*? Tech. rep., Federal Trade Commission.

Rawls, J. (1971). *A theory of justice*. Cambridge: Harvard University Press.

Ribeiro, M., Singh, S., & Guestrin, C. (2016). Why should I trust you?: explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Roemer, J.E. (1996). *Theories of distributive justice*. Cambridge: Harvard University Press.

Roemer, J.E. (1998). *Equality of opportunity*. Cambridge: Harvard University Press.

Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, *29*(5), 582–638.

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*(1), 7–59.

San Pedro, J., Proserpio, D., & Oliver, N. (2015). Mobiscore: towards universal credit scoring from mobile phone data. In *Proceedings of the international conference on user modeling, adaptation and personalization (UMAP)* (pp. 195–207).

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: research methods for detecting discrimination on internet platforms. In *Data and discrimination: converting critical concerns into productive inquiry, a preconference at the 64th annual meeting of the international communication association*.

Schermer, B.W. (2011). The limits of privacy in automated profiling and data mining. *Computer Law & Security Review*, *27*(1), 45–52.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv:1312.6034.

Sunstein, C. (2012). *Regulation in an uncertain world*. National Academy of Sciences. https://www.whitehouse.gov/sites/default/files/omb/inforeg/speeches/regulation-in-an-uncertain-world-06202012.pdf.

Sweeney, L. (2013). *Discrimination in online ad delivery*. Available at SSRN: http://ssrn.com/abstract=2208240.

Tobler, C. (2008). *Limits and potential of the concept of indirect discrimination*. Tech. rep., European Network of Legal Experts in Anti-Discrimination.

Tverksy, A., & Kahnemann, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, *185*(4157), 1124–1131.

Wang, T., Rudin, C., Wagner, D., & Sevieri, R. (2013). Learning to detect patterns of crime. In *Machine learning and knowledge discovery in databases* (pp. 515–530). Springer.

Willson, M. (2016). *Algorithms (and the) everyday*. Information, Communication & Society.

Zafar, M.B., Martinez, I.V., Rodriguez, M.D., & Gummadi, K.P. (2015). Learning fair classifiers. arXiv:1507.05259.

Zarsky, T. (1989). Automated prediction: perception, law and policy. *Communications of the ACM*, *4*, 167–186.

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, and Human Values*, *41*(1), 118–132.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2012). Learning fair representation. In *Proceedings of the 2013 international conference on machine learning (ICML)* (pp. 325–333).