Philosophy & Technology (2021) 34:1607–1622
https://doi.org/10.1007/s13347-021-00477-0

RESEARCH ARTICLE

Transparency and the Black Box Problem: Why We Do Not Trust AI

Warren J. von Eschenbach[1]

Received: 20 January 2021 / Accepted: 26 August 2021 / Published online: 1 September 2021

Abstract

With automation of routine decisions coupled with more intricate and complex information architecture operating this automation, concerns are increasing about the trustworthiness of these systems. These concerns are exacerbated by a class of artificial intelligence (AI) that uses deep learning (DL), an algorithmic system of deep neural networks, which on the whole remain opaque or hidden from human comprehension. This situation is commonly referred to as the black box problem in AI. Without understanding how AI reaches its conclusions, it is an open question to what extent we can trust these systems. The question of trust becomes more urgent as we delegate more and more decision-making to and increasingly rely on AI to safeguard significant human goods, such as security, healthcare, and safety. Models that "open the black box" by making the non-linear and complex decision process understandable by human observers are promising solutions to the black box problem in AI but are limited, at least in their current state, in their ability to make these processes less opaque to most observers. A philosophical analysis of trust will show why transparency is a necessary condition for trust and eventually for judging AI to be trustworthy. A more fruitful route for establishing trust in AI is to acknowledge that AI is situated within a socio-technical system that mediates trust, and by increasing the trustworthiness of these systems, we thereby increase trust in AI.

Keywords Artificial intelligence · Deep learning · Black box · Transparency · Trust · Trustworthiness

e-Trust, or trust in the context of digital environments or between humans and artificial agents, has gained considerable attention over the last few decades (Taddeo & Floridi, 2011). Questions about essential features of e-trust, its relation to standard notions of trust, and the extent to which trust in technology, whether considered as autonomous agents or artifacts, continue to be salient, especially as technology

✉ Warren J. von Eschenbach
wvonesch@nd.edu

[1] University of Notre Dame, Notre Dame, IN, USA

becomes more ubiquitous in everyday life and gains considerable power and sophistication.

With increasing automation of routine decisions coupled with more intricate and complex information architecture operating this automation, concerns are increasing about the trustworthiness of these systems. Indeed, some scholars even have argued that the growing dependence on these systems gives rise to "the threat of algocracy—a situation in which algorithm-based systems structure and constrain the opportunities for human participation in, and comprehension of, public decision-making" (Danaher, 2016, 246). These concerns are exacerbated by a class of artificial intelligence (AI) that uses deep learning (DL), an algorithmic system of deep neural networks, which on the whole remain opaque or hidden from human comprehension.[1]

This situation is commonly referred to as the black box problem in AI. Observers can witness the inputs and outputs of these complex and non-linear processes but not the inner workings. How AI reaches its conclusion is opaque or hidden from view. Without understanding how AI reaches its conclusions, it is an open question to what extent we can trust these systems. The question of trust becomes more urgent as we delegate more and more decision-making to and increasingly rely on AI to safeguard significant human goods, such as security, healthcare, and safety.

Technical solutions are being pursued to this problem by developing models that "open the black box" by making the non-linear and complex decision process understandable by human observers. This class of models, referred to as explainable artificial intelligence or XAI, are promising solutions to the black box problem in AI but are limited, at least in their current state, in their ability to make these processes less opaque to most observers.

Questions about why trust is more valuable than reliability also are important to consider in the context of AI and other technologies. No doubt that the appeal of AI is that it promises to be more reliable than humans in carrying out complex operations. Already we have witnessed that by using DL techniques, scientists can develop computer programs, such as AlphaGo, that are superior to human counterparts. But from an ethical perspective, there are important distinctions between trust and reliability we should heed when thinking about implementation of AI in cases where something of moral significance is at stake.

A philosophical analysis of trust will show why transparency is a necessary condition for trust and eventually for judging AI to be trustworthy. In a moral context, how and why something is carried out is as important as reliability. Given this account, the inscrutability of AI explains why many do not trust these technologies. As noted, XAI offers a promising avenue for making AI more transparent and therefore trustworthy, but it has limitations in terms of its applicability. A more fruitful route for establishing trust in AI is to reject the binary distinction between humans and technology and acknowledge the mediating role that technology plays in human life and that it is interwoven in our lives (Kiran & Verbeek, 2010). In other words,

[1] Opacity and the black box problem are not exclusive to DL as other forms of machine learning also can be opaque. Because DL is paradigmatic of the black box problem, it is the focus of this paper.

AI is situated within a socio-technical system that mediates trust, and by increasing the trustworthiness of these systems, we thereby increase trust in AI.

1 Trustworthiness and Transparency

There are at least two related senses in which we think or speak about trustworthiness. In the first sense, we talk about trustworthiness when an individual is deliberating whether to trust another or not. In these cases, one is making judgments about whether another person is trustworthy in the sense of able-to-be-trusted. The trustee's motivational states, interests, character, past performance, competency, and other personal characteristics all factor into the trustor's judgments. Trustworthiness in this sense is the judgment that to trust in a person is fitting or appropriate given the circumstances. We also talk about trustworthiness in the sense of the trustee's responsiveness to the trust placed in her. That one has been entrusted provides the trustee with reasons or motivations to be responsive in the appropriate way. In these cases, the trustee is making herself trustworthy in the sense of being in some way responsive to trust (Pettit, 1995; Hardin, 2004; Jones, 2012). Trustworthiness therefore can either be judgments about the conditions under which one places trust in another or one's being responsive in a certain way to having been entrusted with something.

Though the two accounts are related in that a trustor might consider whether a trustee would have reasons or motivations to be responsive to the trust placed in her in judging whether she is trustworthy, trustworthiness as a judgment about the fittingness of trust is temporally if not logically prior to trustworthiness as a responsiveness to trust. Trust typically is thought to be relational where A trusts B to do X (Flores & Solomon, 1998). Entering into such a relationship usually entails that one makes some judgment or otherwise determines the trustworthiness of another in the sense that one is justified or confident in placing trust. In this important way, A trusts B to X means that A has some judgment about B's trustworthiness or likelihood of making good on X. Trust in this sense requires the judgment that B is capable, disposed, and committed to acting in A's behalf in doing X. Such a judgment requires a well-grounded belief about B's ability and willingness to do X. In other words, A has good reasons to believe that B will do X.

What counts as good reasons for trusting is a matter of considerable debate, but at the very least that B is capable or competent to do X is necessary for A to trust B (Baier, 1986; Jones, 1996, 2012; Simpson, 2012). Those in whom we place trust usually are up to the task. Of course, there might be times in which one places trust in another with little knowledge of the trustee's competence, but these cases typically are either the first stage of an iterative process to gather information about the person's competence or as a means to induce the kind of responsiveness associated with the second sense of trustworthiness (Pettit, 1995). In typical cases however, A's judgment about B's trustworthiness involves some judgment about B's competence. In other words, B is deemed to be reliable in doing or achieving X. At minimum then, trustworthiness as the fittingness of trust can be understood as:

> (1) A trusts B to do X only if A judges B to be trustworthy where trustworthy means that A has good reason to believe that B is competent in doing X.

Many cases of reliance might also satisfy the conditions in (1) to the extent that one has good evidence of another's competence in doing X without necessarily trusting them. Everyday transactions with others, especially with those about whom we know very little, can be cases of reliance rather than trust. The barista at a local coffee shop is competent in fulfilling orders accurately and quickly. The barista certainly is competent in carrying out these duties, but what would justify the claim that one trusts the barista rather than simply relies on her? At the very least, with respect to their competence in fulfilling drink orders, trust in one's barista is indistinguishable from reliance.

Our relationship with technology often is one of reliance rather than trust. We rely on computer and digital technology to procure many goods and services: banking, transportation, healthcare, and, to an increasing degree, even education. Yet many people report a lack of trust in these technologies and the companies that manage and produce them. According to the 2020 Edelman Trust Barometer, trust in technology has declined in 21 of 26 markets surveyed; trust of AI was reported by less than 50% of respondents in the USA, Canada, the UK, Germany, France, and Ireland; and only 44% of respondents globally believe that the use of AI will have a positive impact (Edelman Trust Barometer, 2020). Yet the use of technology and AI continues to rise.

To make the distinction between reliance and trust clearer, several theorists of trust argue that trust, unlike reliance, involves vulnerability on the part of the trustor and so requires that the trustee be aware of and responsive to the trustor's interest or wellbeing (Baier, 1986; Hardin, 2004). We expect that the trustee will not take advantage of our vulnerability even if the opportunity should arise. Responses to violations of trust typically entail feelings of betrayal, not just disappointment, because the trustee not only has failed to be competent but also has failed to consider, honor, or respect the trustor's interests or goods (Baier, 1986).

The definition of trustworthiness in (1), therefore, is insufficient. Because trust requires some acknowledgment or consideration of the goods or interests of another, judgments about another's trustworthiness need to include these factors. Trustworthiness as the fitting of trust can be revised as:

> (2) A trusts B to do X only if A judges B to be trustworthy where trustworthy means that A has good reason to believe that B is competent in doing X and that B would act in A's behalf.

Acting in one's behalf means that one is acting in another's interest or for their good. In the context of trust, it means that one will not betray another should the occasion arise. That trust involves investment in another's interests or goods, and prohibitions against betrayal suggest that trust is also different from reliance in that it is a moral concept. Trust is a relationship of reciprocal duties, obligations, and expectations, and so feelings of betrayal or outrage are apt should trust be violated (von Eschenbach 2019).

Because trust is a moral relationship that takes into account interests and goods, acting in another's behalf requires that one also act in ways consistent with these goods and interests. *How* and *why* B does X is just as important as *that* B does X. B can betray A not only in failing to do X but also if she does X in a way that is not felicitous to A's interests or goods.

I entrust my broker with a sum of money to invest with the expectation that my broker will provide a reasonable return. I trust my broker with my money, judge her to be competent in producing returns on investments, and have reason to believe that my broker would act in my behalf by not betraying me in stealing my money. Suppose also that I am strongly committed to sustainability and counteracting the effects of climate change. Were my broker knowingly to invest my money in fossil fuel companies or coal power plants that are at odds with my commitment to combating climate change, I might still be justified in feeling betrayed even if my broker makes good on providing a return on my investment, especially if my broker is aware of my commitments. Acting in another's behalf, therefore, also entails that one act in ways consistent with, or at the very least not in opposition to, that person's values and commitments and that these provide some reason or justification for one's action.

For one to have good reason to believe that another would act in one's behalf then requires some understanding or insight as to how and why the person would carry out that task. Another's actions or intentions would have to be *transparent* for one to judge that person as trustworthy. Without knowing how another intends to or typically does carry out that which they are entrusted, one is unable to be sure of whether that person will act in one's behalf and therefore is trustworthy.

The standard account of trust, where one is judged to be reliable and competent, is not sufficient for "moral" trust. We need the further requirement that actions are done in the *right way (how)* and for the *right reasons (why)*. The procedures and principles one employs or follows in carrying out what one has been entrusted to do are as important as the reasons one has to do so. For trust to be fitting in a moral sense requires that the trustee follows norms and expectations in carrying out activities related to that which they have been entrusted to do. If one does so in a way that violates conceptions of justice or legality or even our deep commitments, we would not judge that person to be trustworthy and trust would not be fitting.

2 Opacity of Deep Learning

Significant concerns have been raised in recent years about the moral hazards associated with the increasing prevalence of algorithms and their use as a substitute for human judgment in decisions within the criminal justice system, consumer credit ratings, finance services, college admissions, and job applications, among others (Eubanks, 2018; O'Neill, 2020).Worse still, the use of algorithms seems to have a disproportionately adverse impact on the poor and marginalized in society due to implicit biases in these algorithms. Ethical concerns are exacerbated by the fact that often there is a lack of transparency into these algorithms and how they operate. Issues about by what standards we measure or assess the function of algorithms or how we govern their use become more acute (D'Agostino & Durante, 2018).

The Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS, is a commercial tool used by several courts and probation offices to predict recidivism of offenders seeking parole. A recent analysis by ProPublica determined that black defendants were much more likely to be incorrectly judged to be high risk for recidivism than white defendants, and white defendants were much more likely to be incorrectly judged to be low risk for recidivism than their black counterparts (Larson et al., 2016). The analysis also showed that "even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 45 percent more likely to be assigned higher risk scores than white defendants" (Larson et al., 2016). Despite their relatively low level of criminality, female defendants also were almost 20% more likely to receive a high-risk score than men (Larson et al., 2016).[2]

ProPublica's analysis highlights one notable example of widespread use of an algorithm that is flawed because of systematic bias that often reflects existing social, racial, and economic inequities. The dearth of precise data in some instances can introduce bias into algorithms by forcing designers to develop proxies for the information they are seeking. Algorithms used as alternatives to credit agency scores to assess loan applications have come under scrutiny because of the use of proxies, such as zip codes, which disproportionately discriminate against minority communities (O'Neill, 2020).

The use of algorithms itself is not objectionable especially when algorithms can be more accurate than human judgment. Moreover, that algorithms can exhibit bias does not differentiate them from human decisions or evaluations. What is objectionable about the use of these algorithms in making decisions that significantly impact people's lives, however, is that because of proprietary interests or confidentiality, algorithms lack transparency and so results are difficult, if not impossible, to dispute or appeal. The harms that they can potentially perpetrate often have no remedy, and those who suffer these harms consequently lack recourse to address them.

Transparency is an even bigger issue for DL networks that are becoming more prevalent in AI. Machine learning systems and deep learning networks can be opaque due to the absence of any mechanism to reproduce or explain decision-making processes or "reasons" for reaching a decision given that "*ex ante* predictions and *ex post* assessments of the system's operations alike will be difficult to formulate precisely" (Zerilli et al., 2019, 664). This "black box problem" in AI becomes especially worrisome when coupled with outcomes that are ethically problematic, such as biased algorithms or decision models. Without an adequate understanding of how decisions are reached using DL, the criteria for trustworthiness cannot be satisfied.

[2] ProPublica's analysis of COMPAS is useful for heuristic purposes but not without criticism. Subsequent analyses have raised questions about ProPublica's conclusions regarding racial bias but uncovered other serious concerns that remain hidden due to a lack of transparency (Fisher et. al., 2019; Rudin et. al., 2020). Other analysis has suggested that it is no more fair or accurate than human judgment (Dressel & Farid, 2018).

Opacity in machine learning is a complex and nuanced phenomenon that may admit of variation depending upon particular stakeholders, their interests, and sophistication (Zednik, 2019). Systems can be opaque to the extent that epistemically relevant elements are unknown to an agent and so opacity might vary depending on the agent (Humphreys, 2009). Generally speaking, however, machine learning algorithms can be opaque in two senses: (1) the process or mechanism for how machine learning arrives at outputs from given inputs may be inaccessible or unknowable, and (2) inputs themselves may be unknown to programmer or observers. This opacity might be due to proprietary concerns, technical illiteracy, or the characteristics of machine learning (Burrell, 2016). The innate complexity and nonlinear functions stemming from the neural network architecture of DL and that it uses hundreds of billions, if not trillions, of parameters in carrying out its computation is what contributes to DL's opacity in the last sense. This type of opacity differs from the opacity of simpler algorithms that are opaque due to proprietary considerations or concerns about competition in that there is no easy way in which to make DL transparent. Companies can disclose their proprietary algorithms to other experts to make them transparent, but because of DL's scale and complexity, it may never be transparent to anyone, not even to the expert.

Take for example the case of Deep Patient, in which researchers created a deep neural network to examine 12 years of electronic medical records from 700,000 patients. Deep Patient proved to be better able to predict future disease and disabilities, especially diabetes, schizophrenia, and cancer, using information such as previous diagnoses, medications, lab results, and procedures than simpler algorithms or human counterparts (Miotto et al., 2016). Because Deep Patient is opaque, however, researchers do not know how or why it comes to its diagnoses and therefore cannot learn from it nor explain to the patient exactly what causes or factors are contributing to their morbidity and how.

3 Trust Issues

The opacity of "black box" AI systems and some algorithms poses significant challenges from an ethical perspective especially when considering questions of trust. It is clear that in many cases, such as Deep Patient, these systems can be more reliable, if not more competent, than human judgment. Because reliability is only one criterion for trustworthiness, however, predictive power and competency alone are insufficient reasons to trust these systems.

As discussed previously, trust requires that we also appreciate or understand how something or someone will carry out the task at hand. Because these systems are opaque in precisely the sense that we cannot or do not know how the algorithm or neural network reaches its outcome, these systems fail to satisfy this important criterion to be judged trustworthy. We might judge these systems to be reliable, but trust would not be fitting due to the absence of evidence that these systems would carry out these functions in a manner consistent with our goods or interests.

To return to a previous example, the use of algorithms might be reliable as a means for predicting recidivism in some specific cases, but without understanding how the algorithm reaches its conclusion or ensuring that it is free from bias, it cannot be trusted to execute its purpose of securing the aims of our criminal justice system. Because equality and impartiality before the law are foundational principles of our justice system, we have an interest in ensuring that these goods are upheld and honored along with other aims. Without adequate evidence or transparency into the operations of the system, we cannot be assured that these algorithms are acting on our collective behalf.

This example is less concerning because eventually we were able to understand the algorithm and identify through further analysis that there was bias in the system. In a sense, the number of false positives and false negatives rendered the system less reliable as well as inadequate for upholding other values of our criminal justice system and so was seen as deficient even from a technical standpoint. The picture becomes much more complex, however, when we consider systems involving artificial intelligence and machine learning.

Because of their speed and sophistication, DL and AI systems are becoming increasingly more relied upon in carrying out complicated predictions across a number of domains, but as discussed, also much more difficult to understand without adequate transparency into how these systems work. Due to their increasing reliability and accuracy, moreover, these systems are being employed to make decisions involving significant human goods and interests, such as national security, healthcare, transportation, finance, and information systems. The absence of transparency coupled with increased prevalence of DL present potential ethical dilemmas especially when the consequences of these outputs are significant, such as when AI predicts acts of foreign aggression, makes specious medical diagnoses, or causes driverless vehicle accidents (Bleicher, 2017; Guidotti et al., 2018). The inability to understand or explain why errors were made or how conclusions with significant consequences were reached presents considerable challenges and undermines one's confidence, if not trust, in these systems.

With respect to trust, however, not all transparency is equal. Certainly, comparisons between DL's neural networks and human brains can be made, and, in many ways, human cognition can be viewed as a black box as well (Burrell, 2016; Castelvecchi, 2016). Merely inspecting the inner workings of these devices and systems, however, might not be possible for practical reasons and falls short of the kind of transparency required for trust (Dahl, 2018). Just as full knowledge of the inner workings of brain functions of those in whom we have entrusted something of value in itself would not increase our trust or justify judgments of their trustworthiness, so too does full knowledge of decision-making processes of DL fall short of the transparency required for trust.

In addition to assessing competency, when making judgments about another person's trustworthiness, we also seek to understand their reasons or motivations for carrying out the task for which we are entrusting that person. More specifically, we want to have reason to believe that these reasons or motivations take into account our goods and interests and will lead to actions consistent with them. When asking why or how this person will safeguard my interests, I am not asking for a description of the underlying physical mechanisms for arriving at an outcome, but the higher-level principles or motivations for undertaking relevant activities. I want good reasons to believe that the person is trustworthy in the sense that she is competent in carrying out what she has been entrusted to do and will do so in my behalf. In the context of technology, what we are seeking then is interpretability, or the ability to understand why or how a decision was reached by AI, rather than mere transparency into the inner workings of the black box (Guidotti et al., 2018).

It should be noted that we are not seeking a full and comprehensive interpretation of the entire logic and workings of the system or model, what some call "global interpretability," but rather reasons for the specific decision or outcome, or "local interpretability" (Guidotti et al., 2018, 6). This is not unlike situations involving interaction between persons, such as a doctor-patient relationship. To earn the trust of the patient, for example, a doctor needs to demonstrate competency and that the patient's best interests are reasons or motivations for medical decisions. The patient does not need to understand the entire biomedical basis for these decisions, such as disease pathology and treatment interdictions, which might require years of study and practice. Global interpretability is not necessary for trust so long as the decision is understood or interpreted to be competent and consistent with one's goods and interests.

4 Explainable AI

XAI refers to that class of models developed to address the black box problem in AI by making DL more transparent, interpretable, and explainable. XAI provides a simplified model to assist us in understanding AI's decision-making processes, its strengths and weaknesses, and how it might behave in the future. With complex AI systems that use DL, XAI can provide post hoc interpretability, or a means to "approximate deep-learning black box models with simpler interpretable models that can be inspected to explain the black box models" (Rai, 2020, 138). These models are promising for helping to understand black box technologies because they can be constructed to interpret the black box both at the global and the local level. This can be achieved by developing interpretative models that approximate DL processes, such as linear approximations or decision trees, or by using post hoc interpretative methods, such as natural language explanations or functional models (Páez, 2019). The goals for each might differ in that some models seek to understand the decision, while others seek to understand the process or function of the system. An important question, however, is whether XAI gives us reasons to trust the system that is being modeled or what seems more likely, the model itself (Rai, 2020, 139). Answering this question in part will require further analysis of how XAI might render DL technologies more transparent and thus allow for interpretation or explanation of its function.

Some XAI models are transparent in the sense that they reveal the inner workings of the black box system. Other models seek to make the decision itself understandable to observers without necessarily providing an objective description or reconstruction of the processes by which the decision has been reached. These two ways of differentiating between XAI correspond roughly to two different kinds of explanations: explaining what versus explaining why (Páez, 2019; Zednik, 2019).

Both types of explanations are important for rendering the black box more transparent, but each will be relevant for different types of stakeholders: "operators seek to render a system transparent by asking *what* it does and by describing its 'input' and 'output' states, several other agents do so by asking *why* it does what it does and interpreting those states in terms of environmental features and regularities" (Zednik, 2019). Whether we ask why or what AI is doing will depend on the kind of question for which we are seeking an answer. Because trust requires that we make some judgment about how and why someone will fulfill that for which one has been entrusted, why-explanations are the relevant class of explanations we should seek for making AI more transparent.

Though it is true that at the local level and for most decisions, understanding what the black box does is necessary for understanding why it does what it does, the latter is more relevant for judgments about its trustworthiness (Páez, 2019). Indeed, XAI does provide some models that can be used to understand why DL reaches a particular outcome but with limitations and caveats that ultimately having bearing on the question of trust. Heatmapping, whereby visual representation is used to emphasize features that contributed most to a particular classification, for example, can provide insights into how certain outputs were achieved without fully understanding every detail of the process. But this kind of modeling is useful for why-explanations when "the highlighted elements together *look like* some recognizable feature of the environment" (Zednik, 2019, emphasis original).

With this type of XAI, one might come to some partial knowledge of the black box technology without understanding fully its inner workings: "a user's understanding of the key details can provide reasons for and against using the device without requiring that the inner procedures of the web device be totally transparent; instead, it is enough that certain key details are indicated by the web device and thereby understood by the user" (Dahl, 2018, 575). Two important questions, of course, are what is meant by a "key detail" and what are the criteria that give some details special significance or consideration over others. One possibility is that understanding a "key detail" means that an individual understands just enough or only those parts necessary to understand why the technology operates within a defined context, but without understanding the whole. For example, I might understand enough about how automobiles operate to know that they will not function without a fuel supply (and so why I need to fill my gas tank when empty), but without having full knowledge of how an internal combustion engine operates. In the context of information technology, this might mean that I have knowledge of why a device obtained its result (by pointing me to a reliable, known information source) without understanding fully the computer code by which the system operates (Dahl, 2018, 576).

In effect, possessing understanding of "key details," where "key detail" means those limited facts relevant to the task at hand, is to possess transparency, or relevant knowledge of how that key detail factors into the process. Where XAI refers me to a known, reliable information source or a model, such as heatmapping, my trust lies with those sources and only extends to the technology by association. In cases of DL, however, where new knowledge is being generated through a complex and opaque process, there may not be a trusted information source to appeal to in response to an inquiry. It is an open question to what extent and under what conditions trust is transferrable, but more to the point, this model for obtaining reasons to justify judgments of trustworthiness of technology would apply to DL in a limited number of cases.

We might also test black box technology through experience, whereby we develop a "a means of checking whether the web device's outputs are correct, a means independent of the web device itself, and independent of testimony from those who can inspect the inner procedures that produced the outputs" (Dahl, 2018, 584). In other words, by testing the technology and observing its outcome over time, we can come to have greater confidence in the technology's ability to provide consistent and accurate outcomes, save for reasons or factors to the contrary. We can make some inductive generalizations about the technology's performance, compare it to independent sources or processes with which we are familiar, and make functional generalizations (Dahl, 2018, 585; Páez, 2019, 454). An important question, however, is whether XAI gives us reasons to trust the system that is being modeled or what seems more likely, the model itself or black box technology without sufficient justification, especially where understanding why requires understanding how (Rai, 2020, 139; Páez, 2019, 455).

Because XAI models that offer why-explanations can also be relatively complex and difficult to interpret, especially in cases where we risk oversimplifying or misrepresenting that which is being represented, and because why-explanations are dependent to a large extent on how-explanations, they might be transparent only to a limited number of stakeholders. In cases where less technically savvy stakeholders are seeking why-explanations that are akin to folk psychology and where these explanations are unavailable through XAI, these stakeholders would need to rely upon the expert testimony of others. Our trust of experts and their knowledge through XAI of the processes by which the black box technology functions and acceptance of their judgment for the trustworthiness of the technology serve as reasons for our own judgments. Certainly, we often rely on the expert testimony of others in making decisions about complex matters, but strictly speaking, in these instances, our trust is placed in the expert. We lack any means of justifying our judgment of trustworthiness independent of the expert's testimony and so accept his or her judgment as a proxy for ours.

Similarly, we come to trust technology through social vetting. Social vetting is a scaled version of individual understanding and expert testimony whereby we rely on the collective expertise and judgment of others to determine the trustworthiness of the black box technology. In social vetting, a group, presumably of experts, tests a device through collecting their individual experience and judges the device to be trustworthy based on the evidence collected.

Social vetting is an effective means for assessing the experience and judgment of others, especially groups of experts, but in itself may not provide reasons for trusting the black box technology. At best, social vetting offers a substitute for our individual judgment but is effective only if we already have reasons to trust those whose judgment we are substituting for our own. In fact, it is entirely consistent with social vetting that one remains skeptical of the black box technology but nonetheless willing to consider accepting its results based on the recommendation of others. The reasons for so doing have little, if anything, to do with informed judgments about black box technology but rest entirely with reasons for trusting one's community of advisors and experts. Increasing public trust of experts and stewards of technology offers a compelling means to overcome the limitations of XAI and to provide reasons for trusting black box technologies.

5 Trusting the Socio-technical System for AI

For some, to ask whether we can trust technology is akin to making a category mistake because trust can occur only between people or moral agents (Pitt, 2010). Technology, to the contrary, should be seen only as artifacts, to which concepts of trustworthiness cannot apply (Nickel et al., 2010). To ask if one can trust AI, whether or not it involves black box technology, is the wrong question to ask, according to these views. Instead, we need to ask if we can trust the people who design, implement, and use these technologies.

Though trust is understood traditionally to be a moral concept that applies to persons and governs interpersonal interactions, to conclude that trust cannot apply to AI because this class of artifacts are not persons is to oversimplify the case. For one thing, as these technologies become more sophisticated, the line between artifact and person becomes less clear to the point where we might consider them to be artificial agents worthy of moral consideration (Floridi & Sanders, 2004). But more to the point for the purposes of this discussion, we interact with AI in ethically significant ways with increasing power, prevalence, and, ultimately, vulnerability.

For these reasons, we should not think only in binary terms of whether it is coherent to speak of trusting artifacts or whether trust is reserved for persons alone but acknowledge that these interactions occur within a complex and often diffuse context, what many have referred to as the socio-technical system. A socio-technical system sees technology as more than "a collection of devices intermediating between their designers on one hand and their users on the other" but is understood to be a hybrid between the technical and social (Nickel et al., 2010). Treating technology in this way not only is more faithful to the phenomenology of using technology in everyday life but avoids the philosophical problems associated with attributing moral properties to artifacts and with limiting trust strictly to persons.

By extending the domain of inquiry to the social-technical system, the relevant ethical questions then become whether one can trust the socio-technical system of AI and black box technologies, and whether the constituent members of this system are trustworthy. Admittedly, there will be numerous components to this system that would need to be considered, each of which would have different kinds of relationship with AI, just as there are different stakeholders with respect to its use. In addition to end users and the technical device itself, we should consider the designers, programmers, "data subjects," operators, "decision subjects," and examiners, or those tasked with auditing and inspecting a system (Zednik, 2019). Each of these agents will have different roles to play and interests in AI, but what unifies the various components and stakeholders is a shared conception of the purpose or goal of the system as a whole.

To give a somewhat simplified example, DL systems used to diagnose and prognosticate the presence and progression of disease, such as cancer, occur within a socio-technical system that includes the doctor, patient, technicians, hospital administrators, and health insurance companies, as well as AI designers, operators, and AI tools. Though each has a different role, level of understanding of AI, and potential for "opening" the black box, they have the common goal and interest in treating disease in the most efficient and effective manner possible. Trust with respect to technology, therefore, can only be understood in reference to the system as a whole, and each agent's trustworthiness will be judged relative to the differences in roles, interests, and expertise. The patient, as the end user, will trust the doctor based on his judgment about her expertise, reliability, and whether she is acting in the patient's behalf. With respect to the black box technology, the patient then would be willing to trust the judgment of the physician, who in turn trusts the judgment of the technician, who presumably can understand the why-explanations of XAI.

Thus, trust in AI requires what Durante calls a "web of trust" that "is founded on *circular informational causality*" oriented toward a shared goal or purpose (Durante, 2010, 355). From the perspective of the user (i.e., patient), the physician in this case becomes the proximate interface to the entire socio-technical system, of which the black box technology is an essential but component part and may remain opaque to the patient. Trust in technology then is properly understood as a social phenomenon for these reasons rather than a relationship that holds between two agents or an agent and an artifact.

Conceiving the problem of trust of AI in terms of the socio-technical system also helps address two significant skeptical arguments. The first worry is that in requiring transparency for trust, we are holding black box technology to a different standard than we hold agents to in interpersonal interactions. The transparency requirement amounts to "a double standard in which machine tools must be transparent to a degree that is in some cases unattainable, in order to be considered transparent at all, while human decision-making can get by with reasons satisfying the comparatively undemanding standards of practical reason" (Zerilli et al., 2019, 668). By focusing on standards for trust with respect to the socio-technical system, which includes humans and machines, concerns about whether the kinds of transparency required for technology differ from standards of practical reason and the extent to which humans remain opaque to themselves are less relevant. For most of us, our trust in AI will be mediated through our trust in the experts and their testimony about the trustworthiness of these technologies, and in these cases, the reasons for trusting AI will be articulated using familiar folk psychology terms and similar explanations (Zednik, 2019).

The second skeptical worry claims that efforts to make AI transparent are mistaken because such efforts confuse explaining the results of a decision with that entity or process for making the decision and that the explicability requirement is for the purpose of maintaining human control. Explanations are required only in cases where something moral is at stake. Decisions that require explanation, the argument concludes, therefore should not be made by AI (Robbins, 2019).

We might agree with the premises of this argument while rejecting its conclusion by recognizing that the decision is the result of a socio-technical system.

The use of AI for low-risk decisions carries little, if any, moral hazard and so does not require explanations, but it does not follow that AI should not be used for high-stake decisions. If AI were to remain completely opaque and without explanation, then it would be problematic from a moral point of view to trust these decisions to AI. XAI promises to make these processes more transparent to some, but more to the point, by embedding the decision in a socio-technical system, AI is *only part* of the decision-making process. If we are justified in trusting the socio-technical system, of which AI is a part, then we can still use AI for high-stake decisions because AI is not the sole decision-maker.

6 Conclusion

Trust requires that we have reason to believe both that AI is reliable and acting in our behalf, and so transparency into how AI operates and reaches its outcomes and predictions is needed in order for us to be able to judge AI to be trustworthy. The black box problem associated with deep learning machines threatens to thwart our ability to make such judgments in spite of best efforts to model these systems either through XAI or other means. This is not to say that we would never be justified in accepting the authority or outcome of black box AI. Though AI remains opaque to many of us, we have seen that we can have good reasons to trust the experts and organizations who use these technologies in our behalf. We have seen how XAI can offer limited success in opening the black box and how this can justify trust in the socio-technical system in which this technology is embedded.

Because XAI alone is insufficient to justify trust of black box technology for most stakeholders, especially the less technically savvy end users, ethicists and practitioners alike should shift their attention to ways in which the socio-technical system can increase trustworthiness. Making black box technology more transparent will remain an essential part of this process, but other avenues of inquiry, such as ethics-based auditing of AI, also are critical to this project and should be explored further (Mökander & Floridi, 2021).

Black box AI should be recognized as being embedded in a larger context of institutional or organizational norms and standards that safeguard the interests and goods of those it serves. Rather than only seeking ways in which AI can be made trustworthy, companies and institutions that develop and use them and who have great control over the socio-technical system need to ensure that they themselves earn our trust. Because trust not only has the power to obligate but also provides motivation or reasons to make good on these obligations, then a very promising avenue to investigate would be the extent to which technology leaders and organizations might wield trust, as one would wield power, to restore our faith in these technologies.

References

Baier, A. (1986). Trust and antitrust. *Ethics*, *96*(2), 231–260. https://doi.org/10.1086/292745

Bleicher, A. (2017). Demystifying the Black Box that is AI. *Scientific American.* https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/. *Accessed 6/4/2020.*

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1). https://doi.org/10.1177/2053951715622512

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature, 538*, 21–23. https://doi.org/10.1038/538020a

D'Agostino, M., & Durante, M. (2018). Introduction: The governance of algorithms. *Philosophy and Technology, 31*(4), 499–505. https://doi.org/10.1007/s13347-018-0337-z

Dahl, E. S. (2018). Appraising black-boxed technology: The positive prospects. *Philosophy and Technology, 31*(4), 571–591. https://doi.org/10.1007/s13347-017-0275-1

Danaher, J. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy and Technology, 29*, 245–268. https://doi.org/10.1007/s13347-015-0211-1

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1). https://doi.org/10.1126/sciadv.aao5580

Durante, M. (2010). What is the model of trust for multi-agent systems? Whether or not e-trust applies to autonomous agents. *Knowledge Technology & Policy*, *23*, 347–366.

Edelman Trust Barometer. (2020). *Special report: Trust in Technology.*

Eubanks, V. (2018). *Automating inequality : How high-tech tools profile, police, and punish the poor* (1st ed.). St. Martin's Press.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*, 1–8.

Flores, F., & Solomon, R. C. (1998). Creating trust. *Business Ethics Quarterly, 8*(2), 205–232.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machine, 14,* 349–379.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Hardin, R. (2004). *Trust and Trustworthiness*. Vol. 4. The Russell Sage Foundation. https://doi.org/10.4324/9781315542294-2

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese, 169*(3), 615–626. https://doi.org/10.1007/s11229-008-9435-2

Jones, K. (1996). Trust as an affective attitude. *Ethics, 107*(1), 4–25. https://doi.org/10.1086/233694

Jones, K. (2012). Trustworthiness. *Ethics, 123*(1), 61–85. https://doi.org/10.1086/667838

Kiran, A. H., & Verbeek, P-P. (2010). Trusting our selves to technology. *Knowledge Technology & Policy*, *23*, 409–27.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. *ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Accessed 6/4/2020.

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, *6*(May), 1–10. https://doi.org/10.1038/srep26094

Mökander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. *Minds and Machines*. Springer Science and Business Media B.V. https://doi.org/10.1007/s11023-021-09557-8

Nickel, P. J., Franssen, M., & Kroes, P. (2010). Can we make sense of the notion of trustworthy technology? *Knowledge Technology & Policy*, *23*, 429–44.

O'Neill, O. (2020). Trust and accountability in a digital age. *Philosophy, 95*(1), 3–17. https://doi.org/10.1017/S0031819119000457

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines, 29*(3), 441–459. https://doi.org/10.1007/s11023-019-09502-w

Pettit, P. (1995). The cunning of trust. *Philosophy & Public Affairs, 24*(3), 202–225. https://doi.org/10.1111/j.1088-4963.1995.tb00029.x

Pitt, J. C. (2010). It's not about technology. *Knowledge Technology & Policy*, *23*, 445–54.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*, 137–141. https://doi.org/10.1007/s11747-019-00710-5

Robbins, S. (2019). A misdirected principle with a catch: Explicability for AI. *Minds and Machines, 29*(4), 495–514. https://doi.org/10.1007/s11023-019-09509-3

Rudin, C., Wang, C., & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, *2*(1), 1–54. https://doi.org/10.1162/99608f92.6ed64b30

Simpson, T. W. (2012). What is trust? *Pacific Philosophical Quarterly, 93*(4), 550–569. https://doi.org/10.1111/j.1468-0114.2012.01438.x

Taddeo, M., & Floridi, L. (2011). The case for e-trust. *Ethics and Information Technology*, *13*, 1–3. https://doi.org/10.1007/s10676-010-9263-1

von Eschenbach, W. J. (2019). Trust as a public virtue. In J. Arthur (ed.), *Virtues in the public sphere: Citizenship, Friendship, Public Duty*. Routledge. https://doi.org/10.4324/9780429505096

Zednik, C. (2019). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, *34*, 265–288. https://doi.org/10.1007/s13347-019-00382-7

Zerilli, J., J. Maclaurin, A. Knott, and C. Gavaghan. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy and Technology*, *32*(4), 661–683. https://doi.org/10.1007/s13347-018-0330-6