AI & SOCIETY (2024) 39:1397–1409
https://doi.org/10.1007/s00146-022-01616-7

OPEN FORUM

# The paradoxical transparency of opaque machine learning

**Felix Tun Han Lo**[1]

Received: 27 January 2022 / Accepted: 29 November 2022 / Published online: 19 December 2022


**Abstract**
This paper examines the paradoxical transparency involved in training machine-learning models. Existing literature typically critiques the opacity of machine-learning models such as neural networks or collaborative filtering, a type of critique that parallels the black-box critique in technology studies. Accordingly, people in power may leverage the models' opacity to justify a biased result without subjecting the technical operations to public scrutiny, in what Dan McQuillan metaphorically depicts as an "algorithmic state of exception". This paper attempts to differentiate the black-box abstraction that wraps around complex computational systems from the opacity of machine-learning models. It contends that the degree of asymmetry in knowledge is greater in the former than the latter. In the case of software systems, software codes are difficult to understand as only software experts with sufficient domain knowledge are equipped to formulate a sound critique. In contrast, the meanings of trained parameters in a machine-learning model are obscure even to the data scientists who configure and train the model. Hence, the asymmetry of knowledge lies only in how data examples are collected, the choice and configuration of machine-learning models, and the specification of features in model design. Under the trend of algorithmic decision-making proliferating with machine-learning heuristics, the paper contends that the more symmetric distribution of knowledge in machine learning could lead to a more transparent production process if proper policies are in place.

**Keywords** Machine learning · Algorithmic governance · Algorithmic opacity · Algorithmic state of exception · Asymmetry of knowledge · Digital democracy

## 1 Introduction

Facing a future of widespread infiltration of algorithmic decision-making in our sociotechnical milieu, scholars and the public are increasingly concerned over the lack of transparency in how automated decisions actually come about. Brown et al. (2021, p. 5) developed a scheme for auditing AI systems based on scores assigned to a variety of characteristics that include the transparency of architecture, the explainability and interpretability of the algorithm, the transparency of whether an algorithm is used, and the transparency of how well the collection and the use of data for the algorithm are known to stakeholders. Similarly, German Federal Office for Information Security recently published the white paper, "Towards Auditable AI Systems" (Berghoff et al. 2021), which has a section called "Explaining Black Box AI Models" (2021, pp. 17–18). This section in the white paper raises the concern that the "inner workings of [AI] models …. do not usually lend themselves to human interpretation." It then contends that "being able to explain and interpret the decisions of an AI model can be important for a number of reasons. These reasons range from finding faults, weaknesses and limitations of a model … to fulfilling requirements for transparency, as for instance mandated by the EU General Data Protection Regulation, and to gaining new insights from large data sets in science and in the economy" (2021, p. 17). To the authors, the opacity of AI models can shield and conceal the infringement of data protection regulation. Overcoming these current limitations of AI models requires "new methods … for explaining complex AI models like neural networks," giving rise to an emerging research field called Explainable AI (XAI) (2021, p. 17).

Perhaps the most convincing reason behind advocating algorithmic transparency concerns the problem of trust. It is difficult for people to put their trust in automated decisions that do not provide human-understandable rationale behind these decisions. This problem of trust is especially

✉ Felix Tun Han Lo
lofelixl@sfu.ca

[1] School of Communication, Simon Fraser University, Burnaby, Canada

relevant in a judicial context where due processes traditionally require human understanding on how decisions come about, or in medical practices where doctors and patients find it difficult to trust algorithmic diagnoses associated with insufficient scientific rationale (Fainman 2019; Longoni et al. 2019; Sullivan 2020). The issue of algorithmic opacity is exacerbated by the increasing adoption of machine learning (ML) models. These models come with varying degrees of comprehensibility on the human-understandable reasons behind their decision-makings. Some of these models, such as neural networks or collaborative filtering, are inherently more opaque than simpler models such as decision-tree or linear regression. Due to this contradiction between the demand for human-scale understanding and the inherent opacity of complex ML models, there is an ample amount of scholarly literature on problematizing the ethical impacts of machine learning (Datta et al. 2015; Mittelstadt et al. 2016), on auditing algorithms (Diakopoulos 2016; Sandvig et al. 2014), and on overcoming the issue of algorithmic opacity (Carabantes 2020; Chan 2021; Fainman 2019; Lee et al. 2021; Müller 2021; Watson 2021; Zednik and Boelsen 2021). It seems awfully wasteful not to adopt the technique of opaque machine learning over the issue of opacity when these algorithms appear to outperform human decision-making in numerous contexts (Longoni et al. 2019; McKinney et al. 2020; Silver et al. 2018; Stokes et al. 2020). Thus much effort has been devoted to XAI, an emerging field of research that develops new techniques for deriving "post-hoc" explanations, which do not attempt to open the black box but to conduct a-posteriori analyses, deducing human-understandable factors from the models' performances in actual practices (Chan 2021; Lee et al. 2021; Watson 2021).

In the midst of this negativity surrounding the opacity of ML models, this paper wants to bring out an alternative perspective on the issue of opacity, one that inquires about the symmetry of knowledge between the producers and consumers of technical systems, in the sense that knowledge is power (e.g., Jarrahi et al. 2021, p. 8). Rather than vilifying the increasing adoption of ML models as a precursor to digital authoritarianism, this paper contends that the lack of transparency in the inner workings of a technical system, to the producers and the consumers alike, actually undermines the asymmetry of knowledge between them. In this respect, an increasing dependence on opaque ML models can paradoxically render a more transparent production process if proper policies and practices for third-party audits can be put in place. From this alternative perspective, the research community ought not to be overly concerned with interpreting what goes on inside the black box of ML models. Instead, this paradoxical transparency can serve as the basis for aiming toward a more democratic digital milieu.

In the following, I begin by expounding on the ideal of algorithmic transparency through the analysis by Dan McQuillan (2015), who uses the metaphor "algorithmic state of exception" to characterize an algorithmic governance that lacks transparency. I then discuss the philosophical critique of automatic societies by Bernard Stiegler and that of algorithmic governance by Antoinette Rouvroy and Thomas Berns. Their critiques are in line with McQuillan's, but their Simondonian philosophy also opens up the possibility of emancipation. Contemplating the practical steps toward emancipation requires a deeper understanding about the opacity associated with digital algorithms. Thus I draw from the distinction between three forms of opacity by Jenna Burrell (2016). The form of opacity she calls cognitive mismatch is typically regarded as the "most worrisome" of the three (Carabantes 2020, p. 310), and can be found in both complex large-scale software and opaque ML models. I then argue, the opacity of cognitive mismatch is "most worrisome" because it appears to lend support to the opacity of institutional concealment, but this relationship between the two forms of opacity is actually contingent and can be subverted if proper policies for third-party audits are put in place. My contention is that, given equal access to the training process of a complex ML model that appears opaque to human understanding, insiders do not possess a distinct advantage in understanding the model than outsiders. If governments can stipulate regulations to enforce third-party audits, the lack of transparency in complex ML models may paradoxically bring a greater degree of process transparency in the training of the models. I conclude by discussing some practical considerations in working toward the goal of process transparency in machine learning.

## 2 Algorithmic state of exception

In his paper "Algorithmic states of exception" (2015), Dan McQuillan puts forth a thorough analysis of the issue of opacity for automated algorithmic decisions. He uses a powerful metaphor, the state of exception, to elucidate the social condition wherein automated decisions substitute human decisions in the subtle enforcement of social regulations and governance. A "state of exception," first introduced by German philosopher Carl Schmitt and further developed by Italian philosopher Giorgio Agamben (2005), refers to the covert scheme of a government to disguise its transition into an authoritarian regime by suspending the rule of law in the name of public good. Agamben investigates the way a government uses crises as excuses to suspend civilians' rights, which are normally protected under the constitutions in ordinary times (2005, p. 32). His prime example comes from Nazi Germany: After the Reichstag Fire, the Third Reich entered and continued to operate as a state of exception over the entire 12-year regime under Hitler's rule (2005, p. 15). In today's sociotechnical milieu that increasingly

jettisons human judgment in favor of algorithmic predictions, McQuillan wants to raise awareness of an emerging form of governance analogous to a state of exception. To him, "pervasive tracking and data-mining are leading to shifts in governmentality that can be characterised as an algorithmic state of exception," and these shifts come in the "[accelerated] use of prediction as a form of governance" (McQuillan 2015, p. 564).

According to McQuillan, we need to look beyond the questioning of data practices (Boyd and Crawford 2011) and examine "the nature of the material–political apparatus that connects data to decision-making and governance" (McQuillan 2015, p. 565). This material–political apparatus is "undergoing a significant shift in the system of relations at several levels: in architectural forms (forms of database structures), administrative measures (as algorithms), regulation (as algorithmic regulation) and laws (as states of exception)" (2015, p. 566). At the architectural level, the form of database structure is shifting from relational database to NoSQL database. In a relational database, data are structured according standardized schema. But the structuring of raw data into pre-specified formats would cut out nuances in the original raw data. These nuances, which would be treated as noise in traditional computing systems, are now regarded as constituents of complex data patterns that ML algorithms can recognize. To preserve nuances, structured data can be replaced by schemaless data blobs, which can be stored in NoSQL databases. Because structured data are ordered with human-interpretable features while schemaless data blobs would appear meaningless to the human intellect, the new form of database structure can be characterized as more opaque than a relational database.

The next level up is "the forms of algorithmic processing known as data-mining and machine learning" (2015, p. 567), which draw inferences from datasets to make probabilistic predictions. As McQuillan points out, this "predictive turn introduces new risks because of the glossed-over difference between correlation and causation" (2015, p. 567). A probabilistic algorithm may produce results that are false positives, but when the reasoning behind false positives is obscure, as is the case with machine learning, differences between subpopulations would be glossed over and certain subpopulations may be exposed to elevated risk. This predicament is amplified as algorithmic predictions become pervasive in the social domain, where "correlation becomes a basis for correction or coercion" (McQuillan 2015, p. 568). McQuillan calls this "algorithmic regulation," in which algorithmic predictions become preemptive as a form of social regulation. For instance, "[i]n Chicago, an algorithmic analysis predicted a 'heat list' of 420 individuals likely to be involved in a shooting, using risk factors like previous arrests, drug offences, known associates and their arrest records. They received personal warning visits from a police commander" (2015, p. 568). In this practice, the police can act with the force of the law without complying with the law: "[P]reemptive measures are applied without judicial standards of evidence and police are sometimes prepared to act on the basis of an algorithm while asserting that they do not understand the reasoning process it has carried out" (2015, p. 568).

Finally, when a society opts for algorithmic regulation as a generic mechanism of social governance, it would become an algorithmic state of exception. In this approach to algorithmic governance,[1] government agencies would allegedly identify and modify social problems in the same way Google and Facebook use statistical feedback loops to police their systems against malware and spam (2015, p. 568). For instance, "[w]hen Facebook's algorithms decide that an unlucky user has violated their Terms of Service, that person discovers he or she has no recourse; there is no real explanation of why they were excluded, and no one to whom he or she can appeal" (2015, p. 569). McQuillan contends that "predictive algorithms increasingly manifest as a force-of [the law] which cannot be restrained by invoking privacy or data protection" (2015, p. 570). When this approach is extended to social governance, the operations of an algorithmic apparatus "have the potential to create social consequences that are unaddressed in law" (2015, p. 570). The algorithmic actions would have the force of the law even though they are not of the law, which according to Agamben is the signature of a state of exception as opposed to a dictatorship (McQuillan 2015, p. 570).

Over the past few years, algorithmic governance is an emerging trend around the world. An epitome of this trend is China's plan to build a Social Credit System (Zou 2021), but we can also see it in numerous other contexts. The UK-integrated health and social care teams transformed the transactions of everyday care work into big data, which in turn enabled the governance of complex service arrangements (Huby and Harries 2021). The Australian cities Perth and Darwin are interlaced with an increasing number of sensing technologies in the governance of the two cities, which are experimenting with algorithmic analytics as a measure for improving efficiency and security (Smith 2020). There are also questions on whether contact tracing and data-driven surveillance in the recent pandemic outbreak are forms of intrusive monitoring introduced in the name of public security and social need (Pūraitė et al. 2020). In fact, Agamben himself sees such measures as manifesting "the growing tendency to use the state of exception as a normal governing paradigm," listing as an example "[t]he enforcement of

[1] Note that McQuillan does not employ the term "algorithmic governance" in "Algorithmic State of Exception" (2015) but he does use it in "Algorithmic Paranoia and the Convivial Alternative" (2016).

quarantine and active surveillance on individuals who had close contact with confirmed cases of infection" (Agamben 2020). In an algorithmic state of exception, when predictive ML algorithms become preemptive, and when preemptive algorithms are adopted as the ideal mechanism of governance, the algorithmic apparatus effectively operates above the law. Social biases and discrimination can be encoded in opaque algorithms without being subjected to the law.

## 3 Algorithmic governance and the path toward emancipation

McQuillan was not the first to raise concerns about algorithmic governance. According to Bernard Stiegler (2016), this term was first coined by Antoinette Rouvroy and Thomas Berns, who formulated an insightful Simondonian and Foucauldian critique of computational capitalism in "Algorithmic governmentality and prospects of emancipation" (2013). They point out that "our behaviours have never been so processed—observed, recorded, classified, evaluated—, underpinned by codes of intelligibility and criteria that are completely opaque to human understanding, as it is now on this statistical basis" (Rouvroy and Berns 2013a, b, p. XIX). This opaque behavioral processing results in a new form of governance that can be characterized as 'a-normative'—as having a normative effect without producing recognisable social categories of the 'normal' (Crogan 2019). It is a new form of social control that "focuses not on individuals, on subject, but on relations" (Rouvroy and Berns 2013a, p. V). Accordingly, social relations between individual subjects would be analyzed as non-subjective statistical correlations. Rather than classifying individual profiles into recognizable social categories, an enormous amount of computational resources today is devoted to statistical computing that identifies correlations between data points across individual profiles. Such statistical processing creates "a sort of statistical 'double' of both subjects and 'reality'" (2013a, p. V), substituting the living relations between individual subjects with the statistical correlations between thousands or millions of these statistical "doubles".

Stiegler compares the algorithmic governmentality of Rouvroy and Berns with the governmentality of Foucault. Whereas Foucault's biopolitics "consists, to an extent at least, in 'taking care' of life so as to be able to exploit it," today's digital algorithms and infrastructures aim "not at mass-deception, nor at 'neutralizing or inactivating' the masses, but at *exploiting them as resources of which they take no care*, and, from this perspective, exploiting them without any biopolitics" (Stiegler 2016, p. 97, emphasis in original). In our digital milieu, the masses are reduced to the metadata perpetually produced through online behavior, and digital algorithms exploit these digital traces as resources. This new form of governmentality does not take care of life as Foucault's biopolitics does, and Stiegler characterizes this "carelessness" as "(a)biopolitical" (2016, p. 97).

According to Stiegler, our capitalist society has been transitioning from the proletarianization of the industrial labourers to that of the consumers in the late twentieth century. Now, in the early twenty-first century, it is transitioning from the proletarianization of consumers to that of theoretical knowledge: "[H]yper-industrial societies have now been undergoing the proletarianization of theoretical knowledge, just as broadcasting analogue traces via television resulted in the proletarianization of life-knowledge" (Stiegler 2016, p. 25). The predictions of scientific theories to care for life are being supplanted by the predictions of algorithmic computations capable of identifying patterns and correlations in heterogeneous data. Citing from Rouvroy and Berns (2013b, p. 173), Stiegler wrote, "these continuously traced and collected statistics constitute and mobilize an '(a)normative and (a)political rationality based on the harvesting, aggregation and automatic analysing of data in massive quantities to model, anticipate and affect in advance possible behaviours'" (2016, p. 106).

But in the process of isolating meaningful knowledge from heterogeneous data that appear to be statistically correlated, computations discard the rest of the data or any meaning in life that do not belong to the correlations. Rouvroy and Berns, as well as Stiegler, deliberate on the philosophical implication of this technical operation by referring to the philosophy of Gilbert Simondon. In their arguments, what has been discarded are the inconsistencies, tensions, and disparations inherent in human nature and in the social milieu, and such tensions are the sources of potentiality for the co-evolution of the social milieu and technology. The social milieu filled with conflicts and tensions on the one hand, and technological development, on the other hand, need to maintain a symbiotic relationship to maintain a healthy relationship between humans and machines. The term Simondon uses to denote this sociotechnical co-evolution is "transindividuation".[2] For Rouvroy and Berns, because digital algorithms discard the inconsistencies and tensions inherent in the social milieu, algorithmic governmentality would lead to the elimination of "transindividuation." Stiegler concurs, contending that the 'a-normative character of

[2] "The technical object taken according to its essence, which is to say the technical object insofar as it has been invented, thought and willed, and taken up [*assumé*] by a human subject, becomes the medium [*le support*] and symbol of this relationship, which we would like to name *transindividual*" (Simondon 2016, p. 252). The transindividual reality is an inter-individual collective reality in which inter-human relations are "created through the intermediary of the technical objects" (2016, p. 254), and the relations with technical objects create "a coupling between the inventive and organizational capacities of several subjects" (2016, p. 258).

algorithmic governmentality' would lead to "an annihilation of transindividuation (as a process of the everyday realization of normative life in the sense of Canguilhem), insofar as the latter results from constant interpersonal co-individuations, consolidated by impersonal retentional systems, but ones that are visible and open to critique—but that are also internalized by instituted knowledge" (2016, p. 107). In the past, publications and legislations consolidated "constant interpersonal co-individuations" into normative behavior. Yet, despite the issues of the normativity and of the instituted knowledge, these tertiary retentional systems[3] were "visible and open to critique." There was enough time for people to digest and critically examine such social changes and their effects to formulate substantial critiques. As such, a "process of everyday realization of normative life" still leaves room for transindividuation, in which meanings and consensus are destructed and reformulated. But now, "[t]he moment of reflexivity, critique and recalcitrance necessary for subjectification to form seems to constantly become more complicated or to be postponed" (Rouvroy and Berns 2013a, p. X). As a result, "intermittence, that is, the time of individuation, has been suspended, and that this occurs through the constitution of a technology that automatically and performatively generates protentions" (Stiegler 2016, p. 101). In algorithmic governmentality, social critique and reflexivity give ways to algorithmically induced "protentions," as people's intuition, anticipation of the future, and decision-making are becoming increasingly manipulated and governed by digital algorithms.

Up to this point, the social critique by Stiegler and that by Rouvroy and Berns seem to be in line with McQuillan's critique, which identifies the parallel between the opacity of algorithmic governance with the political theory on the state of exception by Agamben. Where Stiegler differs from McQuillan is on how emancipation may come about. In *Automatic Societies, vol 1* (2016), Stiegler characterizes technology as *pharmakon*, which can mean either poison or remedy in Greek. Thus he poses this question about algorithmic governance, "is it nevertheless possible to effect a reversal, through which the trace could again become an object of social investment" (2016, p. 24)? While Stiegler never fully developed what this "reversal" may be, he did suggest that "systems must be built and implemented that are dedicated to the individual and collective interpretation of traces—including by using automated systems that enable analytical transformations to be optimized, and by supplying new materials for synthetic activity" (2016, p. 141). The phrase "individual and collective interpretation of traces" seems to imply a democratization of apparatuses that collect and analyze data traces. If people can get hold of their own digital profiles or if they can configure the pattern recognition apparatuses that analyze their daily behaviors, then the technical assemblage can serve their own self-understanding or self-improvement. Such a democratization of apparatuses would then address his philosophical critique on transindividuation, as it would restore the symbiotic relationship between the social milieu and how technology evolves.

## 4 Three forms of opacity

Stiegler hints at a vision of what a democratic digital milieu may look like, but in order to come up with practical steps "to effect a reversal," we need to first gain a deeper understanding about opacity and transparency. Jenna Burrell (2016, p. 2) draws a distinction between three forms of opacity: (1) opacity as intentional corporate or institutional self-projection and concealment with the possibility for knowing deception; (2) opacity as technical illiteracy, in a social context where writing and reading code is a specialist skill and; (3) an opacity that stems from the cognitive mismatch between mathematical optimization in a high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of semantic interpretation. Kathleen Creel (2020, p. 587) also distinguishes three types of transparency in computational systems, between (1) the transparency of algorithms, (2) the static software code that implements the algorithms, and (3) the runtime processes that correspond to the static code. If we compare the two distinctions, we can see that Creel's distinction can be read in parallel with Burrell's: Overcoming the opacity of intentional corporate or state secrecy demands transparency in the specification of algorithms; social improvement in technical literacy would result in the transparency of software code; the high-dimensionality of machine learning operates within the runtime processes of an algorithm. Note that the public approval of an algorithmic state of exception can be strengthened by confounding the three forms of opacity, for instance, by concealing corporate deceptions in the name of opaque computer systems.

Establishing policies and regulations can help alleviate opacity as intentional corporate concealment, and technical illiteracy can be curtailed by a widespread educational curriculum in which young kids learn how to code software. But as Manual Carabantes (2020, p. 310) remarks, it is the third form, the opacity as "cognitive mismatch," that is "the most worrisome, since it also prevents the engineers who develop certain ML models to understand how their own creation work." This third form of opacity may be exhibited in the complexity of large-scale computer systems or in the

[3] Tertiary retention is Stiegler's term for denoting a type of permanent social memory that is possible through technology. Writing, printing, database, YouTube, Facebook, are all examples of tertiary retentional systems.

incomprehensibility of how ML models operate (Burrell 2016, pp. 4–5). Indeed, existing literatures often discuss the opacity of software code and that of ML models. Such research can be categorized into two strands. One strand (e.g., Brill 2015; Diakopoulos 2016; Pasquale 2015) advocates technical transparency, covering the transparency of platform designs, algorithmic mechanisms, software coding and runtime behavior. The other strand is critical of the over-emphasis on technical transparency. Some of these critics (e.g., Ananny and Crawford 2018; Seaver 2017) argue from the socio-technical perspective that sees an algorithmic system not just as "code and data but an assemblage of human and non-human actors" (Ananny and Crawford 2018, p. 983). Other critics engage with the problem from the perspective of scientific understanding, for whom understanding is only meaningful at the appropriate level of abstraction. For instance, Emily Sullivan (2020, p. 1) argues that "it is a lack of scientific and empirical evidence supporting the link that connects a model to the target phenomenon that primarily prohibits understanding." To illustrate, she shows that the "link uncertainty" is much smaller in a neural network that successfully distinguishes malignant tumors from benign ones based on visual patterns, than in a neural network that identifies sexual orientation based on profile images from online dating sites. The emerging research on Explainable Artificial Intelligence (XAI) is situated somewhere in-between, moving gradually from the first toward the second strand, with practical successes being found mostly in post-hoc interpretability of the algorithms (Carabantes 2020; Fainman 2019).

This paper is more closely associated with the second strand of research, but rather than "enact[ing] algorithms ethnographically as heterogeneous and diffuse sociotechnical systems" (Seaver 2017, p. 1), discussing how to attain empirical evidences for scientific understanding at a high level of abstraction (Sullivan 2020), or exploring post-hoc explanations of technological black boxes, I want to draw attention to the potential of algorithmic opacity for bringing about a more democratic digital milieu. In McQuillan's algorithmic state of exception, there is an implicit assumption that the third form of opacity as "cognitive mismatch" would exacerbate the first form of opacity as corporate or institutional concealment. Likewise in Stiegler's critique, automated algorithms opaque to human understanding would empower the proletarianization of human knowledge, thought, and decisions, which also conceals institutions' intention to manipulate people's protention and behavior. But is this assumption about the two forms of opacity necessarily true? In the following sections, I will problematize this assumption by arguing that the opacity of "cognitive mismatch" may actually contribute toward undermining the opacity of institutional concealment. In Stiegler's philosophy, it is the opacity of concealed institutional manipulation that needs to be addressed in order for social critiques to have their say on the evolving development and deployment of digital algorithms. Addressing this opacity would in turn restore the symbiotic relationship between the social milieu and how technology evolves. Therefore, we can "effect a reversal" on the social implications of the opaque digital algorithms if the opacity of "cognitive mismatch" can be made to weaken the opacity of institutional concealment. This can possibly be achieved by stipulating policies on third-party audits, which would then be the first practical step toward the democratization of apparatuses that Stiegler envisions.

## 5 The unique opacity of machine learning

As mentioned earlier, Burrell's third form of opacity as a cognitive mismatch may come from either the lack of transparency in ML models or the complexity of large-scale computer systems built primarily via "programming by hand." Even though the operations for both types of systems may not be fully comprehensible for human-scale understanding, we can still differentiate the types of opacity associated with the two modes of technical actions. In this section, I will expound on this difference. Doing so will help us understand why the training process of ML models is associated with a greater degree of symmetry between the knowledge of the programmers and that of the users, and why the process of training an ML model, which encapsulates and absorbs much of the programming complexity, can be made more transparent than the traditional process of software development.

When the technology of computer software was still in its infancy, Marvin Minsky already commented on its opacity: "The programmer himself … may have only a very incomplete understanding of when and where in the course of the program's operation these procedures will call on each other. … For try as we may, we rarely can fully envision, in advance, all the details of their interactions" (Minsky 1967, p. 6). And such ignorance would only grow after numerous lifecycles of patches and fixes, as "the programmer begins to lose track of internal details and can no longer predict what will happen" (1967, p. 8). Drawing from Minsky's observations, Joseph Weizenbaum (1976) laid out the risk in society's growing reliance on computer systems. He lamented that "decisions are made with the aid of, and sometimes entirely by, computers whose programs no one any longer knows explicitly or understands. Hence no one can know the criteria or the rules on which such decisions are based" (1976, p. 236). Because no one can have a detailed understanding of the inner workings of a computer system, programmers would tend to avoid substantial modifications. Such computer systems are "immune to change" and "can therefore only grow. And their growth and the increasing reliance

placed on them is then accompanied by an increasing legitimation of their knowledge base" (1976, pp. 236–237). Minsky's and Weizenbaum's arguments on the complexity and opacity of software were further extended by Fred Brook in *The Mythical Man-Month* (1975). Brook argues that additional manpower may not necessarily overcome a schedule delay if the software development project is large and complex. The new programmers, lacking tacit knowledge of the code, may introduce unnecessary complexity and software bugs that would further delay the project.

A couple of decades later, these arguments about the intrinsic opacity of software systems were challenged by proponents of the open-source software movement. According to what Eric Raymond dubs as the "Linus's Law," "[g]iven enough eyeballs, all bugs are shallow" (2001, p. 30). Thus "every problem will be transparent to somebody" (2001, p. 30), and "many eyeballs tames complexity" (2001, p. 33). Raymond uses the metaphors cathedral and bazaar to put in contrast the traditional understanding of programming from the open-source perspective. In the cathedral-builder view of programming, "bugs and development problems are tricky, insidious, deep phenomena. It takes months of scrutiny by a dedicated few to develop confidence that you've winkled them all out" (2001, p. 31). In the bazaar view, "bugs are generally shallow phenomena, —or, at least, that they turn shallow pretty quickly when exposed to a thousand eager co-developers pounding on every single new release" (2001, p. 31). According to Raymond, the proven robustness of Linux operating system over the years is a substantial evidence that confirms the validity of the bazaar view and the "Linus' law."

Nevertheless, only the most popular open-source projects would attract "enough eyeballs," leading to Elias Levy's remark that "[s]ure, the source code is available. But is anyone reading it" (2000).[4] Andy Ozment and Stuart Schechter (2006) report that in the OpenBSD source, foundational vulnerabilities have a median lifetime of at least 2.6 years, which seems to refute Raymond's argument. In "Is Open Source Security a Myth" (2011), Guido Schryen conducts an empirical study that further challenges the belief that open-source software is inherently more secure than proprietary software. The study compares 17 well known and widely deployed software packages regarding published vulnerabilities and software vendors' patching behavior, and concludes that "open source and closed source software do not significantly differ in terms of the severity of vulnerabilities, the type of development of vulnerability disclosure over time, and vendors' patching behavior" (2011, p. 140).

From the early pioneers in computer sciences to the controversial debates on open-source software, I have provided a brief survey of the extensive social science and critical engagement around the controversies of software and its making, as well as the collective and organized efforts of open-source proponents to dispel the opacity of software systems. Regardless of whether software systems are inherently transparent or opaque, the opacity of machine learning works differently. Even when human programmers cannot fully understand how their software programs operate in full details, as per Minsky, they nevertheless have a good grasp of the symbolic logic behind the functional behavior that meets the specified requirement. Software code are textual inscriptions of the virtual objects and functions inside the programmers' minds, and automatic compilers would translate the coded logic into operational software programs. The complex logic is still the result of integrating and aggregating programming statements in humanly understandable syntax. This is not necessarily the case with machine learning. It is not that machine learning does not "reason," but its "reasoning" operate differently from human reasons. In cases where the "reasoning" in machine learning surpasses human reasoning, the two modes of "reasonings" are not directly translatable.

This is most evident in deep learning. As the AI expert Kai-Fu Lee remarks, "deep learning's decisions are based on complex equations with thousands of features and millions of parameters. Deep learning's 'reason' is basically a thousand-dimensional equation, trained from large quantities of data. This 'reason' for producing a given output is too complex to explain fully to a human" (Lee and Chen 2021, p. 56). A deep neural network is composed of many hidden layers of nodes. We can think of these hidden nodes as sub-features that contribute to the predictive probability. Training a neural network would generate weights or parameters associated with each node over many iterations of back propagations. In the case of a neural network with only a hundred nodes, it is somewhat possible to confer meanings to these weights and nodes. Such a small neural network can be trained to recognize handwritten digits from one to ten, and a data scientist may inspect the weight of each node and tries to guess how certain nodes may be associated with certain pen strokes, for instance, a straight line down here, a slant there (see Burrell 2016, p. 6–7). But these are, at best, just educated guesses. In practice, the number of hidden layers and nodes in a deep neural network may be in the thousands, millions, or even more. For instance, GPT-3 (GPT stands for "generative pre-trained transformers"), released by OpenAI in 2020, produced a gigantic model with 175 billion parameters (Lee and Chen 2021, p. 152). Google Brain, released one year later, is a language model with 1.75 trillion parameters (Lee and Chen 2021, p. 158). Such neural networks would presumably discover billions or trillions

[4] The article "Six open source security myths debunked—and eight real challenges to consider" (Heath 2013) also makes a similar argument.

of very fine-grained "features," and it is beyond the human mind to interpret such complex "reasoning".

The root of this inherent opacity of ML models comes from the methodology of machine learning, whose goal is different from the goal of statistical analysis. Even though a number of ML algorithms are built using statistical models such as linear regression or logistic regression, machine learning has a different goal from statistical analysis. Whereas statisticians want to infer scientific understanding about some phenomena, machine learning is an engineering endeavor for repetitively making accurate predictions (Malik 2020, p. 20). The principal objective of a statistical inquiry is to identify possible correlations between variables, which are indicated by the coefficients of the variables in a trained model. Conversely, the principal objective of machine learning is to inductively compute probabilistic predictions under repeatable circumstances. It is to identify data patterns in training sets and to find training sets that are representative of the general population such that discovered patterns can be generalized into predictions. To this goal, the statistical correlations between specific variables are irrelevant. For instance, a statistician may be interested in the correlation coefficient between the color or the size of a tumor and its malignancy, whereas the value of this coefficient is irrelevant to an ML algorithm, which analyzes the entire image of the tumor to compute the probability of malignancy.

The loss in the significance of the coefficient, when statistical methods such as linear regression are appropriated to ML algorithms, is apparent in the common ML technique called "regularization." Regularization is a technique for reducing overfitting and to make the algorithm more generalizable. The coefficients associated with input variables would be different for every chosen regularization rate. For example, a statistician may configure a linear regression model with hundreds of features to study the correlations between certain features and the sales price of condos. An AI engineer may similarly train a linear regression ML model on the same set of features to predict the sale prices of condos. Such features may include the size, the location, the number of rooms, the distance to nearby public transits, and so on. When an AI engineer applies regularization to reduce the problem of overfitting, the values of the parameters or coefficients would be reduced to smaller values, and the relative strengths of the parameters may not stay the same. The coefficient for location may be larger than that for the number of rooms prior to applying regularization, but smaller after applying regularization. Though examining the parameters in regularized linear regression may still reveal some ideas about how various features may impact the final prediction, such knowledge is not as scientifically rigorous as the study of coefficients in a statistical study.

While all ML models are typically trained with regularization, the opacity of models' reasoning is even more noticeable in models that are more algorithmic than statistical. In addition to neural networks in deep learning, collaborative filtering for recommender systems and support vector machines (SVM) are ML algorithms that are more algorithmic than statistical. A recommender system seeks to predict the rating that a user would give to an item. Collaborative filtering is one possible algorithm for implementing a recommender system. It computes rating predictions by autonomously discovering features that characterize the content of an item, and these features may differ from typical human descriptions. For instance, a person may describe a given movie as a romantic comedy while the algorithm would characterize it with more fine-grained features. Both neural networks and collaborative filtering are capable of auto-discovering relevant "features," the meanings of which would be very difficult for humans to decipher. SVM is another ML model designed to recognize complex data patterns by algorithmically transforming a small feature set into a very large feature set. Due to this transformation, it would also be very difficult to trace the impacts of particular input features on the final prediction. Even linear or logistic regression would require the introduction of high-ordered polynomial terms if the engineering objective is to recognize complex data patterns.[5] In all these ML algorithms, the goal is not to study statistical correlations to reveal general scientific principles but to build an automaton that can recognize complex data pattern. All these algorithms can be designed in an opaque fashion because the goal of finding correlations between specific variables is no longer relevant.

Deep learning, collaborative filtering, or support vector machine are complex ML models that can recognize patterns in a massive amount of data with a type of reasoning that surpasses human reasoning. Whereas programming by hand cannot produce applications such as driverless cars, which involve innumerable scenarios that cannot be systematized as software code, machine learning is apt at solving such problems. While theoretically we can limit ourselves to develop simpler ML models that are translatable to human-scale reasoning, this self-limitation would also curtail the power of machine learning. In fact, an ML model translatable to human-scale reasoning would in theory be programmable by hand. This is why, even though certain ML models

[5] The complexity of an ML model is indicated by the number of parameters that the model can be trained with. Every ML model can be made arbitrarily complex. For instance, adding polynomial features can artificially expand the feature set of linear regression. Too many parameters may lead to overfitting, which can be attenuated by training the model with a very large training set. Thus Michele Banko and Eric Brill (2001) did an experiment, comparing the performances between different ML models trained with varying sizes of data set. They found out that all models give remarkably similar performances when there is enough data. Hence, they conclude, "it's not who has the best algorithm that wins. It's who has the most data.".

may be translatable to human-scale reasoning, these models rarely find their ways to the industry in practice, which is the context that this paper is most concerned with.

## 6 Process transparency

The preceding section explains that the opacity of machine learning can be attributed to the incommensurable modes of reasonings between humans and certain ML models. As evident in the examples of GPT-3 and Google Brain, the more powerful ML models become, the greater the gap between these two modes of reasoning. Hence, the prospect of XAI and post-hoc explanations appears to be dim. The opacity that stems from the cognitive mismatch in machine learning, which belongs to Burrell's third form of opacity, seems very difficult to penetrate. But paradoxically, as I will argue, this opacity of cognitive mismatch in machine learning is a characteristic that actually prepares a favorable condition for ensuring a transparent production process when policies for enforcing third-party audits are put in place. A transparent production process would in turn undermine Burrell's first form of opacity, which is the opacity as intentional corporate or institutional concealment with the possibility of knowing deception. Undermining the opacity as institutional concealment allows stakeholders to gain access to previously concealed information or biases, and this granting of access helps restore the feedback channel in the technical politics or machine politics of machine learning.[6]

To proceed with this argument, we first need to recognize the typical segregation between the software process for training an ML model and the software process that embeds the trained ML model in production. In typical industrial practices today, the two processes would run on different hardware platforms because they demand different magnitudes of computational power.[7] The training of a complex ML model demands the processing power of a large server farm or on cloud computing services.[8] Once the model is trained, it can then be deployed to a computing device with much less processing power, such as a smartphone. The set-up is similar to online recommender systems that continually incorporate the incoming flow of user feedbacks to update their models. For these online platforms, the re-training of the models cannot be conducted in real time because there is not enough data in a short span of time for the re-training to be meaningful. So the re-training is typically performed periodically, perhaps once a day or once a week, on server farms or cloud computing services that are physically isolated from the web application servers. Once ready, the re-trained matrices of parameters would be data-transferred to the web application servers running in production. Because training an ML model is a process decoupled from running the model, which is intricately tied to other software systems, it is possible to cleanly separate the practice of training ML models as the target for policies and accountabilities.

As discussed earlier, the reasoning of powerful and complex ML models is typically beyond human comprehension. These models operate in a black box to producers and consumers alike. Correspondingly, the asymmetry of knowledge between producers and consumers, typically found in traditional software development, would also be vastly curtailed. In other words, if we compare the transparency of producers' inclination, the complexity of a large-scale software system can provide a better concealment of producers' inclination than the relatively simpler process of training and building ML models. Now, software code is still required to set up and configure the model, to cleanse and import input data, and to output performance metrics. But in comparison to programming-by-hand, it is the training of ML models that automatically constructs the complex program for recognizing patterns in data. Nowadays, most ML training algorithms are available as built-in functions from off-the-shelf software libraries.[9] Since these functions encapsulate most of the coding complexities, the code for configuring the inputs and outputs of ML models and that for cleansing and importing data are relatively simple to write. The sources of complexities and uniqueness have therefore been shifted from the production of software code to the raw data that train a model. This has a couple of implications for the enforcement of third-party audits. First, because the complexity of an ML model neither resides in code nor in patentable algorithms, it would be difficult for companies to justify the claim that exposing their process of training ML models to third-party auditors may expose their proprietary knowledge or trade secrets. Second, the main targets for analyses are the overall design, the process of data collection, data cleansing, and the

[6] The two terms machine politics and technical politics have similar meanings. The former is taken from "Hard Choices in Artificial Intelligence" (Dobbe et al. 2021) and the latter from Andrew Feenberg's works (e.g., see *Technosystem* (2017)). These works share the view that every technological system is inherently political, and they advocate collective agency and political deliberation during the technical design phase.

[7] For instance, running a neural network model requires just one round of forward propagation, whereas training a neural network model requires thousands of iterations of forward and backward propagations.

[8] E.g. Amazon Web Service (AWS) supports deep learning on their cloud services (https://aws.amazon.com/deep-learning/).

[9] Some companies may indeed develop or customize their code for the algorithms for training machine models. But it is nonetheless possible for regulatory policies to request the separation of this code into a software library that will not be subjected to third-party auditing.

data itself, as well as the model's predictions. In comparison, the main targets for auditing in a large-scale software system are primarily the codebase and system tests. Hence, auditing the process of data collection and data cleansing is more feasible than auditing the codebase of a large-scale software system for the simple reason that the amount of code in the former pales in comparison to the latter. It is true that a large software codebase can be scrutinized by many other programmers, as is the case with open-source software when the number of eyeballs can match the complexity of the software. Nevertheless, this scheme only works well in popular open-source software packages, but not in scenarios where the number of eyeballs does not match the complexity of the software, which is likely the case with a regulated third-party audit.[10]

To sum it up, because the training of an individual ML model is primarily driven by data rather than by a unique collection of programming statements, it is more difficult for corporations to justify the exemption from third-party audits with the reason of trade secret protection. Furthermore, because the bulk of the complexity now resides in the unfathomable trained parameters of an ML model, the software code for training the model is relatively simpler than the code in large-scaled software systems. With the relative simplicity in coding, it does not take thousands of eyeballs to audit the code. A third-party auditor with limited man-powers can analyze the model training process and bring new perspectives, such as that of social critiques, into the production process.[11] It is therefore much more feasible to stipulate policies for enforcing third-party auditing, in a similar fashion to financial audits, than regulations that require companies to open-source their entire codebase. Third-party auditing would render a more transparent production process in part due to the typical opacity of ML models in most industrial practices. Improvement in process transparency is the prerequisite for any kind of technical politics or machine politics that attempts to restore the feedback channel for social critiques. In "Hard Choices in Artificial Intelligence" (2021), Dobbe et al. propose an iterative framework for machine politics based on cybernetic feedback, arguing that designers ought to create channels for marginalized stakeholders to participate in system design and to actively determine the system's specification. But if the system conceals formal biases[12] that stakeholders are unaware of, then even the provision of feedback channels can easily become a means for co-opting uninformed stakeholders. Hence, machine politics in the context of opaque machine learning must go hand-in-hand with an improvement in process transparency to establish a more democratic relationship between the producers and the users of a technical system.

## 7 Democratic implications of opaque machine learning

Realizing the democratic potentials for opaque machine learning has important implications for researchers and policymakers on deciding what problems to tackle. Instead of blindly chasing after the translation of machine learning's reasoning into human-scale reasoning, I contend that they ought to devote their attention on how to gradually transform the current state of affairs, in which personal data traces are owned and controlled by commercial firms, into a more democratic milieu where individuals may have controls over their own digital footprints as well as the apparatus for analyzing these footprints. The first step toward this vision is to have proper policies and regulations for third-party audit in place, as third-party audits are helpful at curtailing the institutional power of behavioral manipulation through data profiling and analyses.

Encouraging progress has been made in recent years on regulations for auditing the development of AI systems. There has been much effort devoted to schemes for regulating and auditing AI, particularly in Europe (see Araujo et al. 2020; Berghoff et al. 2021; Brown et al. 2021; Supreme Audit Institutions 2020). Yet, without incentives or regulations, few private corporations would be willing to allow third-party access to their operations, infrastructure, and data farms under their proprietary domain. It is difficult to imagine how the FAANG (Facebook, Apple, Amazon, Netflix, and Google) will endorse any initiative for auditing AI systems, given their reliance on big data to drive their profit growth as Shoshana Zuboff describes in *The Age of Surveillance Capitalism* (2019). But a different picture is presented in a US National Security Commission (NSC) report published in March 2021. This "Final Report on AI in Defense

[10] It is true though, that auditors may also come up with their own automated tools, embedded with ML models, for scanning anomalies or biases in either data or software code. It is conceivable that auditors can design and train their own ML models for detecting software code with fraudulent motives, similar to those models designed to detect fraudulent behaviour in online transactions. So if the legal issue of proprietary trade secret is resolved and policies for regulating third-party audit of software code are set in place, it is conceivable that these tools for auditing software may become available, making it feasible to conduct third-party audits of a large software codebase.

[11] There are works in the academia that exemplify how critiques become feasible when the design process of machine learning is transparent. One such example is Wendy Chun's critique of the paper "Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images" (Wang and Kosinski 2018) in Chapter 4 of her *Discriminating Data* (Chun 2021).

[12] I am using 'formal bias' as defined in *Transforming Technology* (Feenberg 2002, pp. 80–82).

and Intelligence" (2021), a 756-page report drafted by commission members that include mostly senior executives from the high-tech industry, explains the impact of AI on national defense and intelligence. The report aims at presenting "a democratic model of AI use for national security" (2021, p. 11). It posits that a democratic model of AI would only work if the government can earn the public trust in AI tools, which "will hinge on justified assurance that government use of AI will respect privacy, civil liberties, and civil rights" (2021, p. 11). Over the past decade, big data companies claim ownership over personal profiles that were built by mining digital footprints over the years, and the profiles are inaccessible for the individuals themselves.[13] But with increasing public awareness of the authoritarian tendency in digital surveillance, funding allocation for AI research may shrink due to the growing apprehension of the technology. Hence there are now incentives for high-tech firms to endorse policies that can help them regain public trust, as evident in this NSC report on AI. This report, which represents primarily the voice of the high-tech industry, makes the following recommendation:

> Establish policies that allow individuals to raise concerns about irresponsible AI development and institute comprehensive oversight and enforcement practices. These should include auditing and reporting requirements, a review mechanism for the most sensitive or high-risk AI systems, and appeals and grievance processes for those affected by the actions of AI systems (US National Security Commission 2021, p. 138).

It is apparent from the report that prominent representatives of the high-tech industry are endorsing the legalization of third-party audit on AI systems. This endorsement can be attributed to the growing public distrust of AI systems, which may bring negative consequences, such as the shrinking allocation of research grants, to their respective firms.

This recent progress in public perception and policy regulation across the two major continents are encouraging. But as mentioned earlier in this paper, there is a lingering and misguided notion that ML models need to be made completely transparent to earn people's trust in certain contexts. Yet, as Emily Sullivan (2020) explains, ML models should only be deployed in contexts where there are scientific and empirical evidences for relating the model to the target phenomenon. It follows that the problem of trust surfaces only when such evidences are lacking. Pushing this argument further, I contend that the problem of trust is commonly found in situations where ML models are inappropriately deployed. Seeking transparency in ML models to resolve the problem of trust is a misguided endeavor that misses the scientific and technical ground for ML models to work properly. The attention of the public and the research community should rather focus on specifying the appropriate context for deploying ML algorithms, and on creating means for democratizing the technical apparatuses for pattern recognition. Restoring the feedback channel for social critiques to reach and influence the design and production of ML models, as Stiegler suggests in his philosophy, can bring about a regulative milieu[14] that filter out the inappropriate contexts for deploying ML algorithms. Third-party auditing can be such a channel, with social critiques diffusing into audit criteria. For instance, following Sullivan's argument, third-party auditing can gauge for every unique situation whether there are sufficient scientific and empirical evidences for relating the ML model to the target phenomenon. The stipulation of proper policies and regulations for enforcing third-party audit would represent the first step in the democratization of technical apparatuses.

## 8 Conclusion

In an algorithmic state of exception as per McQuillan, algorithmic actions have the force of the law even though they are not of the law. The state-of-exception metaphor raises critical awareness of the trend of increasing dependence on opaque algorithmic decision-making. The philosophical critique of automatic societies by Stiegler and that of algorithmic governmentality by Rouvroy and Berns express similar concerns over the opacity of digital algorithm, but unlike McQuillan, their Simondonian philosophy also outlines the possibility of emancipation toward a more democratic digital milieu. In this paper, I examine this possibility by drawing from Burrell's distinction between three forms of opacity: the opacity of institutional concealment, the opacity of technical illiteracy, and the opacity of cognitive mismatch. I argue, the critique by McQuillan makes the implicit assumption that the opacity of cognitive mismatch necessarily exacerbates the opacity of institutional concealment, but this assumption is contingent and can be subverted. If proper policies for enforcing third-party audits are put in place, the opacity of cognitive mismatch in machine learning can provide us with a technological environment that is more favorable to the establishment of process transparency.

[13] According to Zuboff (2019, p. 328), "[i]n this future we are exiles from our own behavior, denied access to or control over knowledge derived from our experience. Knowledge, authority, and power rest with surveillance capital, for which we are merely 'human natural resources'.".

[14] Note that Simondon also uses the term "regulative external milieu" to propose the proper relation between the social and cultural milieu and technology development (see 2016, pp. 49, 129).

To support this argument, I explain that ML models can be opaque to human understanding when their high-dimensional numerical-based "reasoning" is so complex and powerful that they become incommensurable with human-scale reasoning. The more powerful they become, the bigger the gap between the two modes of reasoning. Since the burden of the complexity has shifted from programming by hand to the trained parameters in ML models, the process of training ML models is relatively simpler than the traditional process of software development. Furthermore, the software for training models is typically isolated from the software that embeds the model because they typically run on different hardware platforms. It is therefore feasible to target specifically the training process of ML models with policies that enforce third-party audits.

This paper aims to raise awareness about the democratic implication of this process transparency associated with the opacity of ML models. This opacity has the unintended consequence of re-balancing the asymmetry of knowledge between producers and consumers, as both parties have no means to intervene with the inner workings of ML models. In that regard, it is not worth pursuing the rather infeasible goal of finding human-interpretable causes behind the decision-making process of an opaque ML model.[15] Rather, the improved symmetry lays the foundation for fostering a more democratic sociotechnical milieu. In Stiegler's philosophy, restoring the feedback channel for social critiques to intervene the production process of ML models can re-establish a regulative milieu for the proper deployment of machine learning. Establishing policies and regulations for third-party audit represents the first step toward this democratic vision. As Stiegler suggests, we can further imagine how individuals may be allowed to gain access to their own data traces, digital profiles, and the apparatuses for analyzing and finding patterns in their own personal data traces. Such a democratized sociotechnical assemblage can serve individuals' own reflections on self-understanding or self-improvement.

**Funding** The author has no financial or proprietary interests in any material discussed in this article.

**Data availability** The manuscript has no associated data that need to be made available.

[15] Lee et al. (2021, p. 12) discusses the limitations of some post-hoc explanation techniques in XAI. E.g., Local Interpretable Model-agnostic Explanations (LIME) "has been shown not to be robust: given two very similar inputs that result in very similar outputs from the model, LIME is not guaranteed to produce similar explanations." Also, as Watson (2021, p. 10) puts it, it is questionable whether interpretable machine learning "really settle matters, or merely push the problem one rung up the latter".

## References

Agamben G (2005) State of exception. University of Chicago Press, Chicago

Agamben G (2020) Giorgio Agamben, "The state of exception provoked by an unmotivated emergency". In: *positions politics*. https://positionspolitics.org/giorgio-agamben-the-state-of-exception-provoked-by-an-unmotivated-emergency/. Accessed 17 Aug 2021.

Ananny M, Crawford K (2018) Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. New Media Soc 20(3):973–989

Araujo T, Helberger N, Kruikemeier S et al (2020) In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI Soc 35(3):611–623. https://doi.org/10.1007/s00146-019-00931-w

Banko M and Brill E (2001) Scaling to very very large corpora for natural language disambiguation. In: Proceedings of the 39th annual meeting of the Association for Computational Linguistics, 2001, pp. 26–33

Berghoff C, Biggio B, Brummel E et al. (2021) Whitepaper: towards auditable AI systems, pp. 32

Boyd D, Crawford K (2011) Six provocations for big data. In: A decade in internet time: symposium on the dynamics of the internet and society, 2011

Brill J (2015) Scalable approaches to transparency and accountability in decisionmaking algorithms: remarks at the NYU conference on algorithms and accountability. Federal Trade Commission 28

Brooks FP (1975) The mythical man-month: essays on software engineering. Addison-Wesley Publisher Co, Reading

Brown S, Davidovic J, Hasan A (2021) The algorithm audit: scoring the algorithms that score us. Big Data Soc 8(1):2053951720983865. https://doi.org/10.1177/2053951720983865

Burrell J (2016) How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc 3(1):2053951715622512

Carabantes M (2020) Black-box artificial intelligence: an epistemological and critical analysis. AI Soc 35(2):309–317

Chan L (2021) Explainable AI as epistemic representation. In: Overcoming opacity in machine learning, pp. 7–8

Chun WHK (2021) Discriminating data: correlation, neighborhoods, and the new politics of recognition. The MIT Press, Cambridge

Creel KA (2020) Transparency in complex computational systems. Philos Sci 87(4):568–589

Crogan P (2019) Bernard Stiegler on Algorithmic Governmentality: A New Regimen of Truth? New Form 98:48–67. https://doi.org/10.3898/NEWF:98.04.2019

Datta A, Tschantz MC, Datta A (2015) Automated experiments on Ad privacy settings. Proc Priv Enhancing Technol 2015(1):92–112

Diakopoulos N (2016) Accountability in algorithmic decision making. Commun ACM 59(2):56–62

Dobbe R, Krendl Gilbert T, Mintz Y (2021) Hard choices in artificial intelligence. Artif Intell 300:103555. https://doi.org/10.1016/j.artint.2021.103555

Fainman AA (2019) The problem with Opaque AI. Thinker 82(4):44–55

Feenberg A (2002) Transforming technology: a critical theory revisited. Oxford University Press, New York

Feenberg A (2017) Technosystem: the social life of reason. Harvard University Press, Cambridge

Heath N (2013) Six open source security myths debunked—and eight real challenges to consider. https://www.zdnet.com/article/six-open-source-security-myths-debunked-and-eight-real-challenges-to-consider/. Accessed 29 Apr 2022

Huby G, Harries J (2021) Bloody paperwork: algorithmic governance and control in UK integrated health and social care settings. J Extreme Anthropol 5(1):1–28. https://doi.org/10.5617/jea.8285
Jarrahi MH, Newlands G, Lee MK et al (2021) Algorithmic management in a work context. Big Data Soc 8(2):20539517211020332
Lee K-F, Chen Q (2021) AI 2041, 1st edn. Currency, New York
Lee E, Taylor H, Hiley L et al (2021) Technical barriers to the adoption of post-hoc explanation methods for black box AI models. In: Overcoming opacity in machine learning, pp. 12–13
Levy E (2000) Wide open source. SecurityFocus. com. Electronic document, p. 19. http://www.securityfocus.com/news
Longoni C, Bonezzi A, Morewedge CK (2019) Resistance to medical artificial intelligence. J Consumer Res 46(4):629–650
Malik MM (2020) A hierarchy of limitations in machine learning. arXiv preprint arXiv:2002.05193
McKinney SM, Sieniek M, Godbole V et al (2020) International evaluation of an AI system for breast cancer screening. Nature 577(7788):89–94
McQuillan D (2015) Algorithmic states of exception. Eur J Cult Stud 18(4–5):564–576
McQuillan D (2016) Algorithmic paranoia and the convivial alternative. Big Data Soc 3(2):2053951716671340
Minsky M (1967) Why programming is a good medium for expressing poorly understood and sloppily formulated ideas. In: Design and planning II-computers in design and communication. New York, Hastings House, pp. 120–125
Mittelstadt BD, Allo P, Taddeo M et al (2016) The ethics of algorithms: mapping the debate. Big Data Soc 3(2):2053951716679679
Müller VC (2021) Deep opacity undermines data protection and explainable artificial intelligence. In: Overcoming opacity in machine learning, pp. 18–21
Ozment A, Schechter SE (2006) Milk or wine: does software security improve with age? USENIX Secur Symp 2006:10–5555
Pasquale F (2015) The black box society. Harvard University Press
Pūraitė A, Zuzevičiūtė V, Bereikienė D et al. (2020) Algorithmic governance in public sector: is digitization a key to effective management. https://repository.mruni.eu/handle/007/17025. Accessed 17 Aug 2021.
Raymond ES (2001) The Cathedral and the Bazaar: musings on Linux and open source by an accidental revolutionary, Rev. O'Reilly, Cambridge
Rouvroy A, Berns T (2013a) Algorithmic governmentality and prospects of emancipation. Reseaux 177(1):163–196
Rouvroy A, Berns T (2013b) Gouvernementalité algorithmique et perspectives d'émancipation. Reseaux 177(1):163–196
Sandvig C, Hamilton K, Karahalios K et al (2014) Auditing algorithms: research methods for detecting discrimination on internet platforms. Data Discrimination: Converting Critical Concerns into Productive Inquiry 22:4349–4357
Schryen G (2011) Is open source security a myth? Commun ACM 54(5):130–140. https://doi.org/10.1145/1941487.1941516
Seaver N (2017) Algorithms as culture: Some tactics for the ethnography of algorithmic systems. Big Data Soc 4(2):2053951717738104
Silver D, Hubert T, Schrittwieser J et al (2018) A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science 362(6419):1140–1144
Simondon G (2016) On the mode of existence of technical objects. Univocal Publisher, Minneapolis
Smith GJ (2020) The politics of algorithmic governance in the black box city. Big Data Soc 7(2):2053951720933989. https://doi.org/10.1177/2053951720933989
Stiegler B (2016) Automatic society: the future of work. Polity Press, Cambridge
Stokes JM, Yang K, Swanson K et al (2020) A deep learning approach to antibiotic discovery. Cell 180(4):688–702
Sullivan E (2020) Understanding from machine learning models. In: Sps S (ed) The British journal for the philosophy of science. The University of Chicago Press, Chicago
Supreme Audit Institutions (2020) Auditing Machine Learning Algorithms. https://www.auditingalgorithms.net/index.html. Accessed 16 August 2021
US National Security Commission (2021) NSCAI Final Report. https://www.nscai.gov/. Accessed 20 May 2021.
Wang Y, Kosinski M (2018) Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. J Personal Soc Psychol 114(2):246
Watson DS (2021) No explanation without inference. In: Overcoming opacity in machine learning, pp. 9–11
Weizenbaum J (1976) Computer power and human reason: from judgment to calculation. Freeman, San Francisco
Zednik C, Boelsen H (2021) Preface: overcoming opacity in machine learning. In: Overcoming opacity in machine learning, pp. 1–2
Zou S (2021) Disenchanting trust: instrumental reason, algorithmic governance, and China's emerging social credit system. Media Commun 9(2):140–149. https://doi.org/10.17645/mac.v9i2.3806
Zuboff S (2019) The age of surveillance capitalism: the fight for a human future at the new frontier of power, 1st edn. PublicAffairs, New York