AI & SOCIETY
https://doi.org/10.1007/s00146-020-01130-8

OPEN FORUM

Making the black box society transparent

Daniel Innerarity[1,2,3,4]

Received: 28 April 2020 / Accepted: 12 November 2020

Abstract
The growing presence of smart devices in our lives turns all of society into something largely unknown to us. The strategy of demanding transparency stems from the desire to reduce the ignorance to which this automated society seems to condemn us. An evaluation of this strategy first requires that we distinguish the different types of non-transparency. Once we reveal the limits of the transparency needed to confront these devices, the article examines the alternative strategy of explainable artificial intelligence and concludes with the idea that these types of complex realities exceed individual capacities and are only comprehensible in a collective fashion.

Keywords Artificial intelligence · Black box · Algorithms · Transparency

1 Introduction

We live in a society that is full of black boxes, mechanisms, systems, algorithms, robots, codes, automatisms and the mechanisms we use or that affect us even though we do not understand how they work. Niklas Luhmann spoke in this regard about "the symphony of non-transparency" that characterizes contemporary society (2017, 96). We could call ours a "black box society" (Pasquale 2015). The first question to be answered has to do with the continuity or breakdown of these systems regarding classic forms of non-transparency. Are current intelligent systems more opaque than the government procedures and behaviour-control mechanisms that the state has traditionally employed? Can we be certain that opacity increases with new systems or might it be that there is a continuity of opportunistic opacity on the part of those who always exercise power, regardless of the technologies they employ to keep their decisions from being transparent? Answering these questions requires previous clarification about the forms of non-transparency associated with those mechanisms.

If it is true that the intervention of intelligent systems has increased and their influence is more decisive in daily life, there is also a greater need to balance the resulting cognitive asymmetries. The demand to fight opacity seems like an appropriate response to this situation, and it is not surprising that many scholars are demanding greater transparency (Balkin 2016; Benjamin 2013; Citron and Pasquale 2014; Cohen 2016; Mehra 2015). Others are sceptical of the demand for transparency (Kroll et al. 2016; Burrell 2016) and note that it is not a panacea to resolve all the ethical questions that come with new technologies (Mittelstadt et al. 2016; Neyland 2016; Crawford 2016).

It is clear, in any case, that we should build a whole new architecture of justification and control where automatic decisions can be examined and submitted to critical review. This is the direction we see in everything from the idea of "reverse engineering", in other words, the process of "extracting knowledge or design blueprints from anything man-made" (Eilam 2005) to the different initiatives for auditing algorithms (Sandvig 2014).

To assess the scope of these and other strategies to promote transparency, we must also carry out a taxonomy of the kinds of non-transparency that exist, which could be summarized as (1) opacity that is intentional and deliberately produced, (2) opacity that is technical and objective, which stems from the cognitive asymmetry of technical complexity

✉ Daniel Innerarity
dinner@ikerbasque.org

[1] Department of Political Philosophy, University of the Basque Country, Leioa, Spain

[2] Ikerbasque Foundation of Science, Bilbao, Spain

[3] European University Institute of Florence, Fiesole, Italy

[4] Instituto de Gobernanza Democrática, Gipuzkoa, Spain

Published online: 05 February 2021

and (3) an emerging opacity, specific to machine learning, to its unpredictability and unintentionality.

The question of the extent to which transparency is possible presents questions that are too large to be answered with a simple division of that which can be known and that which cannot. The desire for intelligibility that is behind calls for greater transparency must confront the question of whether it makes sense to view algorithms as "knowable known unknows", in other words, as something that can be known (Roberts 2012). When Pasquale affirms that "you can't form a trusting relationship with a black box" (2015, 83), he is implying that this cognitive asymmetry could be eliminated. In more recent literature, on the other hand, there is an increasing insistence that it is on principle impossible to look into the "black box" of artificial intelligence. There are those who have spoken of a "loophole" that no supervision of programmes or ethical advisors can fill: "Any system simple enough to be understandable will not be complicated enough to behave intelligently, while any system complicated enough to behave intelligently will be too complicated to understand" (Dyson 2019, 39). The more capable the intelligent systems, the harder it is to understand their decisions. We should analyse every one of these forms of non-transparency and present some form of explainability before deciding that intelligent systems are incomprehensible and uncontrollable.

2 Intentional non-transparency

The first form of non-transparency is one that is produced intentionally; it is due to a deliberate will to conceal, which is not necessarily censurable. In this case, ignorance is not a problem of technology but of deliberately produced opacity. That opacity may be due to the protection of data, to property rights or questions of security or the common good. We could speak, then, of a strategy of black boxing: the intentional use of ignorance to the extent that it can be more advantageous than cultivating knowledge about the ends being sought (McGoey 2012).

Any aspiration to reduce the opacity of the environment in which we move must differentiate between types of non-transparency. We should identify when we are facing a black swan or a black box, whether it is an unpredictable event or a mechanism that is meant to conceal things. Humans have a rudimentary impulse that leads us to distrust that which is hidden or secret, to try to uncover it and to think that knowledge should provide us with greater control over our environment. But when we are facing this type of non-transparency, that attitude only makes sense if the right to know is more important that the goods protected by the secret. And it is always necessary to keep in mind the nature of the non-transparency we are confronting. The identity of the algorithms is in parts settled and non-settled; the critical analyst's task is to study when the will to know is relevant and to question the separation between the social and the technical. It is true that the exact configuration of the algorithm cannot be easily devised, but that does not free us from the need to interrogate, especially since the allusion to ignorance has become a convenient justification for the platforms, when they suggest that their algorithms operate without human intervention and that they are not designed but discovered.

In any case, we should be able to carefully handle our expectations of disclosure, because sometimes opacity is not intentional and, when it is, it is not always clear where complexity ends and where intentionality begins. When we affirm that algorithms discriminate, we are talking more about a matter of distributed agency than of individual intentionality. Terms such as prejudice, subjectivity, manipulation or neutrality suggest that everything is resolved by discovering who it is who is acting, as if human beings were using algorithms to hide who is really taking decisions. Of course, algorithms are made and maintained by human beings. Everything would be easier if someone specific assumed responsibility, but from a relational perspective, it would be a mistake to determine the origin of an action as if it could be referred to a single source. "Agency is not aligned with human intentionality or subjectivity" (Barad 2003, 826). Not all situations refer back to one agent. As Latour affirms, "to use the word 'actor' means that it is never clear who or what is acting when we act since an actor on stage is never alone in acting" (2005, 46). It is not a question of pointing out the designers or the users. A relational perspective "disavows any essentialist or isolated explanation of either human or nonhuman agency" (Schubert 2012, 126). The attribution of responsibility is not impossible in an environment of distributed agency, but neither is it easy since determining who is acting—humans or technologies—depends on the particular pattern under consideration.

3 Objective non-transparency

The existence of inexplicable spheres in ourselves, our objects and in society is not a recent technological development but is a part of our human condition. We can affirm that our very bodies are black boxes for us and that many of the things we do are not brought about by a specific decision and are certainly not caused by something we can or should explain. Human beings do not really know how we do many of the things we know how to do. It is part of the nature of human intelligence that only a part of it is rationally explainable; to a large degree, it is instinctive, subconscious, implicit or inscrutable. There is a certain degree of persistent opacity in the sensorial and cognitive modalities

with which human beings interpret the world. In what is called System 1 thinking (Kahneman 2011), which is the most automatic and least reflexive thought, filled with all types of biases, unconscious preferences and heuristics, to use Kantian terminology, the self simply accompanies our representations in a tacit fashion.

Let us take as a starting point the observation that darkness is not a prerogative of the technology or the algorithms but a component of the human world. We do not fully understand the functioning of our cognitive apparatus, and many of our decisions do not obey a consciousness that can give a reasonable accounting of it. We know that the mind is often mistaken, that it becomes distracted and even tricks us. Automatisms are also part of the human condition, whether they are biological, cultural or social.

The growth of non-transparency has to do with technological progress. With the advance of the civilizing process, human beings have developed reciprocal comprehension by virtue of which they renounce the explanations that we demand from machines (Yudkowsky 2008). The technologies that are most present in many people's lives end up being so familiar that they disappear from view as such and end up being indistinguishable from life itself. There is a "technological unconscious" (Clough 2000) hidden in the quotidian familiarity. This was suggested by Mark Weiser, the father of "ubiquitous computing" or "calm technology": the success of a technology would be connected to its ability to become invisible when it turns into a constitutive and almost natural part of our life without being invasive or reclaiming the user's attention (Weiser and Brown 1998).

The other side of this familiarity is our lack of understanding of technology. Its technological complexity leads to ignorance on the part of users and cognitive asymmetries between them and the experts. The debate of transparency prefers to revolve around technological businesses and the use of technologies, but little about the properties of technology as such. The decision-making process of intelligent systems is non-transparent and opaque, to a large extent because of technical motives, not because of the express intentionality of its designers. "Analysis based upon mined data, premised on thousands of parameters, may be difficult to explain to humans. (…) Equally, the firm governing through such data analysis would find it difficult to adequately explain the 'real reason' for its automated response—even after making a good faith effort to do so" (Zarsky 2016, 121). When neither the user nor the person affected can know why a system has decided in this way and not in another, the controls can barely verify whether the decision was carried out correctly. The lack of transparency that is not intentional turns into a grave impediment for effective regulation.

However, as was previously pointed out, opacity and invisibility are not an epistemic anomaly, but they are part of daily life; they are not an exception but the norm for many things that seem or truly are hidden, implicit, they are not the object of express deliberation that function precisely because of that, unleashing our obligation to decide or allowing us to pay attention to other things. As Schutz foresaw in the 1940s, we would be using "the most complicated gadgets prepared by a very advanced technology without knowing how the contrivances work" (Schutz 1946, 463). In this regard, Ashby recommends we get used to living with "systems whose internal mechanisms are not fully open to inspection" (1999, 86) and suggests that, when we confront a black box, we not expect to know exactly what is inside but distinguish between the properties that can be discovered and the ones that cannot. Even in the case of apparently hidden or closed systems, there are many things that can be known.

Most of the technologies are designed in such a way that people do not need to know exactly how they function (Hardin 2003). Black box code reduces the cognitive load of programmers, allowing them to design new properties and functions without having to think about every little detail of how the system functions. "A black box contains that which no longer needs to be reconsidered" (Callon and Latour 1981, 285). Any strategy destined to strengthen transparency must keep in mind that the success of black boxes is based on obscuring the networks and pieces of which they are made. Blackboxing is a process by which all technical work makes its own success invisible. It reveals that reality is not something stable but an assemblage of many interrelated parts. "Black boxes never remain fully closed or properly fastened (…) but macro-actors can do as if they were closed and dark" (Callon and Latour 1981, 285; Latour 1999, 183).

With these considerations in mind, any attempt to increase transparency in a system should confront an unsettling hypothesis. To what extent is the will for transparency compatible not only with the benefits of automatization in general but also with the performances of the systems that are due precisely to these forms of opacity given their unspoken, implicit, irreflexive and un-themed nature? It is worth thinking about how much we would be limited if we could only make use of those mechanisms that we understand; we would reduce the benefits of artificial intelligence enormously (House of Lords 2019, 37).

4 Emerging non-transparency

The third type of opacity, the most complex and specific of the new smart devices, is the one that is not hidden (intentionally or because of its technological complexity, like the two previous ones) but one that emerges with development; it is unexpected, obeying the autonomy of its intelligent character. We would be talking about the black box of emerging things: mechanisms whose nature, to the extent

that they learn, is in continuous evolution; they are unstable, adaptative and discontinuous given their permanent reconfiguration, as is the case of the actualizations of continuous design. Non-transparency intensifies when the systems are governed by machine learning (Burrell 2016; Danaher 2016). This opacity can be very resistant in the face of strategies of transparency especially when the mechanisms of machine learning make deductive explanations impossible. A phenomenon that is continuously changing becomes for that very reason incomprehensible.

Humanity has constructed machines that were only understood by their creators, but we had never constructed machines that would operate in a way their creators did not understand. Artificial intelligence seems to imply this type of historical novelty. Machine learning excludes any certainty about the result of its operations; if it is true learning that the machine realizes on its own, we would not be able to know beforehand what it is going to know in the future. The decisional rule emerges in a way no human can explain (Kroll et al. 2016). The fact that they are autonomous systems does not mean that they will be free and rational, but that they have the ability to take decisions that cannot be predicted. That is why the demand for transparency can collide with an insuperable limit: it makes no sense to ask the programmers questions to understand the algorithms, as if the true nature of the algorithms were determined by the intentions of their designers. How can we understand a mechanism, its evolution and decisions, if not even its creators know exactly how it works?

This phenomenon is related to what we could call the paradox of software: there should be innovations and anomalies so that software can exist, but they should be eliminated so the software will be stable. Poor functioning is key to revealing the nature of the code, because "circumstances in which the software does not work, or does not work as expected, can tell us a lot about it" (Frabetti 2015, 144). The verification of an anomaly is when we find ourselves in a critical phase in which a decision should be made about whether it is a dysfunction that must be corrected or an anomaly that must be developed and integrated within the system. Coders know that, at a particular level of sophistication of any system, dysfunctionality is indistinguishable from a new functionality within the system.

The decisive question is how to make it possible for the technologies of deep learning to be more comprehensible for their creators and more accountable for the users. Elucidating the political ontology of algorithms demands "remembering that boundaries between humans and machines are not naturally given but constructed, in particular historical ways and with particular social and material consequences" (Suchman 2007, 1). It makes no sense to demand transparency and responsibility for non-human processes, of course, but it is important not to lose sight of the fact that algorithms "may inherit the prejudices of prior decision makers" and "reflect the widespread biases that persist in society at large" (Barocas and Selbst 2016, 671).

To identify the power of algorithms, we must be capable of understanding what type of operations they help produce and what class of subjects are possible in the algorithmic landscape. Algorithms do not determine the behaviour of people but configure the surroundings in which certain subjective positions are more available. In addition, in the age of machine learning, algorithms *need* us, depend on us, they do not develop without us. What follows from the logic of machine learning is that "what people do in anticipation of algorithms tells us a great deal about what algorithms do in return" (Gillespie 2017, 75). In terms of power and responsibility, we must look both at the machines and at ourselves.

5 Explainable artificial intelligence

Keeping in mind the difficulties presented by the strategy of transparency, the debate has turned in recent years toward another category: that of intelligibility or explainability and its ability to reduce the asymmetry of information, of providing fairness, reliability and trust. It would be a question of designing an "explainable AI", a good example of which would be the project DARPA that was initiated in 2017 (Russell et al. 2015; Datta et al. 2017, 71; Fong and Vedaldi 2017; European Commission's High-Level Expert Group on Artificial Intelligence 2018; Joint Research Center of the European Commission 2020). The explanation refers to the information and the logic employed to adopt the corresponding decisions. Public organizations and institutions should explain the processes and decisions of machine-learning algorithms in a way that is comprehensible for the humans that employ them or are affected by them.

Although "black box testing" cannot be carried out for every concrete decision, the general criteria that are being used should be revealed, thus preparing us to diagnose lack of quality or any implicit discriminations, so that this assessment can motivate interventions to correct the program. Every time we decide to develop an algorithmic system, certain choices are made. Algorithms are trained to navigate a massive data set by making use of certain pre-defined key concepts or variables, such as "creditworthiness" or "high-risk individual." The algorithm does not define these concepts itself; human beings—developers and data scientists—choose which concepts to appeal to, at least as an initial starting point. The rule of law presents the requirement to explain decisions, which seems hard to reconcile with a completely automated administration. And in concrete areas, such as criminal law, decisions cannot be adopted on the basis of statistical correlations, but by virtue of the strict reconstruction of causal relationships. We certainly cannot

apply to system decisions the criteria that are effective for human decisions, but we can situate intelligent systems into a deliberative space in which decisions and arguments are weighed. It is possible, for example, to program these systems in such a way that they report on the motives for their decisions. Explaining the functioning and decisions for a system is not everything; we can also demand responsibility, in other words, looking not only at causality but the values that have guided their behaviour.

European regulations on data protection introduce a right to explanation on the side of data subjects regarding "automated decision-making, including profiling" and "the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" (GDPR 2016, Articles 22.1 and 22.4), but this does not imply that the whole AI system is completely explainable. In addition, we would need to clarify what a "significant" level of explanation and transparency is.

The idea of opening the algorithms for public inspection presents diverse dilemmas and perhaps the metaphor of opening the black box is not the most appropriate for understanding the epistemology of algorithms, nor can their disclosure provide what its critic want to achieve (Bucher 2018). Automated processes limit asymmetrical advantages, but their transparency reconstructs them to the extent that they are more comprehensible for some people than for others. The public has more information, but so do interest groups (Zarsky 2016, 125). We must also expect the possible unintended consequence that transparency might help the most competent actors game the system without inexperienced users benefiting from it.

The "right to an explanation" does not have to be an autopsy of the systems. Instead, it works rather like a principle of self-control. Understood like this, this principle is compatible with the complexity of the system and reduces cognitive asymmetry a little between designers and those affected. In any case, explainability confronts a dilemma that is difficult to resolve. Where should these explanations focus more attention: on respecting the complexity of the system or the ability of the recipients?

Perhaps, it is more useful to pay attention to the concrete ways in which this explanation can be carried out: the who, when, what and why of disclosure, that which Pasquale has called "qualified transparency" (2015, 142). Human supervision against biases can be realized in different moments of a decision-making process: in the selection of data, in the configuration of algorithms or in the activity of the system. We could also intervene to balance out discriminatory consequences. There are those who recommend that given the non-intelligibility of decisions made via deep learning, it would be more feasible to focus on outcomes, and "only license critical AI systems that satisfy a set of standardised tests, irrespective of the mechanism used by the AI component" (House of Lords 2019, 94). It is easier to measure the effects of intelligent systems to the extent that they can evaluate, for example, the ways in which certain groups are discriminated against.

6 Comprehension as a collective matter

This supervision of intelligent systems surpasses the ability of average people; it is in principle within the reach of experts, but even the specialists are hard pressed to understand certain decisions. An ex ante forecast about the decision of a dynamic intelligent system is difficult to the extent to which not all possible interactions are known; neither does an ex post reconstruction make it easy to identify the factors that are responsible for certain results. The decision is instead a function of the probability of some variables examined on the basis of a huge dynamic quantity of data. Even when such functions could be identified, from the point of view of the human observer, there is still a "mismatch between mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of semantic interpretation" (Burrell 2016, 2). That means that the inspection of algorithms should generally be delegated to "some trusted auditor" (Pasquale 2015, 141).

It is not the case that transparency in the code increases intelligibility for the average citizen. It is not possible to offer a general description of the system and of the significant factors in every situation (Tene and Polonetsky 2013, 269; Barocas and Selbst 2016; Datta et al. 2017, 71). The programs whose decisions are based on enormous quantities of data are enormously complex. Individual human beings become overwhelmed when it comes to understanding in detail the decision-making process (Leetaru 2016). It is not unusual for the explanations to be harder to understand than the systems that they are meant to explain (Brauneis and Goodman 2018). Only experts are prepared to understand the logic of the codes and algorithms, so that any operation of making them more transparent has asymmetrical effects; it does not allow for universal accessibility.

If we keep these difficulties in mind and examine some previsions of the GDPR, such as Article 22, the "right to an explanation" seems unrealistic. This right recognizes the individual's ability to demand an explanation about how a fully automated decision that affects them was taken. This prevision is vague; it does not apply if the decision is based on explicit consent or if the process was semi-automatic (House of Lords 2019, 101). In addition, it is not possible for a neural network to explain how a situation was categorized. All of the attempts to explain the functioning of a neural network or to audit the decision-making system do nothing but circle the problem;

they make use of a second network that tries to describe the function from which the learning emerged (Dessalles 2019, 87).

But the fundamental problem when we talk about intelligibility is that the task of auditing algorithms or explaining automatic decisions must be conceived as a collective task, not as a mere individual right, a right that is, additionally, often hard to realize. The idea of informed consent stems more from the private right than from the governance of the common good, from the perspective of the consumer, from the protection of intimacy and non-interference, from a negative liberty (in the sense in which Isaiah Berlin formulated it), not from a relational, social and political perspective. We must go beyond the minimalist requirement of information and consent.

It is not enough to "privatize transparency" (Wishmeyer 2018, 54) and leave in the hands of the citizenry the control of intelligent systems—a control they can barely carry out—and thus renounce public regulation. The practices of transparency have no place in a social void (Beer 2017; Kemper and Kolkman 2018), nor are algorithms objects that are known through observation (Ziewitz 2017). Instead, all of it is connected to concrete practices that make sense out of it (Lowrie 2017). Transparency is a relational good (Felzmann et al. 2019). Individual subjects can only manage massive data streams to a limited extent. We would not be able to decide regarding data and possible decisions unless they are filtered down to a size we can handle. Users need to be supported by systems of accountability (O'Neill 2014). It is crucial to consider not only the disclosed information but the instruments and capacities that are needed to interpret it (Kemper and Kolman 2018). For that, we must understand transparency holistically. Instead of taking an independent user as a starting point, the practices of transparency only make sense in a social context, as signals of a willingness to render an accounting and generate confidence.

7 Concluding remarks

After having examined the various types of non-transparency that characterise the general automation of decision-making processes in our societies, this paper examines a more useful and a promising concept of explainability by placing it in the framework of not so much as an individual but as collective capacities to design a possible comprehensibility. The paper points to a very promising future path of research to think about what kind of capabilities and collective intelligence would be needed in order for us to continue thinking that automation is compatible with the ideals of autonomy and responsibility in a human-centred technological environment.

References

Ashby WR (1999) An introduction to cybernetics. Chapman & Hall, London

Balkin J (2016) Information Fiduciaries and the First Amendment. UC Davis Law Rev 49(4):1183–1234

Barad K (2003) Posthumanist performativity: toward an understanding of how matter comes to matter. Signs 28(3):801–831

BarocasAndrew SS (2016) Big data's disparate impact. Calif Law Rev 104:671–732

Beer D (2017) The social power of algorithms. Inf Commun Soc 20(1):1–13

Benjamin SM (2013) Algorithms and speech. Univ Pa Law Rev 161:1445–1493

BrauneisGoodman REP (2018) Algorithmic transparency for the smart city. Yale J Law Technol 20:103–176

Bucher T (2018) If… Then. algoritmic power and politics. Oxford University Press, Oxford

Burrell J (2016) How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc 1:1–12

Callon M, Latour B (1981) Unscrewing the big leviathan: how actors macro-structure reality and how sociologists help them to do so. In: Knorr-Cetina K, Cicourel A (eds) Advances in social theory and methodology: toward an integration of micro- and macro-sociologies. Routledge, London, pp 277–303

Citron D, Pasquale F (2014) The scored society: due process for automated predictions. Wash Law Rev 89(1):1–33

Clough P (2000) Unconscious thought in the age of teletechnology. University of Minessota Press, Minneapolis

Cohen J (2016) The regulatory state in the information age. Theor Inq Law 17(2):369–414

Crawford K (2016) Can an algorithm be agonistic? Sci Technol Hum Values 41(1):77–92

Danaher J (2016) The threat of algocracy: reality resistance and accommodation. Philos Technol 29(3):245–268

Datta A, Sen S, Zick Y (2017) Algorithmic transparency via quantitative input influence. In: Cerquitelli T, Quercia D, Pasquale F (eds) Transparent data min big small data. Springer, New York

Dessalles J-L (2019) Des intelligences TRÈS artificielles. Odile Jacob, Paris

Dyson G (2019) The third law. In: Brockman J (ed) Possible minds. 25 Ways of looking at AI. Penguin, New York, pp 33–40

Eilam E (2005) Reversing: secrets of reverse engineering. Wiley, Indianapolis

European Commission's High-Level Expert Group on Artificial Intelligence (2018), Draft ethics guidelines for trustworthy AI (2018). https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai. Access 21 Feb 2019

Felzmann H, Villaronga EF, Lutz C, Tamò-Larrieux A (2019) Transparency you can trust: transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data Soc 1:1–14

Fong R, Vedaldi A (2017) Interpretable explanations of black boxes by meaningful perturbation. Proceedings of the IEEE international conference on computer vision, 3429–3437

Frabetti F (2015) Software theory. A cultural and philosophical study. Rowman & Littlefield, London

General Data Protection Regulation (GDPR) (2016) https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

Gillespie T (2017) Algorithmically recognizable: santorum's google problem, and google's santorum problem. Inf Commun Soc 20(1):63–80

Hardin R (2003) If it rained knowledge. Philos Soc Sci 33(1):3–44

House of Lords (2019) Select committee on artificial intelligence, report of session 2017–19, HL Paper 100. AI in the UK: Ready, Willing and Able?

Joint Research Center of the European Commission (2020) https://ec.europa.eu/jrc/communities/en/node/1162/article/interpretability-ai-and-its-relation-fairness-transparency-reliability-and-trust. Accessed 31 Jan 2020

Kahneman D (2011) Thinking farrar fast and slow. Straus and Giroux, New York

Kemper J, Kolkman D (2018) Transparent to whom? No algorithmic accountability without a critical audience. Inf Commun Soc 1:1–16

Kroll J, Huey J, Barocas S, Felten E, Reidenberg J, Robinson D, Yu H (2016) Accountable algoritms. Univ Pa Law Review 165:633–706

Latour B (2005) Reassembling the social: an introduction to actor-network-theory. Oxford University Press, Oxford

Leetaru K (2016) In Machines We Trust: Algorithms Are Getting Too Complex To Understand. www.forbes.com/sites/kalevleetaru/2016/01/04/in-machines-we-trust-algorithms-aregetting-too-complex-to-understand/print. Acessed 4 Jan 2016

Lowrie I (2017) Algorithmic rationality: epistemology and efficiency in the data sciences. Big Data Soc 4(1):1–13

Luhmann N (2017) Die kontrolle von intransparenz. Suhrkamp, Berlin

McGoey L (2012) The logic of strategic ignorance. Br J Soc 63(3):533–576

Mehra S (2015) Antitrust and the robo-seller: competition in the time of algorithms. Minn Law Rev 100:1323–1375

Mittelstadt B, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc 1:1–21

Neyland D (2016) Bearing accountable witness to the ethical algorithmic system. Sci Technol Hum Values 4(1):50–76

O'Neill O (2014) Trust, trustworthiness and accountability. In: Morris N, Vines D (eds) Capital failure: rebuilding trust in financial services. Oxford University Press, Oxford, pp 172–189

Pasquale F (2015) The black box society: the secret algorithms that control money and information. Harvard University Press, Cambridge

Roberts J (2012) Organizational ignorance: towards a managerial perspective on the unknown. Manag Learn 44(3):10–26

Russell S, Dewey D, Tegmark M (2015) Research priorities for robust and beneficial artificial intelligence. J Rec AI Commun 36(4):105–114

Sandvig C, Hamilton K, Karahalios K, Langbort C (2014) An algorithm audit. In: Seeta PG (ed) Data and discrimination: collected essays. New America Foundation, Washington, pp 6–10

Schubert C (2012) Distributed sleeping and breathing: on the agency of means in medical work. In: Passoth J-H, Peuker B, Schillmeier M (eds) Agency without actors: new approaches to collective action. Routledge, Abingdon, pp 113–129

Schutz A (1946) The well-informed citizen: an essay on the social distribution of knowledge. Soc Res 13(4):463–478

Suchman L (2007) Human-machine reconfigurations: plans and situated actions. Cambridge University Press, Cambridge

Tene O, Polonetsky J (2013) Big data for all: privacy and user control in the age of analytics. Northwest J Technol Intell Prop 11(5):239–273

Weiser M, Brown JS (1998) The coming age of calm technology. In: Denning P, Metcalfe R (eds) Beyond calculation: the next fifty years of computing. Copernicus, New York, pp 75–85

Wishmeyer T (2018) Regulierung intelligenter systeme. Archiv des öffentlichen Rechts 143:1–66

Yudkowsky E (2008) Artificial intelligence as a positive and negative factor in global risk. In: Bostrom N, Ćirković M (eds) Global catastrophic risk. Oxford University Press, Oxford, pp 308–345

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

Ziewitz M (2017) A not quite random walk: experimenting with the ethnomethods of the algorithm. Big Data Soc 4(2):1–13