Forthcoming in *Social Epistemology*

Final submitted version, November 2023

We have no satisfactory social epistemology of AI-based science

Inkeri Koskinen

University of Helsinki

inkeri.koskinen@helsinki.fi

Abstract: *In the social epistemology of scientific knowledge, it is largely accepted that relationships of trust, not just reliance, are necessary in contemporary collaborative science characterised by relationships of opaque epistemic dependence. Such relationships of trust are taken to be possible only between agents who can be held accountable for their actions. But today, knowledge production in many fields makes use of AI applications that are epistemically opaque in an essential manner. This creates a problem for the social epistemology of scientific knowledge, as scientists are now epistemically dependent on AI applications that are not agents, and therefore not appropriate candidates for trust.*

Keywords: trust, reliance, epistemic opacity, epistemic agency, extended knowledge, extended agency

1. Introduction

The aim of this paper is to argue that currently we do not have a satisfactory social epistemology of AI-based science, and to explicate the problem. Epistemically opaque AI applications are gaining a new kind of a role in scientific knowledge production. This change presents a challenge to a view of relationships and networks of trust that is widely accepted in the social epistemology of scientific knowledge.

It is becoming common in many fields to use AI applications – such as simulations or image classifiers based on machine learning techniques – in knowledge production. Many such applications are epistemically opaque in an essential manner: the computational processes are so fast and complex that it is impossible for human agents to fully grasp how they reach their

results. Their epistemic opacity has given rise to discussions about the trustworthiness of AI systems in many domains in society, not just science. In this article I focus on a specific challenge that the use of epistemically opaque AI applications in science creates for the social epistemology of scientific knowledge, a problem I believe has not yet been discussed in the field. I argue that we currently have no satisfactory way to reconcile the practices of AI-based science with the idea – supported by many in the field – that relationships of trust, not of mere reliance, are unavoidable and indispensable in contemporary science, where knowledge is largely produced collectively and scientists depend on each other.

The kind of relationships of trust that are meant here are possible only between agents who are capable of taking responsibility for their actions. AI applications are not agents in this sense, so we can only rely on them. Due to their epistemic opacity, however, such reliance is blind: scientists' epistemic dependence on the AI applications they use can be permanently opaque. This is in conflict with the idea that epistemic dependence in science is managed with collectively and institutionally controlled relationships of trust between responsible agents. The situation can be analysed in several ways. One that is common in the current literature on AI-based science is to focus on the reliability of the applications, or to develop accounts of trust that allow for trust in AI applications (see e.g. Grodzinsky, Miller & Wolf 2020). However, this literature typically disregards the arguments emphasising the importance of relationships of trust in scientific collaborations. Another option is to acknowledge that AI applications are given a role as some kind of epistemic quasi-agents in scientific knowledge production, even though they cannot be be held responsible, and therefore cannot be trusted, only relied on. And a third option is to argue that a human and an AI application can together form an extended agent, where the human takes responsibility for the workings of the whole even though the AI part of the agent is epistemically opaque to them. I will argue that all of these interpretations challenge the way in which relationships and networks of trust in research groups and scientific communities are generally understood in the social epistemology of scientific knowledge.

I will start with an overview of the claim that relationships of trust are essential in contemporary science, and argue that the notion of trust used in this claim is a thick one: we can only trust agents who can take responsibility for their actions and choices (Hardwig 1985; 1991; Goldberg 2011, 2016; Wilholt 2013; Wagenknecht 2014a, 2015; Reider 2016; Miller & Freiman 2020; Rolin 2020). I continue with a brief discussion of how the role of instruments in scientific knowledge production is understood in this literature. Then I introduce Paul Humphreys' (2009) notion of essential epistemic opacity, and the claim that many widely used AI applications are epistemically opaque to us. When discussing the role complicated instruments have in scientific knowledge production, philosophers have presented both accounts that describe the role of AI applications in science and address the issue of epistemic opacity (e.g. Durán and Formanek 2018), and accounts that explicitly

address the question of agency in situations where we are epistemically dependent on machines and instruments (e.g. Giere 2004; Clark 2015; Palermos and Pritchard 2016), though not both simultaneously. As I argue, none of these accounts is easily compatible both with the current use of epistemically opaque AI applications in science and with the social epistemological understanding of the necessity of controlled relationships and networks of trust in science. This leaves us without a satisfactory social epistemology of AI-based science.

2. The necessary trust view

It is largely accepted in the social epistemology of scientific knowledge that trust plays an indispensable role in scientific knowledge production today. Given the collaborative nature of science, scientists cannot help but trust other scientists. This view takes several slightly different forms, as different philosophers focus on different aspects of the relationships of trust in science. However, the general view that without relationships of trust, the kind of collective knowledge production that characterises contemporary science would not be possible, is largely accepted. I will call this the *necessary trust view*. It virtually always includes descriptions of ways in which such relationships of trust are and must be controlled for them to be epistemically acceptable.

An important part of the new scientific knowledge that is being produced in universities, research institutes, and the like, could not be produced by individual researchers alone. Research is increasingly done by teams where no one is epistemically in a position where they could vouch for the entirety of the collaboratively produced results. Evidence is gathered collectively, scientists rely on each other's expertise, research groups reach results by dividing labour, and training and critical discussions within scientific communities as well as many collectively functioning institutions, such as peer review, are essential in contemporary science.

John Hardwig (1985) started the current philosophical discussion about the role of trust in science by drawing attention to the epistemic significance of trust in collective knowledge production. He argues that in epistemic communities where epistemic labour is divided, the trustworthiness of the members is "the ultimate foundation for much of our knowledge" (Hardwig 1991, 694). No individual researcher has full justification for knowledge claims that are produced collectively, so either only the group or community can be said to have knowledge, or we must accept that individual researchers know even though some (often large) part of the justification of their beliefs is replaced by trust in their colleagues.

Briefly put, trust is necessary because collective knowledge production is characterised by relationships of epistemic dependence. Not everything scientists do can be double-checked; scientific collaborations are in practice possible only if its members accept each others'

contributions without such checks. Not only does a scientist have to rely on the skills of their colleagues, but they must also trust that the colleagues are honest and will not betray them, for instance by intentionally or recklessly breaching the standards and practices accepted in the field, or by plagiarising them or someone else (Frost-Arnold 2013). This is particularly clear when a relation of epistemic dependence on one's peers is opaque: a scientist does not have the expertise that would make it even possible for them to double-check the work of their collaborators (Wagenknecht 2014a).

Torsten Wilholt (2013, 251) has presented additional reasons for thinking that collective scientific knowledge production "has to be bound together by trust". He draws to attention trust and value-decisions in scientific collaborations. Many philosophers of science agree today that value-decisions are unavoidable in science, as scientists must make trade-offs between different values and risks throughout the research process. In scientific collaborations scientists must trust that the decisions their colleagues make are acceptable.

Relationships of trust in science are of course far from unquestioning. What I call the *necessary trust view* does not suggest that scientists could trust each other as easily as people trust their friends. While it acknowledges that relationships of trust are ineliminable in science, the view is not only descriptive, but also normative: virtually all versions of it include discussions about the ways in which scientific communities do and must control relationships of trust.

As Susann Wagenknecht (2015, 160) argues, scientists do not trust "indiscriminately, blindly, deeply, and completely": trust is an indispensable means to manage epistemic dependence in science because it is unavoidable, not because it is particularly sought. Many practices in science are designed so as to minimise the need for trust; scientific communities control the quality of the produced knowledge through critical discussions, peer review and training. And while colleagues are generally trusted, a part of this trust is the trust that other scientists work in ways that enable retrospective control and minimise the risk of error. Scientists are generally supposed to document the research process in a way that makes it possible for other experts to check their work if the need arises. This is particularly important because the epistemic and cognitive processes of individual researchers are not fully transparent even to themselves. Our cognitive capabilities are limited, and, for instance, it is not possible for an individual researcher to be fully aware of all the background assumptions on which they build their work (Longino 1990). Research communities have developed systematic ways to mitigate epistemic risks that arise from our limitations, and ways to ensure that possible errors are recognised and corrected. In other words, while no measures or indicators can completely ensure the trustworthiness of our colleagues, relationships of trust in science must be rational (Origgi 2018; Miller & Freiman 2020). Various collective and institutional measures of control are in place and must be in place for the ineliminable relationships of trust to be epistemically acceptable.

However, even these measures typically require some trust in researchers dutifully following the collectively accepted procedures, and truthfully documenting their own work. And finally, even the diverse quality checks create relations of epistemic dependence, and require trust in the the people who monitor and review other people's work. As Sanford Goldberg (2011, 120) summarises this, we depend on a "wide network of people who 'police' our epistemic communities". Such networks of relations of epistemic dependence are essential in contemporary science, and they always necessitate some trust. (Rolin 2020, Miller & Freiman 2020; Longino 2022.)

3. The necessary trust view uses a relatively thick conception of trust

The necessary trust view states that relationships of trust are a prerequisite for the kind of collective knowledge production and division of epistemic labour that characterise contemporary science. What kind of trust are we talking about?

Philosophers engaged in discussions about trust have suggested various accounts of the notion. One important dimension on which these accounts differ is the amount of moral and even affective weight the notion is taken to carry. According to a widely accepted view, we can distinguish between trust and reliance by saying that trust can be betrayed, whereas reliance can only be disappointed (Baier 1986). In other words, relationships of trust can only be found between agents who are in principle capable of betrayal, and who can be held responsible for their choices and actions. I will call "thick" the accounts of trust that emphasise the moral and/or affective elements of trust, and therefore necessarily take trust to be an appropriate attitude only towards agents. Annette Baier (1986; see also Jones 1996) has defended a particularly "thick" notion of trust, emphasising the affective dimension of trust; when we trust, we depend on the goodwill of others. The kind of trust we have in our friends is a good example of affectively loaded trust that is not easily shaken by evidence pointing to unreliability (Baker 1987; McGeer 2008). Some thick accounts of trust emphasise the responsiveness of the trustee: a trustworthy person is one who "takes the fact that they are being counted on to be a reason for acting as counted on in their motivationally efficacious practical deliberation" (Jones 2012, 66; see also Nguyen 2022). Such an account requires the trustee to be capable of being motivated by moral considerations about issues such as goodwill, conscientiousness, or integrity (Jones 2012). At the other end of the continuum we find less demanding, "thin" accounts that do not emphasise the moral and/or affective dimensions of trust. They often do not make such a sharp distinction between trust and reliance, and/or they do not require the trustee to be an agent. A simple doxastic account of trust, for instance, treats trust merely as a species of reliance: trust is reliance on someone to do something grounded on the belief that they will do it. Or, to give another example, Philip J. Nickel (2013) defends a thin account of trust that allows for trust in mahcines: according to his entitlement account, when we trust something, we believe it is worth relying on, and we

believe that we entitled to rely on it. To summarise, we can identify a continuum of conceptions of trust that vary from "thin" to "thick", depending on how strongly the notion is taken to be morally and/or affectively loaded and, as a result, how clearly trust is taken to be an appropriate attitude only towards agents capable of taking moral responsibility. (Hardin 1991; Petitt 1995; Hawley 2014; Goldberg 2020a.)

The notion of trust used in the the different variants of what I have called necessary trust view is relatively thick. It requires a clear distinction between trust and reliance, and it takes trust to be an appropriate attitude only towards full agents. Trust is needed in contemporary science because scientists depend on each others work, and their dependence is often opaque. Saying that a member of a research group relies on the work of the other group members is not a sufficient description of their relationship. When a scientist trusts his colleague while being epistemically dependent on her work in an opaque manner, he trusts that she will not betray him, for example by fraudulent or blatantly careless work (Hardwig 1991; Wagenknecht 2015; Rolin 2020). Scientists have to trust that their colleagues follow collectively accepted procedures – and this trust includes morally loaded trust in their dutifulness and truthfulness. Our reliance on peer review, such as it is, is dependent on us trusting in the sincerity, honesty, and thoroughness of the anonymous reviewers. As Hardwig (1991) and Goldberg (2011) among others note, the institutional systems and procedures that are in place in order to "police" scientific communities all require some trust in the people doing the policing; trust that is necessarily of a moral kind. Moreover, as Karen Frost-Arnold (2013) has pointed out, when scientists share their work, they must believe that their colleagues will not steal their ideas or materials, or plagiarise them, which requires moral trust. And when Wilholt (2013) argues that in scientific collaborations scientists must trust that the non-epistemic value-decisions their colleagues make are acceptable, this trust is clearly of a moral kind.

The conception of trust needed for expressing these ideas is necessarily a relatively thick one, "trust in the moral sense" (Hardwig 1991, 702), but not unquestioning. Trust in colleagues is expected to be rationally grounded: trust entails epistemic risk, and while scientists deliberately take the risk, they also "resort to a number of measures that can mitigate it" (Wagenknecht 2014b, 85). In other words, it is not the kind of implicit trust we may have in our friends. However, it is necessarily more morally loaded than a thin notion of trust as willingness to rely. While social epistemologists disagree about whether research groups or communities should be seen as agents (see Miller & Freiman 2020; Longino 2022), the different versions of the necessary trust view all use a notion of trust that requires full agency from the trustee.

To summarise, many social epistemologists agree that relationships of trust, understood in a relatively thick sense, are indispensable in science. They are a prerequisite for the kind of collective knowledge production that is ubiquitous in contemporary science. Trust is

necessary when managing relations of epistemic dependence, particularly opaque epistemic dependence, in research groups and scientific communities.

The necessary trust view is largely accepted in the social epistemology of scientific knowledge production, and I find it convincing: trust is the glue that enables collective knowledge production in groups and communities characterised by relationships of opaque epistemic dependence. However, I simultaneously find that it is very difficult to reconcile with the practices of contemporary AI-based science. This is because while relations of opaque epistemic dependence in science have until now existed only between agents capable of taking responsibility for their actions, AI-based science has changed this situation. Scientists have become epistemically dependent, in an opaque manner, on AI applications that are not capable of taking responsibility for their actions. The way in which the role of scientific instruments is generally understood in the literature defending the necessary trust view is not able to accommodate this change.

4. The necessary trust view, epistemic dependence, and instruments

What have the philosophers who argue that relationships of trust are ineliminable in science said about scientific instruments? In fact, often not much.[1] Where relations of opaque epistemic dependence have previously existed in science, they have usually obtained between people. Therefore much of the discussion about trust in research groups and scientific communities has focused on people.

It is, however, possible to sketch a general understanding of how epistemic dependence on instruments is understood in this literature. To summarise: reliance on epistemically opaque instruments is epistemically acceptable in science as long as there is someone who understands how the instruments work, takes responsibility of them in an epistemically acceptable manner, and can be trusted. To better understand this reconstruction of what I take to be the current accepted view, let us distinguish two questions: the question of felt trust and the question of epistemically justified reliance.

In the discussion in epistemology about the notion of trust several philosophers have asked whether an object or an instrument can be a genuine object of felt trust. For instance, C. Thi Nguyen (2022) has recently argued that while most philosophical accounts of trust accept only agents as trustees, another form of trust only involves taking up an unquestioning attitude, and such an attitude can also be held towards objects. For example, a climber can trust a rope. In a similar vein, Ori Freiman and Boaz Miller (2019; Miller & Freiman 2020) suggest that instruments may be subject of "quasi-trust", basing their account on Bruno Latour's idea of nonhuman actants: a person using an instrument can have normative

[1] However, for discussions of the difference between trust in expert testimony and trust in AI applications, see for instance Symons & Alvarado 2019 and Freiman 2023.

expectations about its functioning. However, they agree that such quasi-trust cannot be moral trust in a sense that is central in what I call the necessary trust view. As we can see, here the focus is on the attitude of the trustor: can it differ from mere reliance? Can a person, for instance, feel betrayed if the rope breaks or the instrument malfunctions? Let us call this the question of felt trust. While interesting when trying to understand the notion of trust and human attitudes towards objects, it is not a central question in the necessary trust view.

The necessary trust view includes the demand that trust between scientists must be rationally grounded. As noted, in collective scientific knowledge production the epistemic acceptability of the unavoidable relationships of trust is controlled in various ways: there are different ways of "policing" scientific communities, and established practices meant to ensure that individual researchers or research groups do not behave in epistemically irresponsible ways. For instance, in everyday life we can simply believe the report of an eyewitness because we trust them, but in science, in a similar situation, our trust in our colleague includes our trust in them having followed appropriate, collectively accepted procedures when making the observation and producing the report. Whether an instrument can be a genuine object of felt trust or quasi-trust for an individual is, as a question, distinct from the question under what conditions an individual, as a responsible member of a scientific community, is justified in putting weight on some possible felt trust or quasi-trust when doing research, and relying on the instrument. It is the latter question – the question of epistemically justified reliance – that matters when we talk about the necessary trust view. In their research, a responsible scientist will not rely on an instrument implicitly, without good reasons, and without following collectively accepted procedures (see Record & Miller 2018).

Miller and Freiman (2020) describe as the current "orthodox" view in epistemology that objects, including instruments, cannot be objects of genuine trust. I take this to be the current accepted view also in the literature on the necessary trust view, for two reasons. First, as noted, many of the arguments that have been presented to defend different versions of the necessary trust view use a relatively thick notion of trust that necessitates an agent as the trustee: it does not make sense to say that we trust that the non-epistemic value-decisions made by an instrument are acceptable, as an instrument cannot take responsibility of value-decisions. Secondly, there are normative reasons for questioning trust or quasi-trust in instruments if the attitude is not based on trust in the people who designed and built it. Regardless of whether an instrument can be an object of genuine felt trust or quasi-trust, a responsible scientist must not adopt an "unquestioning attitude" (Nguyen 2022) towards an instrument without good reasons. The scientist should have reasons to believe that the instrument is epistemically reliable, and that if its design has necessitated value-decisions, those decisions have been acceptable. Of the philosophers who have taken part in discussions about the role of trust in science Goldberg (2020b) has explicitly compared ways in which we depend on other people and ways in which we depend on instruments. He argues that these

forms of dependence are different, precisely because an instrument is not an agent or a subject with responsibilities. However, instruments are "the *designed product* of a good deal of epistemic work by others who *are* epistemic subjects in their own right" (Goldberg 2020b, 2785). This view is in line with the accepted view in the epistemological literature on trust: when we cross a bridge, it is not the bridge we trust, but the engineers and builders responsible for it. Similarly, when scientists use instruments that are epistemically opaque to them, they trust the people who designed and built the instruments, and for whom the instruments are not epistemically opaque. These people are capable of taking responsibility of the working of the instruments: we can trust that they have dutifully followed accepted procedures in the design and building of the instrument, and that it is therefore reliable, and we can trust that all possible non-epistemic value-decisions that are reflected in the functioning of the instrument are morally acceptable. In other words, depending on epistemically opaque instruments in science is seen just as a common way in which scientists depend on other scientists. Trust in other agents is the glue that allows collective knowledge production characterised by relationships of opaque epistemic dependence.

By now relations of opaque epistemic dependence in science have existed – or at least they have been taken to exist – between agents who are "*responsible* for the assessments, acquisition, dissemination, and retention of knowledge" (Reider 2016, x). Goldberg (2016) expresses the same basic idea by talking about "epistemic sensibility" and our expectation that epistemic agents act in an epistemically responsible manner: we have normative expectations regarding their actions (see also Collin 2016; Faulkner 2016; Fuller 2016). If we accept the necessary trust view, then only agents who are aware of being answerable to such expectations are taken to be acceptable nodes in the networks of relations of epistemic dependence that are central in contemporary science. But as I argue, this view, the necessary trust view, is built on a contingent amalgamation which has ensured that we are truly epistemically dependent only on agents who are appropriate candidates for trust. This connection between opaque epistemic dependence and candidacy for trust has now unravelled.

5. Essential epistemic opacity

AI applications such as image classifiers or simulations based on machine learning techniques are today used for diverse purposes in scientific knowledge production. They are highly useful, and their use is becoming common in many fields. My claim is that the use of such applications in scientific knowledge production undermines the picture of relationships and networks of trust presented in the previous sections. This is because many AI applications are epistemically opaque.

Paul Humphreys (2009) noted over a decade ago that computer simulations had become epistemically opaque in an essential manner to human agents. He argued that already by then computational processes had become so fast and complex that it was beyond our human cognitive capabilities to understand their details. Thus computational science pushes "humans away from the centre of the epistemological enterprise" (Humphreys 2009, 616). Central for this claim is a distinction Humphreys (2004; 2009) made between ordinary and essential epistemic opacity. A process is epistemically opaque to agent X if they do not know everything that is relevant to it. Such opacity is accidental, and the situation can be remedied: X can learn. With essential epistemic opacity, however, this is not the case: "A process is essentially epistemically opaque to X if and only if it is impossible, given the nature of X, for X to know all of the epistemically relevant elements of the process." (Humphreys 2009, 618.) In other words, a process that is essentially epistemically opaque for X can never become transparent for them. According to Humphreys, the computations involved in simulations used in science were already over a decade ago too fast and complex for any human individual or group of individuals to understand or reproduce. Our cognitive abilities are limited, and the simulations have surpassed our limits.

As noted in the previous section, accidental, remediable epistemic opacity of an instrument used by scientists is not a problem for a social epistemologist arguing that relationships and networks of trust are indispensable in the management of epistemic dependence in contemporary science. There is always someone in the research group or in the larger community who knows all the epistemically relevant elements of how the instrument functions, and insofar as the user of the instrument trusts that person, they can continue using the instrument. Such a situation is just an example of opaque epistemic dependence on other scientists: a scientist does not have the expertise that would enable them to fully grasp how the instrument works, but they trust in their collaborators who do (see Wagenknecht 2014a).

However, if the instrument is essentially epistemically opaque, things change. Humphreys (2009, 691) made comparisons to dependencies discussed in social epistemology, but did not raise the question about relationships of trust. He noted that the situation in which a scientist relies on a simulation that is essentially epistemically opaque resembles one where a group of scientists collaborate and "no one person understands all of the process" – in other words, the members of the research group are epistemically dependent on each other in an opaque manner in the way discussed in section 2 – and added that the sources of the epistemic opacity in the two situations are, however, quite different. But he did not address the issue of trust that highlights the stark differences between the two situations. When AI applications that are essentially epistemically opaque are used in scientific knowledge production, the scientists who use them are epistemically dependent on the instruments, but there is no accountable agent who could fully grasp how the instrument works and therefore be able to take full responsibility of it, and thus no one the scientist could trust, in a rationally grounded

way, to have all the relevant knowledge. Instruments have decisions built into their functioning, and these decisions can involve non-epistemic value-decisions. According to the necessary trust view, if we want to use an epistemically opaque instrument in research, we should be able to trust the person who made these decisions. When we use an epistemically opaque AI application, there is no such person. There is no one who can take full, informed responsibility of value-laden choices that can happen when image classifiers based on machine learning techniques are used in research, and the current rapid progress of large language models may well soon raise questions of plagiarism. Scientists rely on AI applications, as they are highly useful, but this reliance is blind in a way that does not seem to cohere with the necessary trust view. The essential epistemic opacity of AI applications used in science today breaks the networks of trust assumed in social epistemology.

The past decade has only reinforced Humphreys' claim. Many widely used AI applications are not only becoming more important both in society in general[2] and in scientific knowledge production, but they also continue to be epistemically opaque in an essential manner to us human beings. Such AI applications are black boxes not only to their users, but also to their developers.

6. Relying on opaque processes

The essential epistemic opacity of AI applications, or the black box problem, is naturally being discussed in many fields. As Kathleen Creel (2020) notes, scientists who rely on opaque computational systems because of their usefulness can nevertheless find the lack of transparency a serious problem. Davide Castelvecchi (2015) reports that for some scientists the opaqueness of the systems they use is even "a nightmare". Not surprisingly, developers have devised various ways to circumnavigate it. AI applications are tested continually in order to find weaknesses, and interpretability and explainability have become important topics in computer science (Lipton 2018; Beisbart & Räz 2022). When trust or reliance are mentioned or discussed, the focus is typically either on ensuring the reliability of AI applications, or the notion of trust that is used is either undefined or very thin, one that does not require agency from the trustee (Grodizky, Miller & Wolf 2020; see also Hakli and Mäkelä 2019). I will now briefly look into this multifaceted and rapidly growing literature,

[2] It should be noted that AI applications can be opaque in various ways, not all of which follow from them being epistemically opaque in an essential manner. For instance, opacity can result from the need of businesses to maintain their trade secrets. If an AI application is used when deciding whether an applicant is given a bank loan or not, the decision-making process can remain opaque to the applicant simply because the bank will not divulge details about the algorithms they use (Burrell 2016). While this and some other forms of AI opacity are not relevant for the argument of this paper, many of them are societally noteworthy.

arguing that while it has a good grasp of the practices of current AI-based science, the suggestions made in it are typically not compatible with the necessary trust view.[3]

Let us start with an illustrative example. When dealing with epistemically opaque applications, testing them often takes the form of attempts to attack the application. For example, Tom Brown and his colleagues (2018) describe how they have created "universal, robust, targeted adversarial image patches in the real world".[4] In other words, they designed an image that one can print, set somewhere, then take a picture of the scene – and it completely confuses image classifiers. Instead of attempting to transform items in images that are to be classified, Brown and his colleagues generated an image-independent patch that a neural network would find extremely salient:

> We believe that this attack exploits the way image classification tasks are constructed. While images may contain several items, only one target label is considered true, and thus the network must learn to detect the most "salient" item in the frame. The adversarial patch exploits this feature by producing inputs much more salient than objects in the real world. (Brown et al. 2018, 3.)

For a human being, the patch looks like a small, round, somewhat psychedelic jumble of colours. But for an image classifier it apparently looks very much like a toaster. So much so that once the patch is present in an image, no matter what the "lighting conditions, camera angle, type of classifier being attacked, or even the other items within the scene" (Brown et al. 2018, 1), the classifier will deem the image to represent a toaster. In other words, this attack exposes a vulnerability in image classifiers. The reason why the attack is successful is not entirely clear, but now that the vulnerability is known, it can be taken into account in further development.

Clearly the epistemic opacity of AI applications used in science is a concrete problem in the fields where they are used. While they are extremely useful, they do make mistakes.[5] There is something wrong with a process of belief formation that leads to the claim that everything is a toaster. How then can such applications be relied on – not to mention trusting them? A popular answer today is to develop "explainable AI", that is, to either attempt to increase the transparency of the AI applications, or to build simpler, "post hoc" models that retrospectively analyse or try to replicate the results that a black box application has produced. Neither solution is without its problems. Increasing transparency can be interpreted

[3] For a recent, welcome paper that acknowledges the mismatch, see Freiman's (2022) discussion of the "conceptual nonsense 'trustworthy AI'".
[4] I am grateful to Anna-Mari Rusanen for suggesting this example.
[5] Here it is important to distinguish between errors originating in human error or bias, and errors originating in the functioning of an epistemically opaque AI application. Some of the errors that AI applications based on machine learning techniques make are caused by biased data (see e.g. Caliscan et al. 2017), but not all. For instance, the vulnerability Brown et al. (2018) identified has nothing to do with biased data. (See Rusanen & Koskinen 2018.)

in many ways, but if it is taken to mean reducing the opacity of AI applications, it risks reducing their usefulness. The post-hoc models, on the other hand, face a problem of reliability: if the original AI application remains opaque, it cannot be guaranteed that an explanation produced with a post-hoc model accurately represents how the black box model actually reached a conclusion. However, developments in the field of explainable AI are increasing the number of ways in which scientists can test the reliability of the epistemically opaque applications they use, and to some degree increase their understanding of the applications. (Lipton 2018; Zerilli et al. 2019; Creel 2020; Fleisher 2022; Beisbart & Räz 2022.)

The focus on reliability, however, means that it is difficult to find or make meaningful connections between the discussions in computer science and the necessary trust view. The same is the case in much of the philosophical work on the problems epistemic opacity creates in science: the literature focuses on reliance, not trust in a thick sense. Several philosophers who have addressed the issue hold that while the epistemic opacity of many AI applications is unavoidable, reliance on them can nevertheless be justified (e.g. Humphreys 2013; Duran 2017). We can identify unreliable processes: testing, verification and validation allow for reliance on opaque applications. I will now examine more in detail one argument about the reliability of epistemically opaque computational processes. It captures the current practices well, but is incompatible with the necessary trust view.

Juan M. Durán and Nico Formanek (2018) have argued that we can rely on epistemically opaque computational processes because they consistently produce reliable results. Starting from process reliabilism as defined by Goldman (1979; Goldman and Beddor 2016), they address the problem of epistemic opacity and develop a stance they call "computational reliabilism". They argue that scientists are justified in relying on epistemically opaque simulations as long as it can be shown that the simulations consistently produce reliable results. Instead of focusing on the epistemically opaque processes within the simulations, they emphasise the importance of thorough verification and validation – while a simulation is epistemically opaque, the procedures used when verifying its results are not: "Thus understood, computational reliabilism requires a ‘retrospective reliability chain,’ one that conditions the sources that attribute reliability to computer simulations to be reliable in and by themselves." (Durán and Formanek 2018, 656.) They then identify verification and validation methods, robustness analysis, histories of (un)successful implementations, and expert knowledge as sources for attributing reliability to computational processes.

Durán and Formanek undoubtedly describe well the ways in which epistemically opaque AI applications are currently used in science, as well as the reasonings of many of the scientists who use them. But their account is not compatible with the necessary trust view, as they fail to differentiate between trust and reliance, and do not address the issue of agency or epistemic responsibility at all. The necessary trust view states that trust in a thick sense is

needed for the management of opaque epistemic dependence in collaborative knowledge production. While Durán and Formanek suggest various retrospective and indirect ways to attribute reliability to epistemically opaque computational processes, these processes remain epistemically opaque, and scientists remain epistemically dependent on them, while they are not agents who could be trusted. Thus, from the social epistemological viewpoint sketched in the previous sections, the account Durán and Formanek present is not a satisfactory basis for a social epistemology of AI-based science.

As I have failed to find in the philosophical work on AI an account that would try to reconcile the current practices with the necessary trust view, I will now return to philosophical accounts of the epistemic relationships between scientists and their instruments. As noted, there is relatively little discussion about these relationships in the literature defending different versions of the necessary trust view. In the next section, I will therefore focus on arguments made in the philosophical literature on distributed cognitive systems and extended agency.

7. Quasi-agents or extended agents

Philosophers of science and epistemologists have paid abundant attention to the role of things, such as instruments and machines, in knowledge production. We human beings are, after all, able to tackle cognitive tasks with the aid of a pen and paper much better than without. With a microscope we can observe things we could not see without one, and with a computer we are capable of carrying out feats that would otherwise be quite impossible for us. These facts have led to discussions about extended knowledge and distributed cognitive systems: knowledge-conducive cognitive abilities being extended to artifacts people use in cognitive tasks and knowledge production (Clark and Chalmers 1998; Carter et al. 2018).

It is not clear whether it follows that epistemic agency too should be conceived as being sometimes extended beyond the boundaries of a human being. Do the boundaries of an epistemic agent necessarily correspond with the boundaries of an individual? Can a researcher and their instrument sometimes form a single, extended epistemic agent? One's answer to this question seems to depend on the aspect of epistemic agency one finds central: the cognitive processes without which the agent could not know, or the responsibility we assign to agents. I will now examine three notable takes on the question, asking whether they can reconcile the necessary trust view with AI-based science. Ronald Giere (2004; 2007; 2012) argued that while cognitive systems are distributed, agency is not. Orestis Palermos and Duncan Pritchard (2016; Pritchard 2013) are more sympathetic to the idea of extended agents consisting of humans and instruments, but they require that the human member of such an agent must be able to monitor the the working of the instruments. Andy Clark (2015; Clark and Chalmers 1998) notes that this requirement is not compatible with the basic idea of

extended knowledge, and argues that the human part of an extended agent does not in fact need to be able to continually monitor the processes going on in the non-human parts. The views defended by Giere and by Palermos and Pritchard are both compatible with the necessary trust view, but not with the current use in science of AI applications that are epistemically opaque in an essential manner. Clark's view, on the other hand, could perhaps be reconciled with the practices of current in AI-based science, but if this is done, it is no longer compatible with the necessary trust view.

Distributed cognitive systems without extended agency

Giere (2004; see also 2007, 2012) argued that while much of scientific knowledge production happens in distributed cognitive systems that include both humans and various instruments, such systems are not agents. For instance, the Hubble space telescope is a necessary component of a complex cognitive system that incorporates both many other instruments and many people. It is the whole "part physical-causal, part computational and part social-cultural" (Giere 2004, 714) hybrid system that performs cognitive tasks and produces new knowledge. The distribution of cognition, however, does not mean that distributed cognitive systems should be regarded as cognitive or epistemic agents. Giere pointed out that the concept of agency cannot be separated from related concepts such as that intention, responsibility, and consciousness. In short, agents are those we can praise or blame; those who take responsibility (or should take it) when something goes wrong:

> We are generally very good at assigning responsibility, and thus praise or blame, for peoples' actions, including their epistemic actions. Thus, in the operation of a complex distributed cognitive system such as the HST, we can often enough reliably determine who is or is not performing well at their assigned tasks, including the task of drawing theoretical conclusions from the final images, just as we can often enough reliably determine whether a piece of equipment is functioning correctly. That seems sufficient to distinguish the epistemic agency of humans from the causal agency of other parts of the system. (Giere 2004, 717.)

Giere's view is easy to reconcile with the necessary trust view. While the whole system is needed for it to be possible to perform certain tasks, there is always someone whose responsibility it is to take care of the functioning of the instruments and to understand them; someone whose job it is to make sure that a piece of equipment is functioning correctly – someone who can be trusted.

Simultaneously, Giere's account of agency in distributed cognitive systems does not quite fit in with current AI-based science. If an AI application is epistemically opaque in an essential manner even to its developers, there is no one responsible for fully understanding how it works. If we insist that agency in such a situation is not extended, it seems that the

epistemically opaque AI application – say, an image classifier – has been granted the role of some kind of an epistemic quasi-agent in scientific knowledge production. We accept the results it produces, even though we cannot know how it produces them. Scientists rely on such applications, but this reliance, as well as the epistemic dependency that follows from it, is blind in a way that resembles the trust we can have in epistemic agents who are able to take responsibility for their actions – which is something AI applications cannot do.

Extended epistemic agency with monitoring

Palermos and Pritchard (2016; see also Pritchard 2010) have argued that an agent "may extend his cognitive character by incorporating epistemic artifacts to it" (Palermos and Pritchard 2016, 117). While they, unlike Giere, claim that the hypothesis of extended cognition can "open up the possibility of extended epistemic agents" (*ibid.*, 118), they nevertheless retain the demand for responsibility. They argue that cognitive processing can be extended beyond a human individual, thus incorporating artifacts. Nevertheless they demand that cognitive success must be a manifestation of cognitive ability. This means that the human component of an extended agent must be able to monitor the knowledge-producing process – not necessarily all the time, but to a certain extent: the human must be "in a position to be responsive were there something wrong" (*ibid*., 115) with the process.

This account of agency and the epistemic relationship between scientists and their instruments is also easy to reconcile with the necessary trust view. Palermos and Pritchard explicitly demand that the human component at least passively monitors the processes going on in the non-human components. But to be able to do this, the human agent must be in a position where they can understand the functioning of the whole. A scientist using an image classifier that is epistemically opaque in an essential manner is not in such a position. If the classifier wrongly identifies a seagull as a toaster, the scientists will likely notice the error, but if it does something equally problematic, but less easily noticeable, the scientist is not in a position to appropriately monitor the knowledge-producing process and to recognise when something goes wrong. The account Palermos and Pritchard present cannot cover AI-based science as it is today.

Extended epistemic agency without monitoring

Both in Giere's account, and in the one by Palermos and Pritchard, human agents (or agents that are moral similarly to human ones) are required to take epistemic responsibility for the functioning of the whole, and this presupposes that they can monitor all of relevant the processes. This makes the accounts both consistent with the necessary trust view and incompatible with the current practices of AI-based science. Clark (2015) seems to offer us an alternative view. Building on Clark and Chalmers's (1998) analysis of extended cognition,

he argues that epistemic agency can be extended, and contra Palermos and Pritchard, he claims that this "does not require any kind of conscious or personal-level engagement between the agent and the cognitive process on the part of the agent at all" (Clark 2015, 3766).

Clark notes that the demand for even minimal monitoring in fact works against the idea of extended mind. He wishes to defend the view that an instrument can become part of a mind. In order for it to really be part of a mind, it should be treated as any other part of the mind. As we do not usually require the background monitoring of our minds, the situation should not change when the mind is extended. For instance, "our biological memory is not typically subject to agentive scrutiny as a process at all, much less as one that may or may not be reasonably judged to be reliable by the agent" (Clark 2015, 3763). If an instrument is treated differently, it does not look like a part of the mind, but like a piece of external equipment.

To use the terminology of this paper, Clark points out that an important part of the cognitive processes that go on in our heads are epistemically opaque to us. This opaqueness may be partly remediable, but usually we do not require that such processes be made transparent. This seems true: we rely on our memories and even the memories of other people without fully understanding how human memory works. So could we argue that if a human and an epistemically opaque AI application form an epistemic agent, the situation does not need to change: the internal processes of the epistemic agent can remain epistemically opaque even to the agent themself? Why should we treat the opacity of the AI part of such an agent in a different way we treat the opacity of the human part (see Zerilli et al. 2019; Kawamleh 2022).

This could perhaps be compatible with the practices current in AI-based science. It is, however, not quite compatible with the necessary trust view.

First, this solution requires stretching or renouncing some of the conditions Clark sets for the responsible practices and the formation of extended epistemic agents, as well as the conditions Clark and Chalmers give for an object to be a constitutive part of a cognitive process.

As Clark notes, the conscious awareness of our thought processes that Pritchard and Palermos (2016; see also Pritchard 2013) emphasise, becomes important when we start to systematically transmit knowledge to others: "we don’t (or ought not) simply teach facts. Rather, we install methods and practices that help students probe and test their beliefs and knowledge sources, and deepen their understandings" (Clark 2015, 3772). It is not entirely clear that an extended agent consisting of a human and an epistemically opaque AI application could do this.

And, as Clark continues, even in everyday situations our first encounters with new instruments usually differ from the kind of "fluid unreflective use" (*ibid.* 3773) that is typical when the instrument has become part of our extended cognitive architecture. When we

approach new tools and technologies, we usually exercise "agentive epistemic care" when learning to use them, and we certainly exercise "due epistemic caution" (*ibid*. 3774) when developing new technologies. In science, the criteria of due epistemic caution can be heightened, as the developer is responsible to the whole scientific community. It is not entirely clear that epistemically opaque AI applications can become part of a scientist's extended cognitive architecture through such a process, because she is in no point in a position where she or anyone she trusts could fully understand the workings of the application she uses.

It may indeed be that when scientists are well acquainted with an instrument, it becomes a part of their extended cognitive architecture, and they use it without monitoring it. However, when the instrument is epistemically transparent, they can allow this to happen because they either have become familiar first with how the instrument works, or trust its developers. In such a case their behaviour is compatible with the necessary trust view: they can use the instrument in a fluid and unreflective way because if needed, they can either take a step back and make the process transparent to their colleagues, or they trust the developers of the instrument to be able to do this. But if the instrument is an epistemically opaque AI application, a scientist cannot take a step back and make the process transparent. In other words, an extended agent consisting of a scientist and an epistemically opaque AI application is qualitatively different from an extended agent consisting of a scientist and an epistemically transparent instrument.

Moreover, Clark and Chalmers (1998) famously give some criteria – the "trust and glue" conditions – that an object must meet before it can be included into an individual's cognitive system. (The notion of trust used here is a thin one, and basically amounts to willingness to rely.) One of the conditions is the following: "Any information retrieved or gained via it should be more-or-less automatically endorsed. It should not usually be subject to critical scrutiny." (Clark 2010, 64.) Even a brief glance on the current discussion about explainable AI makes it clear that computer scientists do not automatically endorse any information gained by using epistemically opaque AI applications. Therefore, such applications do not seem to meet the trust and glue conditions.

There are at least two more reasons why the idea of extended agents consisting of humans and AI applications is hard to reconcile with the necessary trust view. One is that the ways in which scientists rely on their own, epistemically opaque minds in scientific knowledge production differ from the ways in which scientists rely on epistemically opaque AI applications.

In scientific communities, we require an agentive, responsible approach both to our own thought processes and to the technologies we use. As noted in sections 2 and 4, processes that in everyday life and on the individual level can remain unnoticed, are in science made as transparent as possible. While in everyday situations we do not question the reliability of our

memory, historians will not unquestioningly accept recollections as evidence, as they are all too aware of the many ways in which our memory fails us. We trust human agents in science even though a part of what goes on in their heads remains epistemically opaque to them, and cannot be checked even retrospectively. We can do so because we have collectively established methods for preventing or detecting typical human errors and biases, and we trust that the human agent will be responsible and use those methods – and, in the end, this is the best we can do. Perhaps we can similarly trust extended agents if the same can be said of them? As the image classifier case illustrates, AI applications are not infallible; rather, attempts to verify their results and test their robustness have repeatedly shown that they can be fooled (e.g. Nguyen, Yosinski and Clune 2015). But if we know the ways in which they are fallible, and have established methods for mitigating the ensuing epistemic risks, and we trust that the extended epistemic agent uses such methods, then perhaps our trust in the agent does not need to falter? Perhaps the use of verification and validation methods, robustness analysis, explanatory post-hoc models, and other similar ways to ensure the reliability of the AI applications discussed in section 6, could be compared to the procedures scientific communities have developed to avoid many known failings of our epistemically opaque human minds?

John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan (2019) have argued that demands of transparency in automated decision-making are excessively high, as human decision-makers cannot reach the level of transparency required of the AI applications. Can a similar argument be made in the context of AI-based science? I believe not, because in science, the opaque processes of human minds are not typically accepted as justifications for claims, whereas in the kind of decision-making Zerilli and his colleagues discuss, the opaque processes of human minds can be the basis for generally accepted decisions. While various epistemically opaque processes happening in human minds influence scientific knowledge production in many ways, the justification of a claim is independent of the actual cognitive processes that led to it; it does not need to resemble the original cognitive process. The justification has to be public, as it has to be scrutinisable. But if a claim is the result of an epistemically opaque process within an AI application, there may be no way to produce an independent, public justification – a more or less accurate, post-hoc analysis of some of the central features of the opaque process may well be all we can have.[6] If such claims are accepted and used in science, the processes happening within epistemically opaque AI

[6] Eamon Duede (2023) argues that certain uses of epistemically opaque AI applications in science are epistemically unproblematic: if the application is used in the context of discovery, in order to come up with interesting hypotheses, and justification happens separately, without the use of epistemically opaque AI applications, there is no reason to be worried about the opacity. I agree; the questions and worries I discuss here pertain to the use of epistemically opaque AI applications in the context of justification.

applications are treated differently than processes happening in the epistemically opaque minds of human beings.

Finally, if we think of a human using an epistemically opaque AI application as an extended agent, we must accept that the relative proportion of the cognitive processes that are opaque to the agent themself is increased. We have not even started to think about whether such a change can be reconciled with the necessary trust view.

What does all this mean for the necessary trust view? Is such an extended agent trustworthy? Can we, for instance, reasonably trust that the value-decisions the agent makes in the different stages of a research process are acceptable (Wilholt 2013)? This is by no means obvious. Introducing extended agents consisting of humans and epistemically opaque AI applications to science significantly alters the role of machines in science in ways that social epistemologists have yet to address.

8. Conclusions

In the social epistemology of scientific knowledge it is generally accepted that relationships of trust are necessary in science, that such relationships are possible only between agents who can be held accountable for their actions, and that the trust must be collectively controlled and rationally grounded. The notion of trust used in this view is a thick one: it has a moral component and requires full agency from the trustee. Trust is taken to be necessary, as it is indispensable in the management of epistemic dependence in research groups and scientific communities. To some degree, scientists have to trust the moral character of their colleagues; otherwise research groups and scientific communities could not function. I named this view the *necessary trust view*.

However, this generally accepted view presupposes that we are epistemically dependent, at least in an opaque manner, only on agents who can take responsibility and who can be trusted. This is no longer the case; scientists are now epistemically dependent on epistemically opaque AI applications. In contemporary science, AI applications such as simulations or image classifiers based on machine learning techniques are used in knowledge production. Many of such applications are epistemically opaque in an essential manner (Humphries 2009): it is impossible for humans to follow and fully understand how they reach their results. As I argue, the essential epistemic opacity of such applications creates a problem for the social epistemology of scientific knowledge. It is difficult to reconcile the use of such applications with the necessary trust view, as is it not clear that any agent that we could reasonably deem trustworthy could take responsibility for the workings of the epistemically opaque application. At the same time, the arguments for the necessary trust view are so strong that no account that completely disregards them is a promising basis for a satisfactory social epistemology of AI-based science.

When looking for potential starting points for a satisfactory social epistemology of AI-based science, I have focused on three different accounts of agency that note the complicated relationship that scientists have with their instruments (Giere 2004; Clark 2015; Palermos and Pritchard 2016). None of the accounts is easily compatible with both the necessary trust view and the current practices in AI-based science.

If we argue that a scientist and an AI application can form an extended epistemic agent, and that it is epistemically acceptable in science to trust such an agent, we have to accept that this happens without an initial stage in which the human would be able to fully exercise "agentive epistemic care" (Clark 2015, 3774) when learning to use the application, and without the human being able to trust that at least the developer of the application has understood the processes going on in it. We need to accept that the full justification of a claim produced by such an agent can remain epistemically opaque both to the agent themselves and to the research community. Finally, we must accept that a significant part – that is, clearly larger than in the case of human agents – of the processes of belief formation within the agent are epistemically opaque in an essential manner both to the agent and their colleagues. All this at least raises the question to what degree and under what conditions we should trust such agents in science. Without much further work this does not seem a firm basis for a social epistemology of AI-based science. And finally, it is not clear that in all cases where scientists use epistemically opaque AI applications we can really observe the kind of "fluid incorporation of those tools and technologies deep into our cognitive repertoires" that Clark (2015, 3774) highlights when arguing for extended knowledge and extended agency, as the "trust and glue" conditions are not met.

The other option is to agree that when scientists use applications that are epistemically opaque in an essential manner, agency is not extended to the applications. As no agent can take full, informed responsibility for them in the way in which scientists can take responsibility of traditional instruments, they become what I have called epistemic quasi-agents in research groups and scientific communities: scientists are epistemically dependent on them in an opaque manner in which scientists have previously depended only on human agents. Such quasi-agents, however, are not full agents capable of taking responsibility for their workings. The moral aspect of epistemic agency that makes trust possible is missing. This breaks the idea of relationships and networks of trust that is central in the necessary trust view.

In other words, we currently have no satisfactory social epistemology of AI-based science.

Acknowledgements

This paper grew out from a still unfinished manuscript on which Anna-Mari Rusanen and I worked together some years ago; its main point is mentioned in footnote 5. I am very grateful for Anna-Mari's generosity and expertise; this paper would not exist without her encouragement.

When presenting our joint paper – the one that remains unfinished – in the SAS workshop 'Epistemic opacity in computer simulation and machine learning' in Stuttgart in 2018, I was lucky enough to meet the late Paul Humphreys. I am grateful for his insightful comments, as well as for numerous illuminating discussions with the other participants of the workshop and with its organisers, Michael Herrmann and Andreas Kaminski.

I have presented various iterations of the paper at hand at TINT Centre for Philosophy of Social Sciences' PoS seminar in 2020, at the meeting of the project 'Social and cognitive diversity in science' in 2022, in the workshop 'AI in research' in Helsinki in 2023, and in the Values in Science reading group in Cambridge in 2023. I would like to thank the organisers, audiences, and participants of these meetings, as well as several other people who have kindly read the manuscript, particularly Carl Craver, Raul Hakli, Stephen John, Phillip Hintikka Kieval, Jaakko Kuorikoski, Kaisa Luoma, Pekka Mäkelä, Cristian Larroulet Philippi, Samuli Reijula, Kristina Rolin, Paul Teller, and Eric Winsberg for useful questions and comments. I would also like to thank Arttu Kataja for our numerous discussions about AI over several years – without them this paper would not exist.

Finally I would like to thank three anonymous reviewers for their thorough work – the paper is much better thanks to their comments. All remaining errors are my own.

This work was supported by the Academy of Finland under Grants 316695 and 349051.

Literature

Baier, A. C. (1986). Trust and Antitrust. *Ethics*, 96*,* 231–260.

Baker, J. (1987). Trust and Rationality. *Pacific Philosophical Quarterly*, 68, 1–13.

Beisbart, C.; and Räz, T. (2022). Philosophy of science at sea: Clarifying the interpretability of machine learning. *Philosophy Compass*, 17(6), e12830.

Brown, T.; Dandelion, M.; Aurko, R.; Abadi, M.; and Gilmer, J. (2018). Adversial Patch. arxiv.org/abs/1712.09665v2.

Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1). DOI: 10.1177/2053951715622512.

Caliskan, A.; Bryson, J.; and Narayanan, A. (2017). Semantics Derived Automatically from Language Corpora Contain Humanlike Biases. *Science*, 356(6334), 183–186.

Carter, J. A.; Clark, A.; Kallestrup, J.; Palermos, S. O.; and Pritchard, D. (2018). Extended Epistemology: An Introduction. In Carter et al. (eds.), *Extended Epistemology*. Oxford University Press, 1–16.

Castelvecchi, D. (2015). Artificial Intelligence Called in to Tackle LHC Data Deluge. *Nature News*, 528(7580), 18–19.

Clark, A. and Chalmers, D. (1998). The Extended Mind. *Analysis*, 58(1), 7–19.

Clark, A. (2010). Memento's revenge: The extended mind, extended. In Richard Menary (ed.), *The Extended Mind*. MIT Press, 43–66.

Clark, A. (2015). What 'Extended Me' Knows. *Synthese*, 192, 3757–3775.

Collin, F. (2016). Two Kinds of Social Epistemology and the Foundations of Epistemic Agency. In P. J. Reider (ed.), *Social Epistemology and Epistemic Agency: Decentralizing Epistemic Agency*. Rowman & Littlefield, 43–59.

Daston, L. and Galison, P. (2007). *Objectivity*. Zone Books.

Douglas, H. (2000). Inductive Risk and Values in Science. *Philosophy of Science*, 67, 559–79.

Durán, J. M. (2017). Varying the Explanatory Span: Scientific Explanation for Computer Simulations. *International Studies in the Philosophy of Science*, 31(1), 27–45.

Durán, J. M. and Formanek, N. (2018). Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism. *Minds and Machines*, 28, 645–666.

Duede, E. (2023). Deep Learning Opacity in Scientific Discovery. *Philosophy of Science,* 1–13. DOI: 10.1017/psa.2023.8.

Faulkner, P. (2016). Agency and Disagreement. In P. J. Reider (ed), *Social Epistemology and Epistemic Agency: Decentralizing Epistemic Agency*. Rowman & Littlefield, 75–90.

Fleisher, W. (2022). Understanding, Idealization, and Explainable AI. *Episteme,* 1–27. DOI: 10.1017/epi.2022.39.

Freiman, O. and Miller, B. (2019). Can Artificial Entities Assert? In S. Goldberg (ed.), *The Oxford Handbook of Assertion*. Oxford University Press, 414–434.

Freiman, O. (2022). Making sense of the conceptual nonsense 'trustworthy AI'. *AI Ethics*. DOI: 10.1007/s43681-022-00241-w.

Freiman, O. (2023). Analysis of Beliefs Acquired from a Conversational AI: Instruments-based Beliefs, Testimony-based Beliefs, and Technology-based Beliefs. *Episteme,* 1–17. DOI: 10.1017/epi.2023.12.

Frost-Arnold, K. (2013) Moral Trust and Scientific Collaboration. *Studies in History and Philosophy of Science*, 44, 301–310.

Fuller, S. (2016). A Sense of Epistemic Agency Fit for Social Epistemology. In P. J. Reider (ed.), *Social Epistemology and Epistemic Agency: Decentralizing Epistemic Agency*. Rowman & Littlefield, 21–39.

Giere, R. (2004). The Role of Agency in Distributed Cognitive Systems. *Philosophy of Science*, 73(5), 710–719.

Giere, R. (2007). Distributed Cognition without Distributed Knowing. *Social Epistemology*, 21(3), 313–320.

Giere, R. (2012). Scientific Cognition: Human Centered but not Human Bound. *Philosophical Explorations*, 15(2), 199–206.

Goldberg, S. (2011). The Division of Epistemic Labour. *Episteme*, 8(1), 112–25.

Goldberg, S. (2016). A Proposed Research Program for Social Epistemology. In P. J. Reider (ed.), *Social Epistemology and Epistemic Agency: Decentralizing Epistemic Agency*. Rowman & Littlefield, 3–19.

Goldberg, S. (2020a). Trust and Reliance. In J. Simon (ed.), *The Routledge Handbook of Trust and Philosophy*. Routledge, 97–108.

Goldberg, S. (2020b). Epistemically Engineered Environments. *Synthese*, 197, 2783–2802.

Goldman, A. I. (1979). *Justification and Knowledge*. Springer.

Goldman, A. and Beddor, B. (2016). Reliabilist Epistemology. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Winter 2016 edition. plato.stanford.edu/archives/win2016/entries/reliabilism.

Grodizky, F.; Miller, K.: and Wolf, M. J. (2020) Trust in Artificial Agents. In J. Simon (ed.), *The Routledge Handbook of Trust and Philosophy*. Routledge, 298–312.

Hakli, R. A. and Mäkelä, P. A. (2019). Moral Responsibility of Robots and Hybrid Agents. *The Monist*, 102(2), 259–275.

Hardin, R. (1991). Trusting Persons, Trusting Institutions. In R. J. Zeckhauser (ed.), *Strategy and Choice*. MIT Press, 185–209.

Hardwig, J. (1985) Epistemic Dependence. *The Journal of Philosophy*, 82(7), 335–349.

Hardwig, J. (1991). The Role of Trust in Knowledge. *The Journal of Philosophy*, 88(12), 693–708.

Humphreys, P. (2004). *Extending Ourselves: Computational Science*, *Empiricism, and Scientific Method*. Oxford University Press.

Humphreys, P. W. (2009). The Philosophical Novelty of Computer Simulation Methods. *Synthese*, 169(3), 615–626.

Humphreys, P. W. (2013). What are Data About? In J. M. Durán and E. Arnold (eds.), *Computer Simulations and the Changing Face of Scientific Experimentation*. Cambridge Scholars Publishing, 12–28.

Jones, K. (2012). Trustworthiness. *Ethics*, 123(1), 61–85.

Kawamleh, S. (2022). Against explainability requirements for ethical artificial intelligence in health care. *AI Ethics*. DOI: 10.1007/s43681-022-00212-1.

Lipton, Z. C. (2018). The Mythos of Model Interpretability. *Queue*, 16(3), 31–57.

Longino, H. E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Longino, H. E. (2022) What's Social about Social Epistemology? *The Journal of Philosophy*, 119(4), 169–195.

Miller, B. and Freiman, O. (2020). Trust and Distributed Epistemic Labor. In J. Simon (ed.), *The Routledge Handbook of Trust and Philosophy*. Routledge, 341–353.

Nguyen, A.; Yosinski, J.; and Clune, J. (2015). Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. 2015 version, arxiv.org/abs/1412.1897.

Nguyen, C. T. (2022). Trust as an Unquestioning Attitude. In T. S. Gendler, J. Hawthorne, and J. Chung (eds), *Oxford Studies in Epistemology*, vol. 7, Oxford University Press, 214–244.

Nickel, P.J. (2013). Trust in Technological Systems. In: de Vries, M., Hansson, S., Meijers, A. (eds) *Norms in Technology. Philosophy of Engineering and Technology*, vol. 9. Springer, 223–238.

Origgi, G. (2018). *Reputation: What It Is and Why It Matters*. Princeton University Press.

Palermos, O. and Pritchard, D. (2016). The Distribution of Epistemic Agency. In P. J. Reider (ed.), *Social Epistemology and Epistemic Agency: Decentralizing Epistemic Agency*. Rowman & Littlefield, 109–126.

Pettit, P. (1995). The Cunning of Trust. *Philosophy & Public Affairs*, 24(3), 202–225.

Pritchard, D. (2010). Cognitive Ability and the Extended Cognition Thesis. *Synthese* 175(1), 133–151.

Record, I., and Miller, B. (2018). Taking iPhone Seriously: Epistemic Technologies and the Extended Mind. Jn J. A. Carter et al. (eds), *Extended Epistemology*. Oxford University Press, 105–126.

Reider, P. J. (2016). Introduction: What is Social Epistemology and Epistemic Agency? In P. J. Reider (ed.), *Social Epistemology and Epistemic Agency: Decentralizing Epistemic Agency*. Rowman & Littlefield, vii–xvi.

Rolin, K. (2020). Trust in science. In J. Simon (ed.), *The Routledge Handbook of Trust and Philosophy*. Routledge, 354–366.

Rusanen, A.-M. and Koskinen, I. (2018). Tiede, tekoäly ja tiedolliset riskit. *Futura,* 37(4), 47–52.

Symons, J. and Alvarado, R. (2019). Epistemic Entitlements and the Practice of Computer Simulation. *Minds & Machines*, 29, 37–60.

Wagenknecht, S. (2014a). Opaque and Translucent Epistemic Dependence in Collaborative Scientific Practice. *Episteme*, 11(4), 475– 492.

Wagenknecht, S. (2014b). Four Asymmetries Between Moral and Epistemic Trustworthiness. *Social Epistemology Review and Reply Collective*, 3(6), 82–86.

Wagenknecht, S. (2015). Facing the Incompleteness of Epistemic Trust: Managing Dependence in Scientific Practice. *Social Epistemology*, 29(2), 160–184.

Wilholt, T. (2013) Epistemic Trust in Science. *The British Journal for the Philosophy of Science*, 64(2), 233–253.

Zerilli, J., Knott, A., Maclaurin, J., and Gavaghan, C. (2019). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy & Technology*, 32, 661–683.