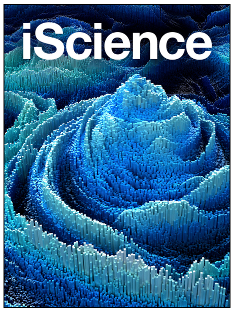

Ethical Machines: The Human-centric Use of Artificial Intelligence

B. Lepri, N. Oliver, A. Pentland

PII: S2589-0042(21)00217-0

DOI: https://doi.org/10.1016/j.isci.2021.102249

Reference: ISCI 102249

To appear in: *ISCIENCE*


Black-box models
Privacy violations
Bias and Discrimination
Human-Centric AI
Algorithmic transparency
Human understandable explanations
Privacy-preserving algorithms
Data Cooperatives
Algorithmic fairness
Risks
Requirements
Icons MIT licensed from: icons.getbootstrap.com. Copyright (c) 2019-2020 The Bootstrap Authors

Ethical Machines: The Human-centric Use of Artificial Intelligence

B. Lepri[1,3,5,⋆], N. Oliver[2,3], and A. Pentland[4,3]

[1]Digital Society Center, Fondazione Bruno Kessler, Trento, 38123, Italy
[2]ELLIS (the European Laboratory for Learning and Intelligent Systems) Unit Alicante, Alicante, 03690, Spain
[3]Data-Pop Alliance, New York, NY, USA
[4]MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA
⋆Correspondance: lepri@fbk.eu

Summary

Today's increased availability of large amounts of human behavioral data and advances in Artificial Intelligence are contributing to a growing reliance on algorithms to make consequential decisions for humans, including those related to access to credit or medical treatments, hiring, etc. Algorithmic decision-making processes might lead to more objective decisions than those made by humans who may be influenced by prejudice, conflicts of interest, or fatigue. However, algorithmic decision-making has been criticized for its potential to lead to privacy invasion, information asymmetry, opacity, and discrimination. In this paper, we describe available technical solutions in three large areas that we consider to be of critical importance to achieve a human-centric AI: (1) privacy and data ownership; (2) accountability and transparency; and (3) fairness. We also highlight the criticality and urgency to engage multi-disciplinary teams of researchers, practitioners, policy makers, and citizens to co-develop and evaluate in the real-world algorithmic decision-making processes designed to maximize fairness, accountability and transparency while respecting privacy.

Introduction

Nowadays, the large-scale availability of human behavioral data and the increased capabilities of Artificial Intelligence (AI) are enabling researchers, companies, practitioners and governments to leverage machine learning algorithms to address important problems in our societies (Gillespie 2014, Willson 2017). Notable examples are the use of algorithms to estimate and monitor socio-economic conditions (Eagle et al. 2010, Soto et al. 2011, Blumenstock et al. 2015, Venerandi et al. 2015, Steele et al. 2017) and well-being (Hillebrand et al. 2020), to map the spread of infectious diseases (i.e. influenza, malaria, dengue, zika and more recently SARS-CoV-2) (Ginsberg et al. 2009, Wesolowski et al. 2012, 2015, Zhang et al. 2017, Jia et al. 2020, Lai et al. 2020), and to quantify the impact of natural disasters (Ofli et al. 2016, Pastor-Escuredo et al. 2014, Wilson et al. 2016).

Moreover, machine learning algorithms are increasingly used to support humans or even autonomously make decisions with significant impact in people's lives. The main motivation for the use of technology in these scenarios is to overcome the shortcomings of human decision-making. In the last decades, several studies in psychology and behavioral economics have highlighted the significant limitations and biases characterizing the human decision-making process (Tverksy & Kahnemann 1974, Samuelson & Zeckhauser 1988, Fiske 1998). Compared to humans, there are advantages that can hardly be denied in the use of machine learning algorithms: they can perform tasks in a shorter amount of time, they are able to process significantly larger amounts of data than humans can, they don't get tired, hungry, or bored and they are not susceptible to corruption or conflicts of interest (Danziger et al. 2011). Furthermore, the increasing tendency in adopting algorithms can be seen as an answer to the request of a greater objectivity and reduced error in decisions. Thus, it is no suprise to see a growth in the use of machine learning-based systems to decide whether an individual is credit worthy enough to receive a loan (Kleinberg et al. 2017), to identify the best candidates to be hired for a job (Siting et al. 2012, Raghavan et al. 2020) or to be enrolled in a specific university (Marcinkowski et al. 2020), to predict if a convict individual is inclined to re-offend (Berk et al. 2018), to recommend products or content (including news) to consume (Jannach & Adomavicius 2016, Noble 2018, Oyebode & Orji 2020), and so on.

However, researchers from different disciplinary backgrounds and activists have identified a range of social, ethical and legal issues associated with the use of machine learning in decision-making processes, including violations of individuals' privacy (Crawford & Schultz 2014, de Montjoye, Hidalgo, Verleysen & Blondel 2013, de Montjoye et al. 2015, Ohm 2010), lack of transparency and accountability (Citron & Pasquale 2014, Pasquale 2015, Zarsky 2016), and biases and discrimination (Barocas & Selbst 2016, Eubanks 2018, Noble 2018, Benjamin 2019). For example, Barocas and Selbst (Barocas & Selbst 2016) have shown that the use of AI-driven decision-making processes could result in disproportionate adverse outcomes for disadvantaged groups (e.g. minorities, individuals with lower income, etc.). In 2016, the non-profit organization ProPublica analyzed the performance of the COMPAS Recidivism Algorithm, a tool used to inform criminal sentencing decisions by predicting recidivism (Angwin et al. 2016). The results of the conducted analysis found that COMPAS was significantly more likely to label black defendants than white defendants as potential repeat offenders, despite similar rates of prediction accuracy between the two groups (Angwin et al. 2016). More recently, Obermeyer *et al.* (Obermeyer et al. 2019) have shown that an algorithm widely used in the health system exhibits a racial bias. Specifically, for a given risk score this algorithm labels black patients as significantly sicker than white patients. As authors pointed out the racial bias arises because the algorithm is predicting health care costs rather than the health status of the individual.

As a consequence, national governments and international organizations (e.g. the European Commission and the European Parliament, the Organisation for Economic Cooperation and Development, etc.), major tech companies (e.g. Google, Amazon, Facebook, Microsoft, IBM, SAP, etc.), and professional and non-profit organizations (e.g. Association for Computing Machinery, Institute of Electrical and Electronics Engineers, World Economic Forum, Amnesty International, etc.) have recently responded to these concerns by extablishing ad-hoc initiatives and committees of experts. These initiatives and committees have produced reports and guidelines for an ethical AI. In a recent paper, Jobin *et al.* (Jobin et al. 2019) have analyzed these guidelines showing that a global convergence is emerging around five ethical principles, namely *transparency*, *justice* and *fairness*, *non-maleficence*, *responsibility*, and *privacy*.

Similarly, the human-computer interaction (HCI) research community has proposed, for over two decades, principles and guidelines for the design of an effective human interaction with AI systems (Norman 1994, Horvitz 1999, Parise et al. 1999, Sheridan & Parasuraman 2005, Lim et al. 2009). Nowadays, this debate is becoming more and more relevant given the growing use of AI systems in decision-making processes (Lee et al. 2015, Abdul et al. 2018, Amershi et al. 2019, Wang et al. 2019). In a recent paper, Amershi *et al.* (Amershi et al. 2019) have sistematically validated a large number of applicable guidelines for designing the interaction between humans and AI systems. Examples of these guidelines (Amershi et al. 2019) are (i) making clear what the system can do and (ii) how well, (iii) supporting an efficient correction of the system's errors and (iv) an efficient dismissal of undesired AI system's services, (v) mitigating the social biases and (vi) matching relevant social norms, and so on. Along this line, Abdul *et al.* (Abdul et al. 2018) have performed a literature analysis of HCI core papers on explainable systems as well as of related papers from other fields in computer science and cognitive psychology. Their analysis (Abdul et al. 2018) revealed some trends and trajectories for the HCI community in the domain of explainable systems, such as the introduction of rule extraction methods in deep learning (Hailesilassie 2016), the demand for a systematic accountability of the AI systems (Shneiderman 2016), the exploration of interactive explanations (Patel et al. 2011, Krause et al. 2016), and the relevance of the human side of the AI systems' explanations (Doshi-Velez & Kim 2017, Lipton 2018, Miller 2019).

In addition, a recent scientific mass collaboration, involving 160 teams worldwide, evaluated the effectiveness of machine learning models for predicting several life outcomes (e.g. child grade point average, child grit, household eviction, etc.) (Salganik et al. 2020). This work used data from the Fragile Families and Child Wellbeing Study (Reichman et al. 2001). The obtained results have shown serious limitations in predicting life outcomes of individuals. Indeed, the best machine learning predictions were not very accurate and only slightly better than the ones obtained by simple baseline models. Therefore, the authors recommend that policymakers determine whether the predictive accuracy, achievable using machine learning approaches, is adequate for the setting where the predictions will be used, and whether the machine learning models are significantly more accurate than simple statistical analyses or decisions taken by human domain experts (Hand 2006, Rudin 2019). Moreover, the perception of algorithms' decisions, regardless of their actual performance, may significantly influence people's trust in and attitudes toward AI-driven decision-making processes (Lee & Baykal 2017, Lee 2018). In a recent work, Lee (Lee 2018) conducted an online experiment in which study participants read the description of a human or an algorithmic managerial decision. These decisions were based on real-world examples of tasks requiring more "human" skills (e.g. emotional capability, subjective judgement, etc.) or more "mechanical" skills (e.g. processing large amount of data, etc.). The study shows that, with the "mechanical" tasks, human-made and algorithmic decisions were perceived as equally trustworthy and fair, whereas, with the "human" tasks, the algorithmic decisions were perceived as less trustworthy and fair than the human ones. In two qualitative laboratory studies, Lee and Baykal (Lee & Baykal 2017) showed that algorithmic decisions in social division tasks (e.g. allocating limited resources to each individual) were perceived more unfair than decisions obtained as a result of group discussions. In particular, the algorithmic decisions were viewed as unfair when they did not take into account the presence of altruism and other aspects related to the group dynamics (Lee & Baykal 2017).

In this article, we build on our previous work (Lepri et al. 2017, 2018) to first provide a brief compendium of risks (i.e. privacy violations, lack of transparency and accountability, and discrimination and biases) that might arise when consequential decisions impacting people's lives are based on the outcomes of machine learning models. Next, we describe available technical solutions in three large areas that we consider to be of critical importance to achieve a human-centric AI: (1) privacy and data ownership; (2) transparency and accountability; and (3) fairness in AI-driven decision-making processes. We also highlight 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, transparency and accountability while respecting privacy, thus pushing towards an ethical and human use of Artificial Intelligence. Detailed reviews and perspectives on these topics can also be found in several recent publications (Pasquale 2015, Mittelstadt et al. 2016, Veale & Binns 2017, Barocas et al. 2018, Cath et al. 2018, Guidotti et al. 2018, Lipton 2018, Jobin et al. 2019, Brundage et al. 2020, Kearns & Roth 2020).

Our ultimate goal is to document and highlight recent research efforts to reverse the risks of AI when used for decision-making and to offer an optimistic view on how our societies could leverage machine learning decision-making processes to build a *Human-centric AI*, namely a social and technological framework that enhances the abilities of individuals and serves the objectives of human development (Letouzé & Pentland 2018). Note that the proposed *Human-centric AI* framework has not the pragmatic and utilitarian objective of improving trustworthiness and of avoiding improper usage of AI-driven decision-making systems in order to increase their adoption. Instead, our envisioned approach has the ambitious goal of building AI systems that preserve human autonomy, complement the intelligence of individuals, behave transparently and help us to increase the fairness and justice in our societies.

The risks of AI-driven decision-making

The potential positive impact of AI –namely, machine learning-based approaches– to decision-making is huge. However, several risks and limitations of these systems have been highlighted in recent years (Crawford & Schultz 2014, Pasquale 2015, Tufekci 2015, Barocas & Selbst 2016, O'Neil 2016, Lepri et al. 2017, Barocas et al. 2018, Brundage et al. 2020), including violations of people's privacy, lack of transparency and accountability of the algorithms used, and discrimination effects and biases harming the more fragile and disadvantaged individuals in our societies. In this section, we turn our attention to these elements before describing existing efforts to overcome and/or minimize these risks and to maximize the positive impact of AI-driven decision-making.

Computational violations of privacy

The use of AI in decision-making processes often requires the training of machine learning algorithms on datasets that may include sensitive information about people's characteristics and behaviors. Moreover, a frequently overlooked element is that current machine learning approaches, coupled with the availability of novel sources of behavioral data (e.g. social media data, mobile phone data, credit card transactions, etc.), allow the learning algorithm to make inferences about private information that may never have been disclosed.

A well-known study by Kosinski *et al.* (Kosinski et al. 2013) used survey information as ground-truth and data on Facebook "Likes" to accurately predict sexual orientation, ethnic origin, religious and political preferences, personality traits as well as alcohol, drugs, and cigarettes use of over 58,000 volunteers. For example, the simple logistic/linear regression model is able to correctly discriminate between African Americans and Caucasian Americans in 95% of cases, between an homosexual and an heterosexual men in 88% of cases, and between Democrats and Republicans in 85% of cases.

More recently, Wang and Kosinski (Wang & Kosinski 2018) used deep neural networks to extract visual features from more than 35,000 facial images. Then, these features were used with a logistic regression algorithm to classify the sexual orientation of the study participants. The authors show that this simple classifier, using a single facial image, could correctly discriminate between gay and heterosexual men in 81% of cases and between gay and heterosexual women in 71% of cases. Human judges, instead, achieved a much lower classification accuracy, namely 61% for men and 54% for women. As pointed out by the authors (Wang & Kosinski 2018), these findings highlight the threats to the privacy and safety of homosexuals given that companies (e.g. recruitment and advertising companies, banks, insurances, etc.) and governments are increasingly using computer vision algorithms to detect people's traits and attitudes.

Along a similar line, Matz *et al.* introduced a *psychological targeting* approach (Matz et al. 2017) that consists in predicting people's psychological profiles (e.g. Big Five personality traits) from their digital footprints, such as Twitter and Facebook profiles (Quercia et al. 2011, Kosinski et al. 2013, Schwartz et al. 2013, Segalin et al. 2017), mobile phone data (Staiano et al. 2012, de Montjoye, Quoidbach, Robic & Pentland 2013, Chittaranjan et al. 2013, Stachl et al. 2020), credit card transactions (Gladstone et al. 2019) and even 3G/4G/Wifi usage patterns (Park et al. 2018), in order to influence people's behaviors by means of psychologically-driven interventions. This technological approach attracted significant attention in the context of the Facebook-Cambridge Analytica scandal, where millions of Facebook users' personal data and psychological profiles were extracted and used without consent by Cambridge Analytica, a British consulting political firm, mainly acting in the domain of political advertising.

Despite the algorithmic advancements in anonymizing data, several works have shown that is feasible to infer identities from pseudo-anonymized human behavioral traces. For example, de Montjoye *et al.* (de Montjoye, Hidalgo, Verleysen & Blondel 2013, de Montjoye et al. 2015) have demonstrated how unique mobility and shopping behaviors are for each individual. Specifically, the authors have shown that four spatio-temporal points are enough to uniquely identify 95% of people in a pseudo-anonymized mobile phone dataset of 1.5 millions people (de Montjoye, Hidalgo, Verleysen & Blondel 2013) and to identify 90% of people in a pseudo-anonymized credit card transactions dataset of 1 million people (de Montjoye et al. 2015).

Furthermore, since machine learning algorithms were often designed without considering potential adversarial attacks, several recent studies are highlighting their privacy vulnerabilities (Papernot et al. 2016, Song et al. 2019). More precisely, adversarial attacks aim at obtaining private sensitive information about the learning model or the model's training data. For example, the attacks targeting the learning model's privacy include (i) the inference of model's hyperparameters using stealing attacks (Wang & Zhenqiang Gong 2018, Song et al. 2019) and (ii) the inference of model's details using model extraction attacks (Tramér et al. 2016, Song et al. 2019). Regarding data privacy, adversarial attacks may also infer, using membership inference attacks (Shokri et al. 2017, Nasr et al. 2019, Song et al. 2019), whether input examples are used to train the target learning model. Additional adversarial attacks targeting data privacy include covert channel model training attacks (Song et al. 2017, 2019) as well as the adoption of property inference attacks to learn global properties of training data (Ganju et al. 2018, Song et al. 2019). As a consequence, the privacy research community has designed and developed defenses to prevent privacy leakage of the target learning model (Kesarwani et al. 2018, Song et al. 2019) and of the model's training data (Shokri & Shmatikov 2015, Abadi et al. 2016, Hayes & Ohrimenko 2018, Song et al. 2019). However, adversarial attacks raise broader risks for the robustness and the trustworthiness of the machine-learning based systems. A notable example is the attack consisting in pasting stickers on traffic signs to fool the computer vision-based signage recognition module in the autonomous vehicles (Eykholt et al. 2018).

Lack of transparency and accountability

Transparency in corporate and government use of AI-driven decision-making tools is of fundamental importance to identify, measure and redress harms (e.g. privacy harms) and discriminatory effects generated by these algorithms, as well as to validate their value for public interest. Moreover, transparency is generally thought as a mechanism that facilitates *accountability*, namely the clarity regarding who holds the responsibility of the decisions made by AI algorithms or with algorithmic support. For this reason, the General Data Protection Regulation (GDPR) framework, launched in 2018 in the European Union (EU), highlighted a “right to an explanation". See `http://eur-lex.europa.eu/eli/reg/2016/679/oj` for more details on the “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 the free movement of personal data, and Directive 95/46/EC (General Data Protection Regulation)".

In "The Mythos of Model Interpretability" (Lipton 2018), the computer scientist Lipton has identified three different notions of transparency: (i) at the level of the whole learning model (i.e. the entire model can be explained and understood), (ii) at the level of individual components (i.e. each component of the model can be explained and understood), and (iii) at the level of the training algorithm (i.e. only the specific algorithm can be explained and understood without any explanation and understanding of the entire model or of its components).

However, different types of opacity or lack of transparency might emerge in AI-driven decision-making tools (Burrell 2016). For example, Datta *et al.* (Datta et al. 2015) have investigated the trasparency provided by Google's Ad Settings using their AdFisher tool and they have found examples of opacity as they encountered cases where there were significant differences in the ads shown to different user profiles while the AdFisher tool failed to identify any type of algorithmic profiling.

Moreover, the inventor and owner of an AI system could intentionally design an opaque system in order to protect the intellectual property or to avoid the gaming of the system (Burrell 2016). Regarding the latter case, network security applications of machine learning remain opaque in order to be effective in dealing with frauds, spams and scams (Burrell 2016). This *intentional opacity* (Burrell 2016) could be mitigated with legislation interventions in favour of the use of open source AI systems (Diakopoulos 2015, Pasquale 2015). However, these interventions often may collide with the interests of corporations that develop and use these systems. For example, when the algorithmic decision being regulated is a commercial one, a legitimate business interest in protecting the algorithm or the proprietary information may conflict with a request of full trasparency.

The second type of opacity is *illiterate opacity* (Burrell 2016), given that a large fraction of the population currently lacks the technical skills to understand how the machine learning algorithms work and how they build models from input data. This kind of opacity might be attenuated by establishing educational programs for e.g. policy makers, journalists, activists in computational thinking and AI, as well as helping the people affected by machine learning decisions to resort to the advice of independent technical experts.

Finally, certain machine learning algorithms (e.g. deep learning models) are by nature difficult to interpret. This *intrinsic opacity* (Burrell 2016) is well-known in the academic machine learning community and it is usually referred to as the *interpretability problem* (Lipton 2018). The main approach to deal with this type of opacity is to use alternative machine learning models that are easier to interpret by humans in order to characterize the decisions made by the black-box algorithm. However, this approach typically does not provide a perfect model of the black-box algorithm's performance.

Biases and discriminatory effects

In legal terms, *discrimination* occurs when two different rules are applied to similar situations, or the same rule is applied to different situations (Tobler 2008). Turning our attention to the use of machine learning in decision-making processes, discriminatory effects and biases could be the result of the way input data are collected and/or of the learning process itself (Barocas & Selbst 2016, Barocas et al. 2018).

First of all, specific features and attributes may be poorly weighted, thus leading to *disparate impact* (Barocas & Selbst 2016, Barocas et al. 2018). For example, predictive policing algorithms may overemphasize the predictive role of the "zip code" attribute, thus leading to the association of low-income African-American and Latino neighborhoods with areas with high criminality. This example highlights an area of ethical ambiguity in current law, known as *indirect discrimination* (Christin et al. 2015), in which social conditions (such as the neighborhood) plays a role in individual decision making, but the algorithm (or law) imputes these social constraints to choices made by the individual.

As before, biased training data can be used both for training models and for evaluating their predictive performance (Calders & Zliobaite 2013), and machine learning algorithms can lead to discriminatory effects as a result of their misuse in specific contexts (Calders & Zliobaite 2013). Indeed, discrimination may occur from the simple decision of when to use an algorithm, a choice that inevitably excludes consideration of some contextual variables (Diakopoulos 2015).

Moreover, the use of AI-driven decision-making processes may also result in the denial of opportunities and resources to individuals not because of their own actions but due to the actions of other individuals with whom they share some characteristics (e.g. income levels, gender, ethnic origin, neighborhoods, personality traits, etc.) (Lepri et al. 2018).

However, as recently argued by Kleinberg *et al.* (Kleinberg et al. 2020), the prevention of discriminatory effects requires the identification of means to detect these effects, and this can be very difficult when human beings are making the decisions. Interestingly, machine learning algorithms require greater levels of detail and specificity than the ones needed in the human decision-making processes. Thus, regulatory and legal changes may potentially force machine learning algorithms to be transparent and to become effective tools for detecting and preventing discrimination (Kleinberg et al. 2020).

Note that these limitations of AI systems are not disconnected from each other. Recent work has explored the relationship between algorithmic fairness and explainability. For example, Dodge *et al.* (Dodge et al. 2019) studied how unbiased, user-friendly explanations might help humans assess the fairness of a specific machine learning-based decision-making system. The authors find that the type of explanation impacts the users' perception of algorithmic fairness; different types of fairness might require different styles of explanation; and there are individual differences that determine people's reactions to different kinds of explanations. Others have developed visualizations of different definitions of fairness in ranking decisions to support human decision-making (Ahn & Lin 2020). Thus, there is a fertile ground for novel research at the intersection of algorithmic fairness, explainability and accountability.

Requirements for a Human-centric AI

In this section, we provide an overview of current research efforts towards the development of a *Human-centric AI*. These efforts include a fundamental renegotiation of user-centric data ownership and management as well as the development of secure and privacy-preserving machine learning algorithms; the deployment of transparent and accountable algorithms; and the introduction of machine learning fairness principles and methodologies to overcome biases and discriminatory effects. In our view, humans should be placed at the center of the discussion as humans are ultimately both the actors and the subjects of the decisions made via algorithmic means. If we are able to ensure that these requirements are met, we should be able to realize the positive potential of AI-driven decision-making while minimizing the risks and possible negative unintended consequences on individuals and on the society as a whole.

Privacy-preserving AI algorithms and data cooperatives

A big question for policy-makers and researchers is the following: *how do we unlock the value of human behavioral data while preserving the fundamental right to privacy?* To address this issue, the computer science and AI communities have over the years developed several approaches ranging from *data obfuscation* (i.e. the process of hiding personally identifiable information and other sensitive data using modified content) (Bakken et al. 2004), *data anonymization* (i.e. the process of removing personally identifiable information and other sensitive data from datasets) (Cormode & Srivastava 2009), *adversarial training* (i.e. a technique adopted in computer vision and machine learning communities to obfuscate features so that an attacker cannot reconstruct the original image or to infer sensitive information from those features) (Feutry et al. 2018, Kim et al. 2019, Li et al. 2020), and the generation of synthetic datasets (Machanavajjhala et al. 2008) to methods for quantifying privacy guarantees, such as *differential privacy* (Dwork 2008, Dwork & Roth 2014, Kearns & Roth 2020), or *privacy-preserving machine learning* (PPML) approaches (Chaudhuri & Monteleoni 2008). PPML is inspired by research efforts in cryptography and it has the goal of protecting the privacy of the input data and/or of the models used in the learning task. Examples of PPML approaches are (i) *federated learning* (Kairouz et al. 2019, Yang et al. 2019) and (ii) *encrypted computation* (Dowlin et al. 2016).

More in detail, *differential privacy* (Dwork 2008, Dwork & Roth 2014, Kearns & Roth 2020) is a methodology that provides a formal quantification of privacy guarantees with respect to an aggregate metric on a dataset due to a privacy protection mechanism. Examples of privacy protection mechanisms that *differential privacy* can be applied to include adding noise, providing a coarser histogram, or learning with adversarial examples. The value of *differential privacy* is that given a particular dataset and privacy mechanism it can quantify the probability of a privacy leak with guarantees. Furthermore, *differential privacy* guarantees that the distribution of aggregate metric values (e.g. database values, model predictions), such as mean, variance, prediction probability distribution, etc., are indistinguishable (to within some bound) between the original dataset and a dataset where any training datapoint is omitted (Dwork 2008, Dwork & Roth 2014, Kearns & Roth 2020).

Federated learning is a machine learning approach where different entities or organizations collaboratively train a model, while at the same time they keep the training data decentralized in local nodes (Kairouz et al. 2019, Yang et al. 2019). Hence, the raw data samples of each entity are stored locally and never exchanged, and only parameters of the learning algorithm are exchanged in order to generate a global model (Kairouz et al. 2019, Yang et al. 2019). It is worth noting that *federated learning* (Kairouz et al. 2019, Yang et al. 2019) does not provide a full guarantee of the privacy of sensitive data (e.g. personal data) as some characteristics of the raw data could be memorized during the training of the algorithm and thus extracted. For this reason, *differential privacy* can complement *federated learning* by providing guarantees of keeping private the contribution of single organizations/nodes in the federated setting (Brundage et al. 2020, Dubey & Pentland 2020).

Finally, *encrypted computation* (Dowlin et al. 2016) aims at protecting the learning model itself by allowing to train and evaluate on encrypted data. Thus, the entity/organization training the model is not be able to see and/or leak the data in its non-encrypted form. Examples of methods for *encrypted computation* are (i) *homomorphic encryption* (Dowlin et al. 2016), (ii) *functional encryption* (Dowlin et al. 2016), (iii) *secure multi-party computation* (Dowlin et al. 2016), and (iv) *influence matching* (Pan et al. 2012).

This is an active and growing area with several open-source frameworks available to perform privacy-preserving machine learning, such as PySyft (https://github.com/OpenMined/PySyft), Tensor Flow Federated (https://www.tensorflow.org/federated), FATE (https://fate.fedai.org/overview/), PaddleFL (https://paddlefl.readthedocs.io/en/latest), Sherpa.AI (https://developers.sherpa.ai/privacy-technology/), and Tensor Flow Privacy (https://github.com/tensorflow/privacy).

Additionally, new user-centric models and technologies for personal data management have been proposed, in order to empower individuals with more control of their own data's life-cycle (Pentland 2012, de Montjoye et al. 2014, Staiano et al. 2014). Along this line, Hardjono and Pentland (Hardjono & Pentland 2019) have recently introduced the notion of a *data cooperative* that refers to the voluntary collaborative sharing by individuals of their personal data for the benefit of their community. The authors underline several key aspects of a *data cooperative*. First of all, a data cooperative member has legal ownership of her/his data: this data can be collected into her/his Personal Data Store (PDS) (de Montjoye et al. 2014), and s/he can add and remove data from the PDS as well as suspend access to the data repository. Members have the option to maintain their single or multiple Personal Data Stores at the cooperative or in private data servers. However, if the data store is hosted at the cooperative, then data protection (e.g. data encryption) and curation are performed by the cooperative itself for the benefit of its members. Moreover, the data cooperative has a legal fiduciary obligation to its members (Balkin 2016, Hardjono & Pentland 2019): this means that the cooperative organization is owned and controlled by the members. Finally, the ultimate goal of the data cooperative is to benefit and empower its members (Hardjono & Pentland 2019). As highlighted by Hardjono and Pentland (Hardjono & Pentland 2019), credit and labor unions can provide an inspiration for data cooperatives as collective institutions able to represent the data rights of individuals.

Interestingly, Loi *et al.* (Loi et al. 2020) have recently proposed *personal data platform cooperatives* as means for avoiding asymmetries and inequalities in the data economy and realizing the concept of property-owning democracy, introduced by the political and moral philosopher Rawls (Rawls 1971, 2001). In particular, Loi *et al.* (Loi et al. 2020) argue that a society characterized by multiple *personal data platform cooperatives* is more likely to realize the Rawls' principle of *fair Equality of Opportunity* (Rawls 1971, 2001), where individuals have equal access to the resources –data in this case– needed to develop their talents.

Algorithmic transparency and accountability

The traditional strategy for ensuring soundness of a decision-making process is *auditing*, and this approach may easily be applied to machine learning decisions. This strategy deals with the decision process as a black-box where only inputs and outputs are visible (Sandvig et al. 2014, Guidotti et al. 2018). However, while this approach can demonstrate the fairness or accuracy of the decisions, it has limitations for understanding the reasons for particular decisions (Datta et al. 2015, Guidotti et al. 2018).

As a consequence, *explanations* are increasingly advocated in the research community (Doshi-Velez & Kim 2017, Adadi & Berrada 2018, Guidotti et al. 2018, Lipton 2018, Wang et al. 2019, Miller 2019, Barocas et al. 2020) as a way to help people understand AI-driven decision making processes (Lipton 2018, Selbst & Barocas 2018, Wachter et al. 2018) and identify when they should object to the decisions made by the algorithms (Wachter et al. 2018, Lipton 2018, Selbst & Barocas 2018). As argued by Adadi *et al.* (Adadi & Berrada 2018), the variety of explainability methods, proposed over years, can be classified according to three criteria: (i) the complexity of providing an explanation (i.e. more complex is a machine learning model more difficult it is to explain), (ii) the type of explanation (i.e. *global* vs *local explanations*), and (iii) the dependency from the adopted machine learning model (i.e. *model-specific* vs *model-agnostic explanations*).

Regarding the complexity-related methods, the most simple and straightforward approach is the design and implementation of machine learning algorithms that are intrisically easy to interpret and explain. Several works have proposed this explainability strategy (Caruana et al. 2017, Letham et al. 2015, Ustun & Rudin 2015). However, a problem with the adoption of this strategy is the tradeoff between explainability and accuracy. Indeed, more simple and interpretable models tend to be also less accurate (Sarkar et al. 2016). To avoid this potential tradeoff, several works have proposed to build complex and highly accurate black-box models and then use a different set of techniques to provide the required explanations without knowing the inner functioning of the original machine learning model. In this way, this approach offers a *post-hoc explanation*, e.g. using examples, visualizations or natural language descriptions (Mikolov et al. 2013, Mahendran & Vedaldi 2015, Krening et al. 2016, Lipton 2018). As an alternative, some works have proposed *intrinsic methods* that modify the structure of a complex black-box model (e.g. a deep neural network) to improve its interpretability (Dong et al. 2017, Louizos et al. 2017).

As previously said, some research efforts have attempted to provide an explanation of the *global behavior* of a machine learning model (i.e. *global explanations*) (Lakkaraju et al. 2016, Adadi & Berrada 2018, Lipton 2018, Brundage et al. 2020), while others have focused on a *specific prediction* of the model given an input (i.e. *local explanations*) (Baehrens et al. 2010, Zeiler & Fergus 2014, Zhou et al. 2016, Fong & Vedaldi 2017, Wei Koh & Liang 2017, Adadi & Berrada 2018, Yeh et al. 2018, Fong et al. 2019, Brundage et al. 2020, Guidotti 2021). Notable examples of building explanations about the global behavior of a machine learning model are (i) the characterization of the role played by the internal components of the model (e.g. visualization of the features) (Bau et al. 2017, Ulyanov et al. 2018, Brundage et al. 2020), and (ii) the approximation of a complex model by means of a simpler one (e.g. a decision tree) (Zhang et al. 2019, Brundage et al. 2020). However, it is worth noticing that *global explanations* are hard to obtain, in particular for machine learning models characterized by a large number of parameters (Adadi & Berrada 2018). Instead, notable examples of building explanations for a specific decision or a single prediction include (i) identifying which training examples (Lakkaraju et al. 2016, Wei Koh & Liang 2017, Yeh et al. 2018) or (ii) which parts of the training data (Dabkowski & Gal 2017, Fong & Vedaldi 2017, Fong et al. 2019) are responsible for the model's prediction. A recent promising line of work is trying to combine the benefits of *global* and *local explanations* (Linsley et al. 2018, Molnar 2019, Pedreschi et al. 2019).

Furthermore, a third way to characterize techniques for explaining machine learning models is whether they are *model-agnostic explanations*, thus applicable to any type of machine learning model, or *model-specific explanations*, thus applicable only to a single class of machine learning algorithms (Adadi & Berrada 2018). As highlighted by Adadi *et al.* (Adadi & Berrada 2018), *intrinsic methods* provide by definition *model-specific explanations*. However, this approach limits the choice of models, often at the expenses of more predictive and accurate ones (Adadi & Berrada 2018). For this reason, there has been a recent growth of *model-agnostic approaches*, which separate prediction and explanation. These *model-agnostic methods* fall into four techniques: (i) *visualizations*, (ii) *influence methods*, (iii) *example-based explanations*, and (iv) *knowledge extraction* (Adadi & Berrada 2018).

The idea behind visualization techniques is to visualize, expecially in deep neural networks, the representations of the learning model. Popular examples of visualization techniques are (i) *surrogate models* (i.e. interpretable models like a decision tree which are trained on the predictions of the black-box model to make easier its interpretation) (Ribeiro et al. 2016, Bastani et al. 2017), (ii) *partial dependance plots* (i.e. graphical representations visualizing the partial average relationships between input variables and predictions) (Chipman et al. 2010), and (iii) *individual conditional expectations* (i.e. plots revealing the individual relationships between input variables and predictions by disaggregating the output of the partial dependance plots) (Casalicchio et al. 2018).

Influence methods, instead, estimate the relevance of an input variable (i.e. feature) by modifying the input data or the internal components of the model, and then recording how the change affects the performance of the machine learning model (Adadi & Berrada 2018). Looking at the state-of-the-art literature, we may find three different approaches to estimate the importance of an input variable: (i) *sensitivity analysis* (i.e. this method evaluates wheter the performance of the model remains stable when input data are perturbed) (Cortez & Embrechts 2013), (ii) *feature importance* (i.e. this approach quantifies the contribution of a given input variable to the model's predictions by computing the increase of the prediction after permuting the input variable) (Casalicchio et al. 2018), and (iii) *layer-wise relevance propagation algorithm* (i.e. this method decomposes the output of a deep neural network into the relevance scores of the input and at the same time keeps the total amount of relevance constant across the layers) (Bach et al. 2015).

Example-based explanations select specific instances of the dataset under investigation to explain the behavior of a machine learning model. Two promising approaches are (i) *counterfactual explanations* (i.e. these explanations are generated by analyzing how minimal changes in the features would impact and modify the output of the learning model) (Wachter et al. 2018, Dhurandhar et al. 2018, Karimi et al. 2020), and (ii) *prototypes* and *criticisms* (i.e. *prototypes* are representative instances from the dataset, while *criticisms* are instances not well represented by those prototypes) (Kim et al. 2014, 2016).

Finally, some techniques aim at extracting, in a understandable form, knowledge from a machine learning model (in particular, from deep neural networks). Examples of these techniques are (i) *rule extraction* (i.e. this approach provides a symbolic description of the knowledge learned by an highly complex model) (Hailesilassie 2016), and (ii) *model distillation* (i.e. distillation consists in a model compression to transfer information from an highly complex model, called "teacher", to a simpler one, called "student") (Hinton et al. 2015, Furlanello et al. 2018, Xu et al. 2018).

Obviously, a relevant challenge about *transparency* and *accountability* is the difficulty in producing explanations that are *human-understandable* (Guidotti et al. 2018). This implies the communication of complex computational processes to humans, and thus it requires a multidisciplinary research effort mixing methodologies and technologies from human-computer interaction and machine learning communities with models on human explanation processes developed in cognitive and social sciences. For example, the AI scholar Tim Miller (Miller 2019) has extensively analysed the research conducted on human explanation processes in cognitive science (Lombrozo 2006), cognitive and social psychology (Hilton 1990) and philosophy (Lewis 1974), and has highlighted four major findings to take into account in order to build explainable AI methods that can be understable and useful for humans. First of all, explanations are *contrastive* (Lipton 1990, Miller 2019); this means that people do not ask why a given event happened, but rather why this event happened instead of an alternative one. Then, explanations are *selective* and thus they focus only on one or few possible causes and not on all the possible ones (Hilton et al. 2010, Miller 2019). Explanations constitutes a *social conversation* for transfering knowledge (Hilton 1990, Walton 2004), and thus the AI-driven explainer should be able to leverage the mental model of the human explainee during the explanation process (Miller 2019). Finally, the reference to statistical associations in human explanations is less effective than referring to causes.

Adopting a similar multidisciplinary approach and drawing insights from philosophy, cognitive psychology and decision science (Lipton 1990, Hoffman & Klein 2017, Miller 2019), Wang *et al.* (Wang et al. 2019) have recently proposed a conceptual framework that connects explainable AI techniques with core concepts of the human decision-making processes. First of all, the authors have identified why individuals look for explanations (i.e. to focus on a small set of causes, to generalize observations in a model able to predict future events, etc.) and how they should reason. Then, Wang *et al.* (Wang et al. 2019) analyzed several explainable AI techniques and how they have been developed to support specific reasoning methods. For example, visualization techniques, such as saliency heatmaps (Ribeiro et al. 2016, Kim et al. 2018), support contrastive and counter-factual explanations (Miller 2019). As a third part of their conceptual framework, the authors have highlighted and discussed how fast reasoning and cognitive biases may negatively impact human decision-making processes, thus inducing errors (Croskerry 2009, Kahneman & Egan 2011). Finally, Wang *et al.* (Wang et al. 2019) described how explainable AI methods can be adopted as strategies to mitigate some decision biases such as the anchoring bias (i.e. it occurs when the decision-maker is not open to explore alternative hypotheses), the confirmation bias (i.e. the tendency of the decision-maker to interpret information in a way that confirms her/his previous beliefs), the availability bias (it occurs when the decision-maker is unfamiliar with the frequency of a specific outcome), etc.

Another relevant aspect for algorithmic *accountability* and *transparency* is how and from where input data are collected. As recently discussed by Hohman *et al.* (Hohman et al. 2020), machine learning applications require an iterative process to create successful models (Amershi et al. 2014). In particular, Hohman *et al.* (Hohman et al. 2020) have shown that *data iteration* (e.g. collecting novel training data to improve model's performance) is equally important as *model iteration* (e.g. searching for hyperparameters and architectures).

Finally, *transparency* is generally thought as a key enabler of *accountability*. However, transparency is not always needed for accountability. For instance, Kroll *et al.* (Kroll et al. 2017) introduced computational methods that are able to provide accountability even when some fairness-sensitive information is kept hidden, and our earlier discussion about privacy-preserving learning, federated learning, and learning on encrypted data suggests additional paths to accountability without disclosing sensitive data or algorithms.

Algorithmic fairness

A simple way to try to avoid *discrimination* and to maximize *fairness* is the *blindness approach*, namely precluding the use of sensitive attributes (e.g. gender, race, age, income level) in the learning task (Calders & Verwer 2010, Kamiran et al. 2010, Schermer 2011, Barocas & Selbst 2016, Kearns & Roth 2020). For example, in order to build a race-blind AI-driven decision-making process we could avoid to use the "race" attribute. However, this approach has several technical limitations: first of all, the excluded attribute might be implicit in the non-excluded ones (Romei & Ruggieri 2014, Zarsky 2016, Kearns & Roth 2020). For example, the "race" attibute might not be taken directly into account as a criterion for granting or not a loan. However, it might implicitly be present via e.g. the applicant's zip code, given that zip code may be a good proxy for race in a segregated urban environment (Schermer 2011, Macnish 2012).

As a consequence, several researchers have proposed alternative approaches of machine learning *fairness* that formalize the notion of *group fairness* (Calders & Verwer 2010, Kamishima et al. 2011, Zemel et al. 2012, Feldman et al. 2015, Kearns & Roth 2020). One of the most used methods is *statistical parity*, which requires that an equal fraction of each group according to a protected attribute (i.e. black vs white applicants) receives each possible outcome (i.e. loan vs no loan) (Calders & Verwer 2010, Kamishima et al. 2011, Zemel et al. 2012, Feldman et al. 2015, Kearns & Roth 2020). However, the *group fairness* approach often fails at obtaining a good accuracy, as illustrated by the following example in a lending scenario: if two groups (group A and group B) have different proportions of individuals who are able to pay back their loans (e.g. group A has a larger proportion than group B), then the algorithm's accuracy will be compromised if we constrained the algorithm to predict an equal proportion of payback for the two groups. Another issue related to *group fairness* is that a creditworthy individual from group A has no guarantee to have an equal probability of receiving a loan as a similarly creditworthy individual from group B.

A different framework, called *individual fairness*, was introduced by Dwork *et al.* (Dwork et al. 2012). This fairness framework is based on a similarity metric between individuals: any two individuals who are similar should be classified in a similar way (Dwork et al. 2012). This definition resembles partly the interpretation of *Equality of Opportunity* (EoP) proposed by the political scientist Roemer (Roemer 1996, 1998). For Roemer, EoP is achieved when people, irrespective of circumstances beyond their control (e.g. birth circumstances, such as gender, race, familiar socio-economic status, and so forth), have the same ability to achieve desired outcomes through their choices, actions, and efforts (Roemer 1996, 1998). In particular, Roemer claims that if inequalities are caused by birth circumstances, then these are unacceptable and must be compensated by society (Roemer 1996, 1998).

Following Dwork *et al.*'s work (Dwork et al. 2012), Joseph *et al.* (Joseph et al. 2016) proposed an approach to *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" (Rawls 1971). Hence, the formalization of machine learning fairness, proposed by Joseph *et al.* (Joseph et al. 2016), requires that the learning algorithm never favors applicants whose attributes (e.g. income level) are lower than the ones of another applicant. Along this line, Hardt *et al.* (Hardt et al. 2016) have proposed a fairness measure, based again on *Equality of Opportunity*, that tries to overcome the main conceptual shortcomings of *statistical parity* as a fairness notion, and to build classifiers with high accuracy. To this end, they have shown how to optimally adjust any supervised learned predictor to remove discrimination against a specific sensitive attribute (e.g. race, gender, etc.).

Another interesting set of results are the ones obtained by Friedler *et al.* (Friedler et al. 2016), Corbett-Davies *et al.* (Corbett-Davies et al. 2017), and Kleinberg *et al.* (Kleinberg et al. 2017), which highlight that it is not enough to simply achieve *algorithmic fairness*. For example, Friedler *et al.* (Friedler et al. 2016) have proven the impossibility of simultaneously satisfying the mathematical constraints of multiple formalizations of fairness, and thus the impossibility of a single universally accepted definition and metric of *algorithmic fairness*. Indeed, each metric embodies a different criterion of equity. A similar result was discussed by Kleinberg *et al.* (Kleinberg et al. 2017). In their paper, they formalized three fairness conditions, namely *calibration within groups*, *balance for the positive class*, and *balance for the negative class*. Interestingly, they proved that, except in highly constrained special cases, there is no method that is able to satisfy these three conditions at the same time (Kleinberg et al. 2017).

Thus, choosing a particular fairness metric involves implicitly committing to a moral and political philosophy (Heidari et al. 2019, Gummadi & Heidari 2019), the role of social context in the selection process of the fairness metric (Grgic-Hlaca et al. 2018, Madras et al. 2018), and issues of human perception of those metrics (Srivastava et al. 2019). This shifts the question of fairness from a purely technical task to a multi-disciplinary problem. In particular, the problems of defining what equity means as well as what is fair in a given context (Barry 1991) become of paramount relevance. Indeed, what constitutes fairness changes according to different worldviews: for example, the moral and political philosopher Nozick in his book "Anarchy, State, and Utopia" (Nozick 1974) proposed a libertarian alternative view to the Rawlsian notion of EoP. In his view, the elimination of the discriminatory biases, present in society, may create new harms to new groups of people. For this reason, it is urgent to bring together, in joint publications, conferences, projects and institutions, researchers from different fields –including law, moral and political philosophy, and machine learning– to devise, evaluate and validate in the real-world alternative fairness metrics for different tasks.

Finally, as previously noted, recent work has also explored the relationship between fairness and explainability of decision-making algorithms, showing that the type of explanation influences the human's perception of how fair an algorithm is (Dodge et al. 2019).

Conclusion

Our society is experiencing an unprecedented historic moment where the availability of vast amounts of human behavioral data, combined with advances in Artificial Intelligence (and particularly machine learning), is enabling us to tackle complex problems through the use of algorithmic decision-making processes. 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 this opportunity. However, we should focus our attention on existing risks related to the use of algorithmic decision-making processes, including computational violations of privacy, power and information assymetry, lack of transparency and accountability, and discrimination and bias. It is important to note that tackling these limitations would entail multi-disciplinary teams working together with expertise in areas, such as machine learning, human-computer interaction, cognitive sciences, social and cognitive psychology, decision theory, ethics and philosophy, and the law. It will only be via multi-disciplinary approaches, as shown for building human-understandable AI systems and for connecting algorithmic fairness approaches with different moral and political worldviews, that we will be able to effectively address the limitations of today's algorithmic decision-making systems.

We have also underlined three extensive requirements that we consider to be of paramount importance in order to enable an ethical and human-centric use of Artificial Intelligence: (i) privacy-preserving machine learning and user-centric data ownership and management; (ii) algorithmic transparency and accountability; and (iii) algorithmic fairness. If we will honor these requirements, then we would be able to move from the feared tyranny of Artificial Intelligence and of algorithmic mass surveillance (Zuboff 2019) to a *Human-centric AI* model of democratic governance for the people.

Acknowledgements

The authors would like to thank Lorenzo Lucchini and Simone Centellegher for their support in preparing the graphical abstract. The work of Nuria Oliver was partly supported by funding from the Valencian government.

Authors' contributions

All authors contributed equally to the manuscript.

Declaration of Interests

The authors declare that they have no competing interests.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H., Mironov, I., Talwar, K. & Zhang, L. (2016), Deep learning with differential privacy, *in* 'Proceedings of the 2018 ACM Conference on Computer and Communications Security (CCS '16)', pp. 308–318.

Abdul, A., Vermeulen, J., Wang, D., Lim, B. & Kankanhalli, M. (2018), Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda, *in* 'Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems', pp. 1–18.

Adadi, A. & Berrada, M. (2018), 'Peeking inside the black-box: A survey on explainable artificial intelligence (xai)', *IEEE Access* **6**, 52138–52160.

Ahn, Y. & Lin, Y. R. (2020), 'Fairsight: Visual analytics for fairness in decision making', *IEEE Trans Vis Comput Graph* **26**(1), 1086–1095.

Amershi, S., Cakmak, M., Bradley Knox, W. & Kulesza, T. (2014), 'Power to the people: The role of humans in interactive machine learning', *AI Magazine* **35**(4), 105–120.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P., Inkpen, K., Teevan, J., Kikin-Gil, R. & Horvitz, E. (2019), Guidelines for human-ai interaction, *in* 'Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems', pp. 1–13.

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

Bach, S., Binder, A., Montavon, G., Klauschen, F. & Müller, K.-R. (2015), 'On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation', *PloS One* **10**(7), e0130140.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K. & Müller, K.-R. (2010), 'How to explain individual classification decisions', *J. Mach. Learn. Res.* **11**(6), 1803–1831.

Bakken, D., Rarameswaran, R., Blough, D., Franz, A. & Palmer, T. (2004), 'Data obfuscation: Anonymity and desensitization of usable data sets', *IEEE Security & Privacy* **2**(6), 34–41.

Balkin, J. (2016), 'Information fiduciaries and the first amendment', *UC Davis Law Review* **49**(4), 1183–1234.

Barocas, S., Hardt, M. & Narayanan, A. (2018), *Fairness and Machine Learning*, fairmlbook.org.

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

Barocas, S., Selbst, A. & Raghavan, M. (2020), The hidden assumptions behind counterfactual explanations and principal reasons, *in* 'Proceedings of the 2020 International Conference on Fairness, Accountability, and Transparency', pp. 80–89.

Barry, B. (1991), *Theories of Justice*, Berkeley: University of California Press.

Bastani, O., Kim, C. & Bastani, H. (2017), 'Interpreting black box models via model extraction', *arxiv preprint arXiv:1705.08504* .

Bau, D., Zhou, B., Khosla, A., Oliva, A. & Torralba, A. (2017), Network dissection: Quantifying interpretability of deep visual representations, *in* 'Proceedings of the 2017 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2017)', pp. 3319–3327.

Benjamin, R. (2019), 'Assessing risk, automating racism', *Science* **366**(6464), 421–422.

Berk, R., Heidari, H., Jabbari, S., Kearns, M. & Roth, A. (2018), 'Fairness in criminal justice risk assessments: The state of the art', *Sociological Methods & Research* **50**(1), 3–44.

Blumenstock, J., Cadamuro, G. & On, R. (2015), 'Predicting poverty and wealth from mobile phone metadata', *Science* **350**(6264), 1073–1076.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Wei Koh, P., Hooker, S., Leung, J., Trask, T., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., Ryffel, T., Rubinovitz, J., Besiroglu, T., Carugati, F., Clark, J., Eckersley, P., de Haas, S., Johnson, M., Laurie, B., Ingerman, A., Krawczuk, I., Askell, A., Cammarota, R., Lohn, A., Krueger, D., Stix, C., Henderson, P., Graham, L., Prunkl, C., Martin, B., Seger, E., Zilberman, N., hÉigeartaigh, S., Kroeger, F., Sastry, G., Kagan, R., Weller, A., Tse, B., Barnes, E., Dafoe, A., Scharre, P., Herbert-Voss, A., Rasser, M., Sodhani, S., Flynn, C., Krendl Gilbert, T., Dyer, L., Khan, S., Bengio, Y. & Anderljung, M. (2020), 'Toward trustworthy ai development: Mechanisms for supporting verifiable claims', *arxiv preprint arXiv:2004.07213* .

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

Calders, T. & Verwer, S. (2010), 'Three naive bayes approaches for discrimination-free classification', *Data Min. Knowl. Discov.* **21**(2), 277–292.

Calders, T. & Zliobaite, I. (2013), Why unbiased computational processes can lead to discriminative decision procedures, *in* B. Custers, T. Calders, B. Schermer & T. Zarsky, eds, 'Discrimination and Privacy in the Information Society', pp. 43–57.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. & Elhadad, N. (2017), Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, *in* 'Proceedings of the 2017 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2017)', pp. 1721–1730.

Casalicchio, G., Molnar, C. & Bischl, B. (2018), Visualizing the feature importance for black box models, *in* 'Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases', pp. 655–670.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M. & Floridi, L. (2018), 'Artificial intelligence and the 'good society': the us, eu, and uk approach', *Sci. Eng. Ethics* **24**, 505–528.

Chaudhuri, K. & Monteleoni, C. (2008), Privacy-preserving logistic regression, *in* 'Proceedings of the International Conference on Advances in Neural Information Processing Systems (NIPS 2008)'.

Chipman, H., George, E. & McCulloch, R. (2010), 'Bart: Bayesian additive regression trees', *Appl. Statist.* **4**(1), 266–298.

Chittaranjan, G., Blom, J. & Gatica-Perez, D. (2013), 'Mining large-scale smartphone data for personality studies', *Pers Ubiquitous Comput* **17**(3), 433–450.

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), Algorithmic decision making and the cost of fairness, *in* 'Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2017)'.

Cormode, G. & Srivastava, D. (2009), Anonymized data: Generation, models, usage, *in* 'Proceedings of the 2009 ACM SIGMOD International Conference on Management Of Data', pp. 1015–1018.

Cortez, P. & Embrechts, M. (2013), 'Using sensitivity analysis and visualization techniques to open black-box data mining models', *Info. Sci.* **225**, 1–17.

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.

Croskerry, P. (2009), 'Clinical cognition and diagnostic error: Applications of a dual process model of reasoning', *Advances in Health Sciences Education* **14**(1), 27–35.

Dabkowski, P. & Gal, Y. (2017), Real time image saliency for black box classifiers, *in* 'Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)', pp. 6967–6976.

Danziger, S., Levav, J. & Avnaim-Pess, L. (2011), 'Extraneous factors in judicial decisions', *Proc. Natl. Acad. Sci. USA* **108**, 6889–6892.

Datta, A., Tschantz, M. C. & Datta, A. (2015), Automated experiments on ad privacy settings, *in* 'Proceedings on Privacy Enhancing Technologies', pp. 92–112.

de Montjoye, Y.-A., Hidalgo, C., Verleysen, M. & Blondel, V. (2013), 'Unique in the crowd: The privacy bounds of human mobility', *Scientific Reports* **3**.

de Montjoye, Y.-A., Quoidbach, J., Robic, F. & Pentland, A. (2013), Predicting personality using novel mobile phone-based metrics, *in* 'Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction', pp. 48–55.

de Montjoye, Y.-A., Radaelli, L., Singh, V. & Pentland, A. (2015), 'Unique in the shopping mall: On the re-identifiability of credit card metadata', *Science* **347**(6221), 536–539.

de Montjoye, Y.-A., Shmueli, E., Wang, S. & Pentland, A. (2014), 'Openpds: Protecting the privacy of metadata through safeanswers', *PloS One* **9**(7), e98790.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K. & Das, P. (2018), Explanations based on the missing: Towards contrastive explanations with pertinent negatives, *in* 'Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)'.

Diakopoulos, N. (2015), 'Algorithmic accountability: Journalistic investigation of computational power structures', *Digit. Journal.* **3**(3), 398–415.

Dodge, J., Liao, Q., Zhang, Y., Bellamy, R. & Dugan, C. (2019), Explaining models: An empirical study of how explanations impact fairness judgment, *in* 'Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI 2019)', pp. 275–285.

Dong, Y., Su, H., Zhu, J. & Zhang, B. (2017), Improving interpretability of deep neural networks with semantic information, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)', pp. 4306–4314.

Doshi-Velez, F. & Kim, B. (2017), 'Roadmap for a rigorous science of interpretability', *arXiv preprint arXiv:1702.08608* .

Dowlin, N., Gilad-Bachrach, R., Laine, K., Lauter, K., Naehrig, M. & Wernsing, J. (2016), Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy, *in* 'Proceedings of 2016 International Conference on Machine Learning (ICML 2016)', pp. 201–210.

Dubey, A. & Pentland, A. (2020), Private and byzantine-proof federated decision making, *in* 'Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2020)'.

Dwork, C. (2008), Differential privacy: A survey of results, *in* 'Proceedings of the International Conference on Theory and Applications of Models of Computation', pp. 1–19.

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.

Dwork, C. & Roth, A. (2014), 'The algorithmic foundations of differential privacy', *Foundations and Trends in Theoretical Computer Science* **9**(3–4), 211–407.

Eagle, N., Macy, M. & Claxton, R. (2010), 'Network diversity and economic development', *Science* **328**(5981), 1029–1031.

Eubanks, V. (2018), *Automating inequality: How high-tech tools profile, police, and punish the poor*, St. Martin's Press, Inc., USA.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T. & Song, D. (2018), Robust physical-world attacks on deep learning visual classification, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)', pp. 1625–1634.

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 (kDD 2015)', pp. 259–268.

Feutry, C., Piantanida, P., Bengio, Y. & Duhamel, P. (2018), 'Learning anonymized representations with adversarial neural networks', *arXiv preprint arXiv:1802.09386* .

Fiske, S. (1998), Stereotyping, prejudice, and discrimination, *in* D. Gilbert, S. Fiske & G. Lindzey, eds, 'Handbook of Social Psychology', Boston: McGraw-Hill, pp. 357–411.

Fong, R., Patrick, M. & Vedaldi, A. (2019), Understanding deep networks via extremal perturbations and smooth masks, *in* 'Proceedings of the IEEE International Conference on Computer Vision (CVPR 2019)', pp. 2950–2958.

Fong, R. & Vedaldi, A. (2017), Interpretable explanations of black boxes by meaningful perturbation, *in* 'Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV 2017)', pp. 3449–3457.

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

Furlanello, T., Lipton, Z., Tschannen, M., Itti, L. & Anandkumar, A. (2018), Born again neural networks, *in* 'Proceedings of the International Conference on Machine Learning (ICML 2018)', pp. 1602–1611.

Ganju, K., Wang, Q., Yang, W., Gunter, C. & Borisov, N. (2018), Property inference attacks on fully connected neural networks using permutation invariant representations, *in* 'Proceedings of the 2018 ACM Conference on Computer and Communications Security (CCS '18)', pp. 619–633.

Gillespie, T. (2014), The relevance of algorithms, *in* T. Gillespie, P. Boczkowski & K. Foot, eds, 'Media technologies: Essays on communication, materiality, and society', MIT Press, pp. 167–193.

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M. & Brilliant, L. (2009), 'Detecting influenza epidemics using search engine query data', *Nature* **457**(7232), 1012–1014.

Gladstone, J., Matz, S. & Lemaire, A. (2019), 'Can psychological traits be inferred from spending? evidence from transaction data', *Psychol. Sci.* **30**(7), 1087–1096.

Grgic-Hlaca, N., Zafar, M., Gummadi, K. & Weller, A. (2018), Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning, *in* 'Proceedings of the 32nd Conference on Artificial Intelligence (AAAI 2018)', pp. 51–60.

Guidotti, R. (2021), 'Evaluating local explanation methods on ground truth', *Artificial Intelligence* **291**, 103428.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. & Pedreschi, D. (2018), 'A survey of methods for explaining black box models', *ACM Comput. Surv.* **51**(5), 1–42.

Gummadi, K. & Heidari, H. (2019), Economic theories of distributive justice for fair machine learning, *in* 'Companion Proceedings of the 2019 World Wide Web (WWW 2019) Conference', pp. 1301–1302.

Hailesilassie, T. (2016), ‘Rule extraction algorithm for deep neural networks: A review’, *arXiv preprint arXiv:1610.05267* .

Hand, D. (2006), ‘Classifier technology and the illusion of progress’, *Stat. Sci.* **21**, 1–14.

Hardjono, T. & Pentland, A. (2019), ‘Data cooperatives: Towards a foundation for decentralized personal data management’, *arxiv preprint arXiv:1905.08819* .

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 2016)’, pp. 3315–3323.

Hayes, J. & Ohrimenko, O. (2018), Contamination attacks and mitigationin multi-party machine learning, *in* ‘Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)’, pp. 6602–6614.

Heidari, H., Loi, M., Gummadi, K. & Krause, A. (2019), A moral framework for understanding of fair ml through economic models of equality of opportunity, *in* ‘Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency’, pp. 181–190.

Hillebrand, M., I., K., Peleja, F. & Oliver, N. (2020), ‘Mobisenseus: Inferring aggregate objective and subjective well-being from mobile data’, *Proceedings of the European Conference on Artificial Intelligence (ECAI 2020)* pp. 1818–1825.

Hilton, D. (1990), ‘Conversational processes and causal explanation’, *Psychol. Bull.* **107**(1), 65–81.

Hilton, D., McClure, J. & Sutton, R. (2010), ‘Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes?’, *Eur. J. Soc. Psychol.* **40**(3), 383–400.

Hinton, G., Vinyals, O. & Dean, J. (2015), ‘Distilling the knowledge in a neural network’, *arxiv preprint arXiv:1503.02531* .

Hoffman, R. & Klein, G. (2017), ‘Explaining explanation, part 1: theoretical foundations’, *IEEE Intelligent Systems* **3**, 68–73.

Hohman, F., Wongsuphasawat, K., Kery, M. & Patel, K. (2020), Understanding and visualizing data iteration in machine learning, *in* ‘Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems’, pp. 1–13.

Horvitz, E. (1999), Principles of mixed-initiative user interfaces, *in* ‘Proceedings of the 1999 CHI Conference on Human Factors in Computing Systems’, pp. 159–166.

Jannach, D. & Adomavicius, G. (2016), Recommendations with a purpose, *in* 'Proceedings of the 10th ACM Conference on Recommender Systems (RecSys)', pp. 7–10.

Jia, J., Lu, X., Yuan, Y., Xu, G., Jia, J. & Christakis, N. (2020), 'Population flow drives spatio-temporal distribution of covid-19 in china', *Nature* **582**, 389–394.

Jobin, A., Ienca, M. & Vayena, E. (2019), 'The global landscape of ai ethics guidelines', *Nat. Mach. Intell.* **1**, 389–399.

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

Kahneman, D. & Egan, P. (2011), *Thinking, fast and slow (Vol. 1)*, Farrar, Straus and Giroux.

Kairouz, P., McMahan, H., Avent, B., Bellet, A., Bennis, M., Bhagoji, A., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S., Sun, Z., Suresh, A., Tramér, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F., Yu, H. & Zhao, S. (2019), 'Toward trustworthy ai development: Mechanisms for supporting verifiable claims', *arxiv preprint arXiv:1912.04977* .

Kamiran, F., Calders, T. & Pechenizkiy, M. (2010), Discrimination aware decision tree learning, *in* 'Proceedings of 2010 IEEE International Conference on Data Mining (ICDM 2010)', pp. 869–874.

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 2011), Part II', pp. 35–50.

Karimi, A.-H., Barthe, G., Balle, B. & Valera, I. (2020), Model-agnostic counterfactual explanations for consequential decisions, *in* 'Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)', pp. 895–905.

Kearns, M. & Roth, A. (2020), *The Ethical Algorithm*, Oxford University Press.

Kesarwani, M., Mukhoty, B., Arya, V. & Mehta, S. (2018), Model extraction warning in mlaas paradigm, *in* 'Proceedings of the 34th Annual Computer Security Applications Conference (AC-SAC 2018)', pp. 371–380.

Kim, B., Khanna, R. & Koyejo, O. (2016), Examples are not enough, learn to criticize! criticism for interpretability, *in* ‘Proceedings of the 29th Conference on Neural Information Processing Systems (NIPS 2016)’, pp. 2280—-2288.

Kim, B., Rudin, C. & Shah, J. (2014), The bayesian case model: A generative approach for case-based reasoning and prototype classification, *in* ‘Proceedings of the 27th Conference on Neural Information Processing Systems (NIPS 2014)’, pp. 1952–1960.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J. & Viegas, F. (2018), Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), *in* ‘Proceedings of the International Conference on Machine Learning (ICML 2018)’, pp. 2673–2682.

Kim, T., Kang, D., Pulli, K. & Choi, J. (2019), ‘Training with the invisibles: Obfuscating images to share safely for learning visual recognition models’, *arXiv preprint arXiv:1901.00098* .

Kleinberg, J., Ludwig, J., Mullainathan, S. & Sunstein, C. (2020), ‘Algorithms as discrimination detectors’, *Proc. Natl. Acad. Sci. USA* **117**(48), 30096–30100.

Kleinberg, J., Mullainathan, S. & Raghavan, M. (2017), Inherent trade-offs in the fair determination of risk scores, *in* ‘Proceedings of Innovations in Theoretical Computer Science Conference’, pp. 1–23.

Kosinski, M., Stillwell, D. & Graepel, T. (2013), ‘Private traits and attributes are predictable from digital records of human behavior’, *Proc. Natl. Acad. Sci. USA* **110**(15), 5802–5805.

Krause, J., Perer, A. & Ng, K. (2016), Interacting with predictions: Visual inspection of black-box machine learning models, *in* ‘Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems’, pp. 5686–5697.

Krening, S., Harrison, B., Feigh, K., Isbell, C., Riedl, M. & Thomaz, A. (2016), ‘Learning from explanations using sentiment and advice in rl’, *IEEE Trans. Cogn. Develop. Syst.* **9**(1), 44–55.

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**.

Lai, S., Ruktanonchai, N., Zhou, L., Prosper, O., Luo, W., Floyd, J., Wesolowski, A., Santillana, M., Zhang, C., Du, X., Yu, H. & Tatem, A. (2020), ‘Effect of non-pharmaceutical interventions to contain covid-19 in china’, *Nature* **585**, 410–413.

Lakkaraju, H., Bach, S. & Leskovec, J. (2016), Interpretable decision sets: A joint framework for description and prediction, *in* ‘Proceedings of the 2016 ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2016)’, pp. 1675–1684.

Lee, M. (2018), 'Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management', *Big Data & Society* **5**(1), 2053951718756684.

Lee, M. & Baykal, S. (2017), Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division, *in* 'Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW 2017)', pp. 1035–1048.

Lee, M., Kusbit, D., Metsky, E. & Dabbish, L. (2015), Working with machines: The impact of algorithmic and data-driven management on human workers, *in* 'Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems', pp. 1603–1612.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A. & Vinck, P. (2018), 'Fair, transparent, and accountable algorithmic decision-making processes', *Philosophy & Technology* **31**, 611–627.

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, *in* T. Cerquitelli, D. Quercia & F. Pasquale, eds, 'Transparent Data Mining for Big and Small Data. Studies in Big Data', Vol. 32, Springer, Cham.

Letham, B., Rudin, C., McCormick, T. & Madigan, D. (2015), 'Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model', *Ann. Appl. Statist.* **9**(3), 1350–1371.

Letouzé, E. & Pentland, A. (2018), 'Towards a human artificial intelligence for human development', *ITU Journal: ICT Discoveries* **1**(2).

Lewis, D. (1974), 'Causation', *J. Philos.* **70**(17), 556–567.

Li, A., Guo, J., Yang, H. & Chen, Y. (2020), Deepobfuscator: Adversarial training framework for privacy-preserving image classification, *in* 'Proceedings of the European Conference on Computer Vision (ECCV 2020)'.

Lim, B., Dey, A. & Avrahami, D. (2009), Why and why not explanations improve the intelligibility of context-aware intelligent systems, *in* 'Proceedings of the 2009 CHI Conference on Human Factors in Computing Systems', pp. 2119–2128.

Linsley, D., Scheibler, D., Eberhardt, S. & Serre, T. (2018), 'Global-and-local attention networks for visual recognition', *arxiv preprint arxiv:1805.08819* .

Lipton, P. (1990), 'Contrastive explanation', *Royal Institute of Philosophy Supplements* **27**, 247–266.

Lipton, Z. (2018), ‘The mythos of model interpretability’, *Commun. ACM* **61**(10), 36–43.

Loi, M., Dehaye, P.-O. & Hafen, E. (2020), ‘Towards rawlsian ’property-owning democracy’ through personal data platform cooperatives’, *Crit Rev Int Soc Political Philos* pp. 1–19.

Lombrozo, T. (2006), ‘The structure and function of explanations’, *Trends in Cognitive Sciences* **10**(10), 464–470.

Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R. & Welling, M. (2017), Causal effect inference with deep latent-variable models, *in* ‘Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2017)’, pp. 6446–6456.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J. & Vilhuber, L. (2008), Privacy: Theory meets practice on the map, *in* ‘Proceedings of the IEEE 24th International Conference on Data Engineering’, pp. 277–286.

Macnish, K. (2012), ‘Unblinking eyes: The ethics of automating surveillance’, *Ethics Inf Technol* **14**(2), 151–167.

Madras, D., Pitassi, T. & Zemel, R. (2018), Predict responsibly: Improving fairness and accuracy by 940 learning to defer, *in* ‘Proceedings of the 2018 International Conference on Advances in Neural Information Processing Systems’, pp. 6147–6157.

Mahendran, A. & Vedaldi, A. (2015), Understanding deep image representations by inverting them, *in* ‘Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)’, pp. 5188–5196.

Marcinkowski, F., Kieslich, K., Starke, C. & Lunich, M. (2020), Implications of ai (un-) fairness in higher education admissions: the effects of perceived ai (un-) fairness on exit, voice and organizational reputation, *in* ‘Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency’, pp. 122–130.

Matz, S., Kosinski, M., Nave, G. & Stillwell, D. (2017), ‘Psychological targeting as an effective approach to digital mass persuasion’, *Proc. Natl. Acad. Sci. USA* **114**, 12714–12719.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. (2013), Distributed representations of words and phrases and their compositionality, *in* ‘Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS 2013)’, pp. 3111–3119.

Miller, T. (2019), ‘Explanation in artificial intelligence: Insights from the social sciences’, *Artificial Intelligence* **267**, 1–38.

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. (2016), 'The ethics of algorithms: Mapping the debate', *Big Data & Society* **3**(2), 2053951716679679.

Molnar, C. (2019), *Interpretable machine learning. A guide for making black box models explainable*.

Nasr, M., Shokri, R. & Houmansadr, A. (2019), Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning, *in* 'Proceedings of IEEE Symposium on Security and Privacy (S&P 2019)', pp. 739–753.

Noble, S. U. (2018), *Algorithms of oppression: How search engines reinforce racism*, NYU Press.

Norman, D. (1994), 'How might people interact with agents', *Commun. ACM* **37**(7), 68–71.

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

Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. (2019), 'Dissecting racial bias in an algorithm used to manage the health of populations', *Science* **366**(6464), 447–453.

Ofli, F., Meier, P., Imran, M., Castillo, C., Tuia, D., Rey, N., Briant, J., Millet, P., Reinhard, F., Parkan, M. & Joost, S. (2016), 'Combining human computing and machine learning to make sense of big (aerial) data for disaster response', *Big Data* **4**, 47–59.

Ohm, P. (2010), 'Broken promises of privacy: Responding to the surprising failure of anonymization', *UCLA Law Review* **57**, 1701–1777.

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

Oyebode, O. & Orji, R. (2020), 'A hybrid recommender system for product sales in a banking environment', *J. Bank. Finance* **4**, 15–25.

Pan, W., Cebrian, M., Kim, T., Fowler, J. & Pentland, A. (2012), 'Modeling dynamical influence in human interaction', *IEEE Signal Process. Mag.* **29**(2), 77–86.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Berkay Celik, Z. & Swami, A. (2016), The limitations of deep learning in adversarial settings, *in* 'Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P 2016)', pp. 372–387.

Parise, S., Kiesler, S., Sproull, L. & Waters, K. (1999), 'Cooperating with life-like interface agents', *Comput. Hum. Behav.* **15**(2), 123–142.

Park, S., Matic, A., Garg, K. & Oliver, N. (2018), 'When simpler data does not imply less information: a study of user profiling scenarios with constrained view of mobile http (s) traffic', *ACM Trans. Web* **12**(2), 1–23.

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

Pastor-Escuredo, D., Torres Fernandez, Y., Bauer, J., Wadhwa, A., Castro-Correa, C., Romanoff, L., Lee, J., Rutherford, A., Frias-Martinez, V., Oliver, N., E., F.-M. & Luengo-Oroz, M. (2014), Flooding through the lens of mobile phone activity, *in* 'IEEE Global Humanitarian Technology Conference (GHTC 2014)', pp. 279–286.

Patel, K., Drucker, S., Fogarty, J., Kapoor, A. & Tan, D. (2011), Using multiple models to understand data, *in* 'Proceedings of the 2011 International Joint Conference on Artificial Intelligence (IJCAI 2011)', pp. 1723–1728.

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S. & Turrini, F. (2019), Meaningful explanations of black box ai decision systems, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2019)', Vol. 33, pp. 9780–9784.

Pentland, A. (2012), 'Society's nervous system: Building effective government, energy, and public health systems', *IEEE Computer* **45**(1), 31–38.

Quercia, D., Kosinski, M., Stillwell, D. & Crowcroft, J. (2011), Our twitter profiles, our selves: Predicting personality with twitter, *in* 'Proceedings of the 2011 IEEE Third International Conference on Social Computing (SocialCom 2011)', pp. 180–185.

Raghavan, M., Barocas, S., Kleinberg, J. & Levy, K. (2020), Mitigating bias in algorithmic hiring: Evaluating claims and practices, *in* 'Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency', pp. 469–481.

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

Rawls, J. (2001), *Justice as fairness: A restatement*, Harvard University Press.

Reichman, N., Teitler, J., Garfinkel, I. & McLanahan, S. (2001), 'Fragile families: Sample and design', *Child. Youth Serv. Rev.* **23**, 303–326.

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 (KDD 2016)', pp. 1135–1144.

Roemer, J. (1996), *Theories of Distributive Justice*, Harvard University Press.

Roemer, J. (1998), *Equality of Opportunity*, Harvard University Press.

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

Rudin, C. (2019), ‘Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead’, *Nat. Mach. Intell.* **1**, 206–215.

Salganik, M., Lundberg, I., Kindel, A., Ahearn, C., Al-Ghoneim, K., Almaatouq, A., Altschul, D., Brand, J., Carnegie, N., Compton, R., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., Morgan, A., Pentland, A., Polimis, K., Raes, L., Rigobon, D., Roberts, C., Stanescu, D., Suhara, Y., Usmani, A., Wang, E., Adem, M., Alhajri, A., AlShebli, B., Amin, R., Amos, R., Argyle, L., Baer-Bositis, L., Buchi, M., Chung, B.-R., Eggert, W., Faletto, G., Fan, Z., Freese, J., Gadgil, T., Gagné, J., Gao, Y., Halpern-Manners, A., Hashim, S., Hausen, S., He, G., Higuera, K., Hogan, B., Horwitz, I., Hummel, L., Jain, N., Jin, K., Jurgens, D., Kaminksi, P., Karapetyan, A., Kim, E., Leizman, B., Liu, N., Möser, M., Mack, A., Mahajan, M., Mandell, N., Marahrens, H., Mercado-Garcia, D., Mocz, V., Mueller-Gastell, K., Musse, A., Niu, Q., Nowak, W., Omidvar, H., Or, A., Ouyang, K., Pinto, K., Porter, E., K.E., P., Qian, C., Rauf, T., Sarsyan, A., Schaffner, T., Schnabel, L., Schonfeld, B., Sender, B., Tang, J., Tsurkov, E., van Loon, A., Varol, O., Wang, X., Wang, Z., Wang, J., Wang, F., Weissman, S., Whitaker, K., Wolters, M., Lee Woon, W., Wu, J., Wu, C., Yang, K., Yin, J., Zhao, B., Zhu, C., Brooks-Gunn, J., Engelhardt, B., Hardt, M., Knox, D., Levy, K., Narayanan, A., Stewart, B., Watts, D. & McLanahan, S. (2020), ‘Measuring the predictability of life outcomes with a scientific mass collaboration’, *Proc. Natl. Acad. Sci. USA* **117**(15), 8398–8403.

Samuelson, W. & Zeckhauser, R. (1988), ‘Status quo bias in decision making’, *J Risk Uncertain* **1**(1), 7–59.

Sandvig, C., Hamilton, K., Karahalios, K. & Langbort, C. (2014), Auditing algorithms: Research methods for detecting discrimination on internet platforms, *in* ‘Proceedings of Data and Discrimination: Converting Critical Concerns into Productive Inquiry, a preconference at the 64th Annual Meeting of the International Communication Association’.

Sarkar, S., Weyde, T., Garcez, A., Slabaugh, G., Dragicevic, S. & Percy, C. (2016), Accuracy and interpretability trade-offs in machine learning applied to safer gambling, *in* ‘Proceedings of CoCo@NIPS’.

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

Schwartz, H., Eichstaedt, J., Kern, M., Dziurzynski, L., Ramones, S., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. & Ungar, L. (2013), ‘Personality, gender, and age in the language of social media: The open-vocabulary approach’, *PloS One* **8**(9), e73791.

Segalin, C., Celli, F., Polonio, L., Kosinski, M., Stillwell, D., Sebe, N., Cristani, M. & Lepri, B. (2017), What your facebook profile picture reveals about your personality, *in* ‘Proceedings of the 25th ACM international conference on Multimedia (ACM MM 2017)’, pp. 460–468.

Selbst, A. & Barocas, S. (2018), ‘The intuitive appeal of explainable machines’, *Fordham Law Review* **87**(1085).

Sheridan, T. & Parasuraman, R. (2005), ‘Human-automation interaction’, *Reviews of human factors and ergonomics* **1**(1), 89–129.

Shneiderman, B. (2016), ‘Opinion: The dangers of faulty, biased, or malicious algorithms requires inde-pendent oversigh’, *Proc. Natl. Acad. Sci. USA* **113**(48), 13538–13540.

Shokri, R. & Shmatikov, V. (2015), Privacy-preserving deep learning, *in* ‘Proceedings of the 2015 ACM Conference on Computer and Communications Security (CCS ’15)’, pp. 1310–1321.

Shokri, R., Stronati, M., Song, C. & Shmatikov, V. (2017), Membership inference attacks against machine learning models, *in* ‘Proceedings of IEEE Symposium on Security and Privacy (S&P 2017)’, pp. 3–18.

Siting, Z., Wenxing, H., Ning, Z. & Fan, Y. (2012), Job recommender systems: A survey, *in* ‘Proceedings of International Conference on Computer Science Education (ICCSE)’, pp. 920–924.

Song, C., Ristenpart, T. & Shmatikov, V. (2017), Machine learning models that remember too much, *in* ‘Proceedings of the 2017 ACM Conference on Computer and Communications Security (CCS ’17)’, pp. 587–601.

Song, L., Shokri, R. & Mittal, P. (2019), Privacy risks of securing machine learning models against adversarial examples, *in* ‘Proceedings of the 2019 ACMSIGSAC Conference on Computer and Communications Security (CCS ’19)’, pp. 241–257.

Soto, V., Frias-Martinez, V., Virseda, J. & Frias-Martinez, E. (2011), Prediction of socioeconomic levels using cell phone records, *in* ‘Proceedings of the International Conference on User Modeling, Adaptation, and Personalization (UMAP 2011)’, pp. 377–388.

Srivastava, M., Heidari, H. & Krause, A. (2019), Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning, *in* ‘Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining’.

Stachl, C., Au, Q., Schoedel, R., Gosling, S., Harari, G., Buschek, D., Völkel, S., Schuwerk, T., Oldemeier, M., Ullmann, T., Hussmann, H., Bischl, B. & Bühner, M. (2020), 'Predicting personality from patterns of behavior collected with smartphones', *Proc. Natl. Acad. Sci. USA* **17**(30), 17680–17687.

Staiano, J., Lepri, B., Aharony, N., Pianesi, F., Sebe, N. & Pentland, A. (2012), Friends don't lie: inferring personality traits from social network structure, *in* 'Proceedings of the 2012 ACM Conference on Ubiquitous Computing', pp. 321–330.

Staiano, J., Oliver, N., Lepri, B., de Oliveira, R., Caraviello, M. & Sebe, N. (2014), Money walks: a human-centric study on the economics of personal mobile data, *in* 'Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing', pp. 583–594.

Steele, J., Sundsoy, P., Pezzulo, C., Alegana, V., Bird, T., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y.-A., Iqbal, A., Hadiuzzaman, K., X., L., E., W., Tatem, A. & Bengtsson, L. (2017), 'Mapping poverty using mobile phone and satellite data', *J. R. Soc. Interface.* **14**, 20160690.

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

Tramér, F., Zhang, F., Juels, A., Reiter, M. & Ristenpart, T. (2016), Stealing machine learning models via prediction apis, *in* 'Proceedings of the USENIX Security Symposium', pp. 601–618.

Tufekci, Z. (2015), 'Algorithmic harms beyond facebook and google: Emergent challenges of computational agency', *Colorado Technology Law Journal* **13**, 203–218.

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

Ulyanov, D., Vedaldi, A. & Lempitsky, V. (2018), Deep image prior, *in* 'Proceedings of the 2018 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2018)', pp. 9446–9454.

Ustun, B. & Rudin, C. (2015), 'Supersparse linear integer models for optimized medical scoring systems', *Mach. Learn.* **102**(3), 349–391.

Veale, M. & Binns, R. (2017), 'Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data', *Big Data & Society* **4**(2).
URL: *https://doi.org/10.1177/2053951717743530*

Venerandi, A., Quattrone, G., Capra, L., Quercia, D. & Saez-Trumper, D. (2015), Measuring urban deprivation from user generated content, *in* ‘Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW 2015)’.

Wachter, S., Mittelstadt, B. & Russell, C. (2018), ‘Counterfactual explanations without opening the black box: Automated decisions and the gdpr’, *Harvard Journal of Law & Technology* **31**(2).

Walton, D. (2004), ‘A new dialectical theory of explanation’, *Philos. Explor.* **7**(1), 71–89.

Wang, B. & Zhenqiang Gong, N. (2018), Stealing hyperparameters in machine learning, *in* ‘Proceedings of the IEEE Symposium on Security and Privacy (S&P)’, pp. 36–52.

Wang, D., Yang, Q., Abdul, A. & Lim, B. (2019), Designing theory-driven user-centric explainable ai, *in* ‘Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems’, pp. 1–15.

Wang, Y. & Kosinski, M. (2018), ‘Deep neural networks are more accurate than humans at detecting sexual orientation from facial images’, *J. Pers. Soc. Psychol.* **114**, 246–257.

Wei Koh, P. & Liang, P. (2017), Understanding black-box predictions via influence functions, *in* ‘Proceedings of the 2017 International Conference on Machine Learning (ICML 2017)’, pp. 1885–1894.

Wesolowski, A., Eagle, N., Tatem, A., Smith, D., Noor, R. & Buckee, C. (2012), ‘Quantifying the impact of human mobility on malaria’, *Science* **338**(6104), 267–270.

Wesolowski, A., Qureshi, T., Boni, M., Sundsøy, P., Johansson, M., Rasheed, S., Engo-Monsen, K. & Buckee, C. (2015), ‘Impact of human mobility on the emergence of dengue epidemics in pakistan’, *Proc. Natl. Acad. Sci. USA* **112**, 11887–11892.

Willson, M. (2017), ‘Algorithms (and the) everyday’, *Inf. Commun. Soc* **20**(1), 137–150.

Wilson, R., Erbach-Schoenengerg, E., Albert, M., Power, D., Tudge, S., Gonzalez, M., Guthrie, S., Chamberlain, H., Brooks, C., Hughes, C., Pitonakova, L., Buckee, C., Lu, X. Wetter, E., Tatem, A. & Bengtsson, L. (2016), ‘Rapid and Near Real-time Assessments of Population Displacement Using Mobile Phone Data Following Disasters: The 2015 Nepal Earthquake’, *PLOS Current Disasters* .

Xu, K., Park, D., Yi, C. & Sutton, C. (2018), ‘Interpreting deep classifier by visual distillation of dark knowledge’, *arxiv preprint arXiv:1803.04042* .

Yang, Q., Liu, Y., Chen, T. & Tong, Y. (2019), ‘Federated machine learning: Concept and applications’, *ACM Transactions on Intelligent Systems and Technology* **10**(12).

Yeh, C.-K., Kim, J., Yen, I.-H. & Ravikumar, P. (2018), Representer point selection for explaining deep neural networks, *in* ‘Proceedings of the 2018 International Conference on Advances in Neural Information Processing Systems (NeurIPS 2018)’, pp. 9311–9321.

Zarsky, T. (2016), ‘The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making’, *Sci. Technol. Hum. Values* **41**(1), 118–132.

Zeiler, M. & Fergus, R. (2014), Visualizing and understanding convolutional networks, *in* ‘Proceedings of the European Conference on Computer Vision (ECCV 2014)’, pp. 818–833.

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 2012)’, pp. 325–333.

Zhang, Q., Sun, K., Chinazzi, M., Pastore y Piontti, A., Dean, N., Rojas, D., Merler, S., Mistry, D., Poletti, P., Rossi, L., Bray, M., Halloran, M., Longini Jr, I. & Vespignani, A. (2017), ‘Spread of zika virus in the americas’, *Proc. Natl. Acad. Sci. USA* **114**(22), 4334–4343.

Zhang, Q., Yang, Y., Ma, H. & Wu, Y. (2019), Interpreting cnns via decision trees, *in* ‘Proceedings of the 2019 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2019)’, pp. 6261–6270.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. (2016), Learning deep features for discriminative localization, *in* ‘Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)’, pp. 2921–2929.

Zuboff, S. (2019), *The Age of Surveillance Capitalism*, Public Affairs.

Highlights

- Artificial Intelligence (AI) algorithms are increasingly used to make or assist in making decisions with significant impact in people's lives.
- Algorithmic decision-making is not exempt from risks and limitations: it has been shown to lead to privacy invasion, opacity, and discrimination.
- We propose three requirements to achieve a human-centric AI: (1) privacy-preserving algorithms and data cooperatives; (2) human-understandable explanations; and (3) algorithmic fairness approaches connected with different worldviews.
- We call for a multidisciplinary effort of researchers from machine learning, human-computer interaction, cognitive sciences, ethics and philosophy, and the law as well as of policy makers and citizens.