Synthese (2022) 200:491
https://doi.org/10.1007/s11229-022-03975-6

ORIGINAL RESEARCH

Instruments, agents, and artificial intelligence: novel epistemic categories of reliability

Eamon Duede[1,2,3,4]

Received: 17 October 2021 / Accepted: 7 November 2022 / Published online: 19 November 2022

Abstract

Deep learning (DL) has become increasingly central to science, primarily due to its capacity to quickly, efficiently, and accurately predict and classify phenomena of scientific interest. This paper seeks to understand the principles that underwrite scientists' epistemic entitlement to rely on DL in the first place and argues that these principles are philosophically novel. The question of this paper is not whether scientists can be justified in trusting in the reliability of DL. While today's artificial intelligence exhibits characteristics common to both scientific instruments and scientific experts, this paper argues that the familiar epistemic categories that justify belief in the reliability of instruments and experts are distinct, and that belief in the reliability of DL cannot be reduced to either. Understanding what can justify belief in AI reliability represents an occasion and opportunity for exciting, new philosophy of science.

Keywords Deep learning · Scientific knowledge · Models · Reliability · Trust and Justification

T.C. : Philosophy of Science in Light of Artificial Intelligence.

This paper is forthcoming in a special issue of *Synthese* titled "Philosophy of Science in Light of Artificial Intelligence".

✉ Eamon Duede
eduede@uchicago.edu

1 Department of Philosophy, University of Chicago, Chicago, USA

2 Committee on Conceptual & Historical Studies of Science, University of Chicago, Chicago, USA

3 Pritzker School of Molecular Engineering, University of Chicago, Chicago, USA

4 Knowledge Lab, University of Chicago, Chicago, USA

1 Introduction

While contemporary deep learning (DL) systems display superhuman capacities for prediction and classification, their opacity (Creel, 2020; Zerilli, 2022) is thought to limit our ability to understand why such systems make the predictions and classifications they do. This limitation is of particular concern in high-stakes decision-making settings such as medical diagnosis and criminal justice, where accountability, value-alignment, and a wide range of ethical considerations are salient (Birch & Creel, 2022; Falco & Shneiderman, 2021; Hoffman, 2017; Rudin, 2019). Yet, the inscrutability of DL models (DLMs) is also of epistemological concern in scientific settings, where explanations and understanding (Räz & Beisbart, 2022; Sullivan, 2019) represent central epistemic virtues (Khalifa, 2017). Of course, in contexts where deep learning systems are used to emulate well-understood but time-consuming tasks (e.g., identifying galaxies or denoising data), purely pragmatic considerations such as accuracy of classification or degree of control are often sufficient or adequate for a given model's purpose (Parker, 2020). Moreover, in many such cases, DL outputs can be independently verified, thereby rendering opacity epistemically irrelevant (Duede, 2022).

Nevertheless, in many contexts, scientists treat DL outputs themselves as claims about the target systems upon which the models were trained. Here, it is reasonable to ask what justifies belief in the reliability of those models. We might feel that, in such settings, we need explanations for and understanding of the underlying network processes by which the outputs (claims) are arrived at to serve, in part, as justification for our belief in them (Creel, 2020). Developing methods for extracting explanations from deep neural networks to bolster our confidence in their reliability is, of course, the goal of much of the work in the growing field of explainable AI (XAI). Importantly, however, progress here has, to date, been quite limited (Ghorbani et al., 2019; Lipton, 2018; Rudin, 2019).

While recent work in the philosophy of science has sought to make sense of how attempts at and need for interpretability fit into a broader nexus of practices surrounding the use of AI in science (Boge, 2021; Buckner, 2019; Duede, 2022; Räz, 2022; Sullivan, 2019; Zerilli, 2022)[1], this paper seeks to understand the principles that underwrite scientists' epistemic entitlement to rely on AI in the first place and argues that these principles are philosophically novel. The question is not whether scientists can be justified in trusting in the reliability of DLMs. I take it that, in principle, they can and that the recent scientific literature provides good evidence to that effect. Instead, the central question of this paper is whether, in general, the epistemic basis for believing in the reliability of DLMs is something philosophically familiar or, rather, something philosophically novel. When scientists use DLMs to solve complex scientific problems that have otherwise been intractable (Senior et al., 2020), bypass costly computations (Wang et al., 2019), and outperform human experts (themselves included) on routine but complex tasks (Chen et al., 2014), what justifies their belief in the reliability of their

[1] Philosophers have also revived interest in what we can learn about cognition from deep learning models. For instance, Buckner (2018) argues that the evaluation of the behavior of deep convolutional neural networks helps us resolve questions going back to Locke concerned with human abilities for abstraction. Others, however, have expressed skepticism about the legitimacy of looking to neural nets as plausible models of human cognition at all (Stinson, 2020).

DLMs? Three philosophically familiar justifications present themselves as plausible answers, though, as I will show, none can, in fact, be made to ground epistemic reliance in deep learning models.

The first seemingly straightforward reason one might have for believing in the reliability of DLMs is through an evaluation of their reliability in the past. Here, the idea is that a scientist can be justified in believing in the reliability of DL based solely on what I will call *brute inductive considerations*. On brute inductive considerations alone, one could attempt to ground the basis for epistemic reliance on DL either on the past reliability of DL in general (e.g., on DL's track record in science) or on the past reliability of specific deep learning models (e.g., on this or that model's performance on out-of-sample data). I explore this justification in Sect. 2, and argue that neither general nor specific brute inductive considerations can justify belief in the reliability of DLMs.

It is not uncommon, however, for scientists to refer to and use DL in ways that are suggestive of other familiar objects of epistemic reliance that do not depend solely on brute inductive considerations to justify belief in their reliability. Many examples from the scientific literature in which DL plays a central role look like paradigmatic cases in which scientists use a *scientific instrument* to learn something they did not previously know. This suggests a justification for epistemic dependence on AI's reliability that is generally reducible to that of scientific instruments. When, for instance, a scientist uses a thermometer to check a temperature or a computer to calculate the product of many large numbers, they, in general, are justified in believing in the reliability of the results without, for instance, needing to independently verify the product of the multiplication. However, as I will show, subtle but fundamental distinctions exist between relying on deep learning models and relying on traditional scientific instruments. These distinctions block a reduction of the justification for believing in the reliability of DLMs to that of instruments. In particular, these distinctions concern the nature of the epistemic relations that scientists stand in with respect to the underlying processes (Goldberg, 2014) that mediate the outputs of their instruments (Goldberg, 2020) and the underlying processes that mediate the output of AI models. As a result, in Sect. 3 I argue that explaining what justifies belief in the reliability of AI models cannot be accomplished by appealing to the general form that such justification takes when scientists can be said to rely on traditional scientific instruments.

Another possibility for justifying belief in the reliability of deep learning models is by appealing to the justification scientists have for trusting *other scientists*. After all, this form of justification is distinct from and not reducible to that of instruments (Goldberg, 2020). Here too, when a scientist asks a domain expert a question concerning that domain, they, in general, are justified in believing in the reliability of the answer without, for instance, needing to independently verify the claim (Goldberg, 2014, 2021; Wilholt, 2020).[2] So, we might think that the best option for characterizing the justification for scientific reliance on deep learning models is by thinking of DLMs as *expert agents*. Indeed, deep learning is synonymous with artificial intelligence, a term that evokes something rather like agential status. Moreover, some have already

[2] Recent research on the trustworthiness of experts and expert claims notwithstanding (Ioannidis, 2005; Wilholt, 2020), throughout, I take it that we are presumptively entitled to the belief that experts are following best practices and are not being dishonest in their claims.

begun theorizing how best to situate human efforts alongside AIs conceived as 'alien' collaborators or interlocutors in scientific investigations (Bommasani et al., 2021; Sourati & Evans, 2021) or to develop a new science dedicated to studying the behavior (or actions) of AI (Rahwan et al., 2019). Nevertheless, In Sect. 4 I show that an approach to justifying our belief in the reliability of DL by way of a reduction to that of other agents also fails. Like with instruments, the failure results from fundamental distinctions concerning the nature of the epistemic relations that scientists stand in with respect to the underlying processes of expert reasoning and the underlying processes that mediate AI outputs.

As a result, we are left with either accepting that there is no justification for belief in the reliability of deep learning models beyond pragmatic considerations (which I deny) or that what can justify belief in their reliability represents a philosophically novel approach. In Sect. 5, I argue for the latter and conclude that contemplation of Artificial Intelligence is an occasion and opportunity for exciting, new philosophy of science.

2 Brute inductive reliability

One way scientists can come to be epistemically entitled to depend on the reliability of some process is the consistent success of that process in producing accurate results (Goldman, 1979). Here, the justification for belief in the reliability of a process is, what I call, brute inductive consideration. How might one establish the reliability of DL through brute inductive consideration?

One approach would be to evaluate the past success of deep learning *in general*. This would involve looking at all instances in which deep learning was deployed in a scientific setting and evaluating the ratio of its successes to failures. I take it as straightforwardly uncontroversial that deep learning's track record, in general, is insufficient to warrant belief in its reliability. For one thing, we do not even have access to the (undoubtedly countless) failed attempts to train a reliable model. Yet, this should not count against the reliability of a particular deep learning model any more than it should count toward it. Scientists do succeed in training highly successful deep learning models. In evaluating their accuracy, scientists do not look to the track record of deep learning in general but, instead, to the accuracy of the *specific* model under evaluation.

Evaluation of the accuracy of specific models does involve inductive considerations. Given a dataset of inputs and outputs, deep learning algorithms train DLMs by minimizing a loss function such that the resulting model has a high degree of accuracy in generating outputs given inputs on test data that were not in the training sample. Minimization is carried out by iteratively updating the weights on all connections in the network through back-propagation of errors from prior iterations. In this way, deep learning algorithms sweep through the space of functions representable by the network to find a model that best approximates the function that generates the training data. The model is assumed to be statistically low risk if it performs well on a randomly selected, independent, and identically distributed test set drawn from the same underlying distribution as the training set. Given that the data used in a test set were

not seen by the model during training and given that each classification a model makes during testing counts toward the model's overall performance, a model that does well on a large test set can be said to have an excellent track record upon which to base a brute inductive assessment of its reliability. Finally, a model's accuracy is expected to generalize to data that was not used for either training or testing. However, when it comes to generalization, there are certain assumptions at play concerning the nature of the underlying distribution that are not always principled.

As a matter of fact, scientists do not yet understand *how* deep learning models generalize nor the conditions under which they will generalize well rather than poorly Räz (2022), Zhang et al. (2021). This is due, in part, to the empirical fact that the space of models representable by a neural network is exceedingly large, and contemporary deep learning models often return many equally good-looking models (e.g., trained to the same level of iid generalization). While these equally good-looking models are often treated as equivalent based on their training and testing domain performance, they can behave very differently in deployment domains. This form of 'underspecification' has been shown to lead to instability and poor model behavior when models are used in practice D'Amour et al. (2020). One reason is that real-world data in many domains often exhibit fat-tailed distributions (e.g., power-law distributions, Cauchy distributions) with undefined variance. This means that events that would be very unlikely (so-called 'corner cases') in normally distributed data can be rather common in distributions with undefined variance. While a model may be highly accurate on a training and test set that captured many common corner cases, the space of possible unlikely events drawn from a fat-tailed distribution is significant. This is not, by itself, a reason to think that deep learning models are, in general, unreliable. Rather, it means that scientists cannot rely solely on model performance on out-of-training-sample data (e.g., brute inductive considerations) to justify belief in model reliability, as they cannot be sure that they have not underestimated statistical risk or because their model is likely underspecified or both.

Discarding underspecified models from sets of equally good-looking models is frustrated in large part by model opacity. A lack of DL model transparency prevents scientists from assessing the degree to which models encode inductive biases (Neyshabur et al., 2014) that generalize well to real world data. As a result, scientists are rarely in a position to assess the conditions under which a given model will fail to generalize. Traditionally, scientists can evaluate the reliability of their models by examining not only their accuracy but the degree to which they accurately represent their targets (Baker, 2021; Giere, 2010; Weisberg, 2012), carry out the purposes for which they were constructed (Parker, 2020), exemplify properties or principles of the target that are under evaluation (Frigg, 2010; Frigg & Nguyen, 2016), and so on. What all of these approaches to evaluation have in common is the necessity of model 'transparency'. Yet, in general, it is well known that deep learning models are not transparent.

Philosophers and AI researchers alike have raised concerns about the epistemological impact of neural network opacity on science and society (the failure of brute inductive consideration counting as just one such concern). Of course, from a strictly mathematical perspective, deep learning models are fully transparent given that the weight matrices that mediate the underlying processes that transform inputs to outputs are

directly observable (Leslie, 2019; Lipton, 2018; Zerilli, 2022). However, in general, the high-level logic (if there is one) of a fully trained deep learning model cannot be interpreted in terms of its target system in a way that would allow someone to understand Räz and Beisbart (2022); Sullivan (2019) or fathom (Zerilli, 2022) how the individual parts interact and contribute to the network's outputs. Yet, this level of transparency would be needed to directly address the concerns outlined above. As a result, Creel (2020) and Zerilli (2022) have argued persuasively that deep learning models are epistemically opaque in Humphreys' sense (Humphreys, 2004) meaning that a lack of DLM transparency prevents scientists from attending to the epistemically relevant factors of the model needed to justify their belief in the reliability of claims made on the basis of a model's outputs. For Creel, these factors are the network's algorithmic and structural interpretations (Creel, 2020) which are, for Zerilli, neither intelligible nor fathomable (Zerilli, 2022). Recently, Duede (2022) has argued that, while there are contexts in which DLM opacity does not prevent justified scientific knowledge, if the outputs of epistemically opaque models are treated as claims that, themselves, stand in need of justification that can only be furnished through an evaluation of the underlying process that mediates model output, then opacity is straightforwardly problematic.

If the underlying process of a deep learning model is epistemically opaque, then a scientist cannot directly evaluate the process to form a judgment about whether it is reliable or not. That does not mean that the underlying process is unreliable. It just means that our justification for believing in the reliability of the process cannot be based on brute inductive considerations alone because we are not in a position to make such an evaluation. This is because, due to epistemic opacity and empirical facts of the world (e.g., fat-tailedness), we do not stand in the right sort of epistemic relation to the underlying process to establish the conditions and limits under which we are and are not justified to believe in its reliability.

Of course, scientists routinely form beliefs about the outputs of processes that they are justified in believing to be reliable on grounds other than brute inductive consideration. Some of these processes are even entirely or partially epistemically opaque (Humphreys, 2004). In what follows, I examine the nature of the epistemic relation that scientists stand in relative to the processes that mediate the outputs of scientific instruments and processes of that mediate expert judgments and ask whether their justified belief in the reliability of either can serve to warrant belief in the reliability of deep learning models.

3 Instrument reliability

It is common to describe deep learning models as scientific instruments. Indeed, many of their applications share at least surface-level similarities to traditional instruments. They detect, measure, predict, control —on and on. Obviously, many scientific instruments are reliable, and the reasons scientists have for believing in their reliability are not based solely on brute inductive considerations. In this section, I consider the question of what justifies belief in the reliability of scientific instruments and ask whether

that justification can also serve to warrant belief in the reliability of deep learning models.

In what follows, I will use a broader sense of the concept 'scientific instrument' than may immediately come to mind. It is intuitive to think of scientific instruments as physical objects that one can touch or put to one's eye (e.g., telescopes and cyclotrons). Indeed, a broad philosophical and historical literature has focused on the strictly material nature of scientific instruments (Baird & Faust, 1990; Galison, 1997; Hacking, 1983; Shapin & Schaffer, 2011). However, the broader sense of scientific instrument I consider in this section also includes instruments such as statistical techniques for detecting and teasing out correlations (e.g., regression techniques), computational instruments for solving or approximating solutions to equations (e.g., simulations), and even instruments for discovering causal relationships and effects (e.g., randomized control trials).

Considering a wider sense of the concept reveals a crucial distinction that obtains between two general categories of instruments. The first are *physically mediated* instruments and the second are *theoretically mediated* instruments. The distinction concerns the underlying processes that mediate the manner by which instruments belonging to each category derive their outputs. Instruments such as the mercury thermometer represent physically mediated instruments. Such instruments work primarily by exploiting causal, law-like, physical processes in the world. Physically mediated instruments are distinct from theoretically mediated instruments, such as computational simulation. Instruments of this latter kind work by carrying out a reliable, theoretically informed procedure to arrive at an output. As I will show, what justifies our belief in the reliability of physically mediated instruments is distinct from that which justifies our belief in the reliability of theoretically mediated instruments [a claim echoed in the broad literature on scientific instruments (Baird, 2004; Charbonneau, 2010; Hacking, 1983)]. So, when it comes to the reliability of scientific instruments, in general, there exist two distinct epistemic categories to which we appeal in seeking justification for belief in their reliability.

The distinction I draw between physically and theoretically mediated instruments mirrors a distinction drawn by other philosophers of scientific instruments. For instance, Harré (2010) argues that scientific instruments can be divided into two categories. On Harré's understanding, an "instrument" is a device for detecting and measuring natural phenomena, while an "apparatus" is used to study natural processes by simulating them. The former are physically mediated, while the latter are theoretically mediated. Similarly, Baird (2004) argues that there are two fundamental epistemic categories of scientific instruments: those that create phenomena and those that are models. The former work reliably in so far as they regularly produce phenomena through their activity in a way that is not dependent on theoretical considerations (e.g., via causation), while the latter are broadly similar to theories.

My conceived distinction between physically and theoretically mediated instruments, while more spartan and general, mostly aligns with Harré's and Baird's categories. It is important, however, to note that many (perhaps most) instruments blur the distinction, and I am not claiming that bright lines can always be easily drawn. Nevertheless, for any given instrument, at bottom, the fundamental process that acts to mediate the result will be either physical or theoretical. Of course, many sophisticated instruments will

weave physical and theoretical processes together. Yet, even here, justification for the reliability of the underlying processes that constitute that fabric must be secured, and the nature of the justification will depend on the type of mediation involved.

3.1 Physically mediated reliability

Physically mediated instruments function by effecting or being affected by (or both) some causal, law-like, physical process in the world (Baird, 2004; Charbonneau, 2010; Goldberg, 2020; Harré, 2010). Consider the Geiger counter. This scientific instrument is designed to detect ionizing radiation. It functions by instantiating conditions under which the presence of ionizing radiation will cause an electrical charge to form and be both conducted and detected by the device. Specifically, the design of the Geiger counter exploits what is known as the Townsend avalanche phenomenon. Here, an electric field applied across an inert gas creates conditions under which an ionized particle passing through the field will liberate an electron which, in turn, liberates more electrons, on and on in a cascading event generative of a detectable and measurable charge. So long as the field over the inert gas is of high enough voltage, an ion passing through will *cause* a charge to form as a matter of physical necessity. This specific physical necessity is exploited in the design and use of the Geiger counter. Reliably, just so long as physical conditions are satisfied, a radioactive particle will cause a charge to form within a Geiger–Müller tube. This exploitative approach generalizes to all physically mediated instruments such that all are designed and used to exploit some causal, physical necessity, or necessities, and this exploitation is the key to their reliability.

Our justification for believing in the reliability of such instruments, then, is based on our having good reasons to believe that a physical process connects the instrument's behavior to the world in such a way that the latter causes the former. In the vast majority of cases, physically mediated instruments are designed, from the start, to exploit well understood causal relations or pathways. Yet, these relations or pathways need not always be based upon well-understood, theorized, or hypothesized principles. What is required is that we have good reason to believe that reliable, spatiotemporally continuous processes connect the behavior of an instrument with an event of interest. Consider that, for quite some time, scientists' use of a mercury thermometer was based on observed and precise correlations between the expansion of mercury and temperature. The use of various types of lenses (e.g., biconvex and biconcave) had been widespread for hundreds of years [if not millennia (Sines & Sakellarakis, 1987)] before the principles of optics that govern their light focusing and dispersing capacities were understood. For instance, seventeenth century astronomers did not know how the lenses in their telescopes magnified, were not sure how to improve the reliability of their lenses, and all available knowledge of optics was insufficient to account for key processes such as refraction. Yet, Galileo was able to refine his telescopes continuously (Zik & Hon, 2017). Similarly, for centuries before anything like modern chemistry or specific knowledge of pH, the use of litmus (by, for instance, *alchemists*) to reliably evaluate the acidity of substances was commonplace.

For seventeenth century astronomers, the underlying processes that mediated the behaviors of their telescopes were epistemically opaque in the sense that optical principles that ensure the reliability of their instruments were not directly evaluable. When it comes to physically mediated instruments, even in cases where the underlying process central to the instrument's reliability is epistemically opaque (e.g., early telescopes), our belief in the reliability of the process is still justifiable. Physically mediated instruments perform reliably partly because their behavior does not rely on theory or, as Baird puts it, their "action has been separated from human agency and built into the reliable behavior of an artifact." (Baird (2004), p. 12). Where physically mediated instruments are concerned, once the limits and conditions under which the instruments function with consistency have been established (either by subjecting them to careful tests via experimentation or known a priori from theory), if the instrument is used within these limits and under the right conditions, then the process of obtaining reliable results is, as it were, out of our hands—mercury will *always* expand at the same rate when heated; a charge will *always* form in an electrified field over an inert gas when exposed to ionizing radiation.

Certainly, there are causal processes at play when a deep learning model is in use. The most obvious is the underlying physical processes unfolding within a digital computer. There is a sense in which the flow of electrons through logic gates causes the algorithm to execute. However, at bottom, the process that mediates the behavior of a DLM is the logic of the learned algorithm itself. No direct causal connection between the world and the DLMs mediates the model's output of a given value.

If we consider the formal representation of a trained DLM, we can see that the entire model (regardless of how many layers there are) is expressible as a single, highly nonlinear, nested function. So, the output of a deep learning model is merely a function of the input, where the function just is the model. Scientists do not need to rely on some particular physical process when they believe in the reliability of a DLM. Of course, they need to rely on a computer to *instantiate* the model. It would not be possible in any reasonable amount of time to calculate an output for a model of any significant size. So, in some sense, the model depends on the computer, but not for its reliability. As a result, if our belief in the reliability of DLMs is justified in the same way as that of a scientific instrument, then it is not of the physically mediated variety.

3.2 Theoretically mediated reliability

Theoretically mediated instruments function by carrying out a theoretically informed, algorithmic or heuristic procedural process for accomplishing some goal or completing some task. They are reliable insofar as the underlying procedural process is reliable. Computational simulation is a good example of a theoretically mediated scientific instrument that bears some resemblance to deep learning in the philosophical literature. It is typical for real-world systems that admit of spatial or stochastic dynamics (e.g., gravitationally bound systems of masses, signaling systems, molecular biological systems, economies—on and on) to be represented by mechanistic models (e.g., differential mathematical equations). Solutions to the equations that describe such models might represent system states evolved from particular initial conditions.

While such mathematical models admit of no analytic solutions, numerically approximate solutions can often be found cheaply by means of computational simulations. Computational simulation functions to solve such models by instantiating algorithmic procedural processes designed for operationalizing the mathematical methods of numerical analysis, such as discretization and numerical integration (e.g., Riemannian summation).

Unlike physically mediated instruments, for theoretically mediated instruments like computational simulation to function, we must know the underlying procedural process, how it works, and how to implement it in the instrument. Moreover, our justification for believing that the procedural process (e.g., algorithm) is reliable requires knowledge of the principles that the procedure operationalizes (e.g., discretization, numerical integration, real-root isolation) and under what conditions these principles apply (e.g., continuous, polynomial functions). In order to ensure that the process can reliably succeed in carrying out its task, specific criteria must be met concerning the quality of the implementation, data structures of inputs, the appropriateness of the application, and the soundness of the underlying assumptions. These qualities serve to condition and constrain the reliability of any given procedure for estimating effects.

When Humphreys argued that computational simulation was philosophically novel (Humphreys, 2009), he joined (Oreskes et al., 1994) in worrying that the underlying procedural process for the numerical approximation of solutions to the equations under simulation cannot, in practice, be checked, verified, or validated by the human scientist even though the solution, itself, is relied upon to license claims about its target system.[3] Here, Humphreys' worry is expressible in terms of Creel's 'run transparency,' which is knowledge of the simulation procedure as it was actually executed on this or that occasion (including the physical processes in the hardware) (Creel, 2020). Roman Frigg and Julian Reiss argued (Frigg & Reiss, 2009) that, in the case of simulation, the opacity concern was not sufficient to warrant new epistemology because both the model (e.g., system of equations) under simulation and the algorithmic procedure for resolving its solution space are still fully interpretable. Moreover, during simulation, the semantics of the model are well articulated and are not lost in the procedure. I argue that, so long as we have good, epistemic justification for the system of equations, are reasonably certain that they have been accurately represented in the computer, and have principled, epistemic justification for belief in the reliability of the procedural processes used in numerical approximation (derived from well understood principles in the applied mathematical sciences), then the fact that the numerical convergence to a solution to the model cannot be checked in practice seems epistemically acceptable. This is because the procedural processes that mediate the output are not, in fact, epistemically opaque. They are both fully interrogable and justifiable on the basis of accepted, well-established principles. Aspects of this claim are either implicit

[3] In fact, many philosophers have argued that simulation requires special philosophical attention (Galison, 1996; Humphreys, 2004, 2009; Oreskes et al., 1994; Rohrlich, 1990; Winsberg, 2001, 2003). In general, I am sympathetic to the view that computational simulation extends the philosophical literature in genuinely fruitful ways and that consideration of simulation deepens our understanding of scientific methodology. It has, nevertheless, proved difficult to articulate precisely in what ways computational simulations give rise to specific philosophical concerns that are *qualitatively distinct* from those already native to the more general literature on models, experiments, or computation.

or explicit in Parker (2008a) and (2008b), Winsberg (2010) and have been used to motivate positive views concerning computational simulation's status as good science (Norton & Suppe, 2001). Moreover, digital computers are well understood, and we have sufficient justification for believing they are reliable under the right conditions and within certain limits (none of which are violated in executing a simulation).

Yet, we need not even concern ourselves with digital computers. If we focus just on the procedural processes that mediate the output of a computational simulation, we see that, though such computations might be tedious, slow, and subject to operational error, they can, in principle, be carried out by hand. So, the underlying processes that mediate computational simulation and require justification are of the theoretically mediated type. Just so long as the various algorithmic processes are carried out correctly and without error, a manual computation would be as accurate and reliable as any other. As a result, when evaluating the reliability of a theoretically mediated instrument like computational simulation, we need not concern ourselves (much) with the physical processes needed to carry out the procedures (e.g., pen and paper, whiteboard, calculator, computer, cloud).

All aspects of theory mediated instruments are designed, implemented, and operationalized from known or hypothesized principles. Recall that, with physically mediated instruments, just so long as the conditions are right, the underlying process that mediates the instrument's reliability is out of our hands. Scientists do not *design* the physical processes. Rather, they, as it were, *discover* them. With theory mediated instruments, nothing is out of our hands. For instance, in the simulation case, it is practically impossible to find solutions to the equations that represent a double pendulum without executing the correct, mathematically justified procedures for numerical approximation. We would never be justified in believing in the reliability of a brute force procedure that guessed random states of the system, as this procedural process is unlikely to *ever* guess the correct solution, and, even if it did, we would never be in a position to know that it had.

It should strike the reader as at least intuitive to think that deep learning models are highly similar to theoretically mediated instruments. After all, what is executed in a DLM can be formalized as a procedural process that maps inputs to outputs. So, it seems reasonable to suspect that deep learning models are theoretically mediated instruments. It is, for this reason, that I have focused more closely on this class of instrument than the physically mediated variety.

Recall that, to justify our belief in the reliability of a theoretically mediated instrument, we need to know what the underlying procedural process is, how it works, and how to implement it in the instrument. Here, everything is in order. We know what the procedural process is (it is the mathematically transparent function), we know how it works (it passes a weighted sum of outputs from layer to layer), and we know how to implement it in a computer (using the relevant software). However, if the DLM is a theoretically mediated instrument, then, in order to justify our belief in the reliability of its outputs, we must also know what principles the procedural process operationalizes and under what conditions these principles apply. Moreover, we must know when the process is appropriate for use and be in a position to justify the soundness of the assumptions that underlie the process and its applications. That is, we need more than just run transparency, we need Creel's other two forms of transparency—algorithmic

transparency (high-level, logical rules instantiated in the procedure) and structural transparency (how the high-level, logical rules are realized in code) (Creel, 2020). Yet, as we saw in Sect. 2, DLMs are epistemically opaque, lacking both algorithmic and structural transparency (Duede, 2022; Räz, 2022; Zerilli, 2022).

So, in order to justify belief in the reliability of a deep learning model conceived as theoretically mediated, scientists would need to establish and agree upon methods for the evaluation of the model itself that allows for justifying the *principles* (e.g., the high-level, logical rules) which the known procedural process operationalizes. As we saw in Sect. 2, measures of accuracy and precision do not get us to justified belief in reliability because they depend on brute inductive considerations that break down under conditions of epistemic opacity that are not physically mediated. In order to justify the global, high-level logical rules instantiated by the model, those rules would need to be known. However, DLMs are not interpretable in this way, so the initial worries are not directly resolvable.

Before conceding, one might, instead, argue that our justification for believing in the reliability of DLMs only looks like it fails to reduce to that of theoretically driven instruments because of the epistemic opacity of the procedure. After all, if it were possible to assign meaningful, global, high-level logical rules to the weight matrices that specify a given DLM, we would be in a position where principles for the evaluation of learned procedures would be plausibly attainable. Such a situation would allow a straightforward reduction of the epistemic status of DLMs to that of theoretically mediated instruments, since the high-level logic of the model could be assessed in the same way as the high-level logic of other theoretically mediated instruments. I take it that this is the standard view that motivates much of the ongoing interpretability research. The idea driving that agenda is that DLMs are, theoretically, interpretable in such a way as to reveal the underlying procedural principles that govern their outcome behaviors. This approach turns on the idea that the network encodes an in-principle-interpretable algorithm which, in turn, implies that we evaluate the principles that govern it.

The problem that this approach faces is that it is merely an assertion that, at this time, comes with no good reason to believe its central premise—namely, that the model structurally encodes (in Creel's sense) global, high-level, logical rules (e.g., a theory) that can be understood. The recent research on interpretability provides plenty of reason to believe that DLMs do not (Leavitt & Morcos, 2020) encode such models. Current interpretability and explainability research focuses on 'local understanding' or 'local interpretability' (as opposed to global understanding). 'Local approaches' such as gradient saliency and saliency maps help researchers understand how the model responds to *particular* inputs and how the model's behavior changes with movement in the input space. Moreover, not only do these approaches not yield global understanding (Räz, 2022), the very approaches, themselves, are theoretically suspect (Nie et al., 2018), demonstrably fragile (Ghorbani et al., 2019), and may not be suitable for tasks that require knowledge of global, high-level logic, such as identifying outliers, explaining the relationship between inputs and outputs, or debugging a model (Adebayo et al., 2018). Finally, even when local approaches reliably work as intended, they cannot give us a global understanding of the model's high-level, logical rules, a point acknowledged by Sullivan (2019) and, more recently, Räz and Beisbart (2022).

Issues of this kind have led some to rethink 'local' interpretability. For instance, in a prominent piece, Cynthia Rudin has urged that scientists avoid epistemically opaque systems in high-stakes settings (Rudin, 2019).

Consider, then, that current approaches to interpretability do not give us reason to believe that DLMs encode interpretable, high-level, logical rules because these approaches do not even aim at revealing them. Instead, they aim to help researchers understand how a particular model will respond to specific inputs. Yet, this is a long way from delivering the kind of knowledge scientists need concerning the principles the model's procedural processes operationalize and under what conditions and limits those principles apply and fail. Given this, it is clear that our justification for trusting in the reliability of deep learning models cannot be reduced to the justification we have for believing in the reliability of theoretically mediated instruments in general.

Importantly, this is not to deny that interpretability and explainability approaches cannot give us justification for believing in the reliability of DLM. It is just that the justification they give us is not of the same form as the justification we have for the reliability of theoretically mediated instruments. It is also not the justification we have for physically mediated instruments. I will, however, return to the discussion of interpretability techniques in Sect. 5 when I consider whether they represent a novel approach to justifying reliability requiring novel philosophy of science. For now, I turn to evaluate whether deep learning models can be considered reliable on the same grounds as the reliability of scientific experts.

4 Expert reliability

While it is common to describe deep learning as an instrument, it has always been more common to describe it in ways that are suggestive of agential status. The monikers Deep *Learning* and Artificial *Intelligence* are certainly suggestive of agency. Scientists and scholars routinely refer to deep learning models operating in scientific settings or AI-infused applications as expert agents (Bommasani et al., 2021; Branch et al., 2021; Sourati & Evans, 2021; Stevens et al., 2020). Yet, this is not new. The vision put forward by the 1950s and 60s cyberneticists like William Ashby (1961) and Douglas Englebart (1962) was of AI conceived as *expert systems* that encode, complement, augment, and amplify *human intelligence* and capacities. Today, deep learning, conceived as AI, increasingly acts as an autonomous participant in collective epistemic tasks. In such 'AI-in-the-loop' contexts, scientists relate to AIs as scientific agents (Rahwan et al., 2019) and situate them within scientific groups. More ambitiously still, completely autonomous, AI-operated, 'self-driving' laboratories are not merely imagined but constitute a national strategic priority for science (Stevens et al., 2020).

If our justification for belief in the reliability of deep learning models is reducible to that of expert scientists, then what justifies our belief in the reliability of experts? Expert reliability is an enormously complicated area of philosophical and social scientific interest. A significant reason why expert reliability is unlike instrument reliability is that, when it comes to experts, their reliability is parasitic on their having good reasons for the claims they make. As we have seen, instrument reliability depends on exploiting causal, physical processes or carrying out reliable, theoretically mediated, procedural

processes. For experts, making consistently reliable claims in accordance with good reasons (and sound reasoning) is a hallmark of expertise. Nevertheless, not all reasons are epistemic, so issues of trust and trustworthiness become central to belief in the reliability of experts.

It is quite common for scientists and AI researchers to talk about trust and trustworthiness concerning AI systems. Indeed, the U.S. *National Science Foundation* recently announced[4] its intention to fund several institutes focused on issues pertaining to individual and community level trust in AI systems. Traditionally, however, trust has been taken to be an attitude and relation that can only be directed toward another agent and plays out in various efforts of cooperation and social relationships. It has been noted that, because trust involves believing that the agent you are trusting has your best interests at heart and is motivated to act accordingly, deep learning models (AI's) cannot stand in the trust relation (Hatherley, 2020). Moreover, it has been argued that, while deep learning can be relied on, it cannot be trusted because it cannot have emotive states and cannot be held responsible for its 'actions' (e.g., outputs) (Ryan, 2020).[5] While this conception of trust has been well theorized in moral, social, and political settings (Baier, 1986; Baker, 1987; Hardin, 1996; Holton, 1994; Jones, 1996, 2012), it has also been noted that without trust among scientists, contemporary science would not be possible (Fricker, 2006; Frost-Arnold, 2013; Gerken, 2015). One critical pathway through which trust enters into general epistemic concerns is by way of the knowledge we might acquire through testimony (Faulkner, 2007; Hardwig, 1985, 1991; Hieronymi, 2008; Hinchman, 2005; Keren, 2014; Lackey, 2010; Nickel, 2012). Here, our justification for believing what we have been told depends, in part, on whether we have good reasons to trust the speaker.

In scientific settings, however, we see that our justification for belief in what another expert testifies to is logically separable from trust, as the justification for the *claim* depends on the *evidence* in support of it, not on the trustworthiness of the speaker (Elgin, 2017; Goldberg, 2014, 2021; Meeker, 2004). Expert reliability, then, is separable from issues of trust and trustworthiness as, at bottom, what justifies belief in the reliability of an expert is the reliability of the underlying process that mediates what claims the expert comes to believe and testify to. In this, I follow Goldberg (2014), (2021) and argue that one is *defeasibly* justified in believing in the reliability of an expert, just so long as the expert's claims are mediated by underlying processes that 'constitute expert judgment within [their] domain of expertise'. The justification is defeasible because the expert can, among other things, be untrustworthy. In this way, expert reliability is less about the individual agent and more about the *process of expert reasoning* that results in the claim.

When a scientific expert makes a claim, they are responsible for providing evidence for and reasons that support the claim. When other scientists evaluate the reliability of the claim, they evaluate not just the evidence, but also the expert's reasoning in light of the evidence. The evidence and expert reasoning in light of it represent the *first-order* reasons for the claim. So, our justification for belief in the reliability of

[4] See https://www.nsf.gov/pubs/2022/nsf22502/nsf22502.htm.

[5] See, however, Nguyen (2020) who argues that trust is an unquestioning attitude which can be taken with respect to, among other things, ropes.

an expert claim depends on whether we have access to good first-order reasons for the claim. Scientists are able to evaluate first-order reasons because, in general, the evidence and reasoning process of an expert are made explicit with the claim. Of course, what counts as evidence and epistemically acceptable reasoning for specific scientific claims is context- and domain-dependent.

However, if justifying belief in the reliability of expert claims in science required evaluating all the first-order reasons in support of those claims, nothing would ever get done. This is why science requires a degree of epistemic trust (Fricker, 2006; Frost-Arnold, 2013; Gerken, 2015) and reliance (Wilholt, 2020). At bottom, however, what that reliance ultimately depends on is a justified belief that there are, in fact, good first-order reasons that support the reliability of the claim. An example is helpful here. Suppose that a mathematician tells a physicist that some theorem is true. The physicist, not expert enough in this area of mathematics, cannot directly evaluate the proof of the theorem for themselves. Nevertheless, the physicist may still have good (justified) reasons to believe that the theorem is true. These *higher-order* reasons might include the fact that the mathematician is well regarded, is the author of the proof, and that the proof has been peer-reviewed. All of these higher-order reasons contribute to the physicist's justification for the belief that good first-order reasons support the reliability of the claim that the theorem is true. In science, appeals to the authority of others (e.g., reliance on higher-order reasons) like this is both ubiquitous and necessary. No one can establish, for themselves, the necessary first-order reasons for all of the claims that constitute the body of scientific knowledge. Importantly, however, this is only acceptable so long as the body of scientific claims is, in fact, supported by evidence of the first-order variety. As John Hardwig put it, "[t]he chain of appeals to authority must end somewhere, and, if the whole chain of appeals is to be epistemically sound, it must end with someone who possesses the necessary evidence, since truth claims cannot be established by an appeal to authority, nor by investigating what other people believe about them." (Hardwig, 1985, p. 337) The brute inductive consideration of the mathematician's reputation cannot justify belief in the reliability of the claim that the theorem is true. That justification can be secured only through a direct evaluation of the first-order reasons for its truth—the proof.

In order to model our trust in the reliability of deep learning models on our trust in the reliability of expert agents, we need to show that either DLMs have good first-order reasons for their outputs that we can evaluate, or we need to show that we have good higher-order reasons for believing that DLMs possess good first-order reasons. To demonstrate that our trust in the reliability of DLMs cannot be reduced to a version of our trust in experts, it would be enough to show that we cannot have good reasons for believing that DLMs have good reasons for their outputs. The most straightforward way to do this is to simply deny that deep learning models are the kinds of things that have reasons. However, this argument is harder to make than it seems. Yet, it is also possible to show that the reduction is blocked without denying that DLMs can have reasons for their claims. Here, it is sufficient to show that, even if they have reasons, those reasons are not evaluable by others.

So, let us posit that deep learning models have reasons for their claims (e.g., outputs). However, due to epistemic opacity, these reasons cannot be directly evaluated by others. As a result, our route to first-order reasons for belief in the reliability of the

model's outputs is blocked, as we cannot say whether those reasons are epistemically acceptable. We might have higher-order reasons, but, as noted above, this is only epistemically acceptable so long as the body of scientific claims is, in fact, supported by good evidence of the first-order variety.

In Sect. 3.2, we considered the efficacy of recent interpretability techniques for grounding our justification in the reliability of DLMs in that of theoretically mediated instruments. We saw that this fails as these techniques are 'local' and cannot supply a global understanding of the model, which blocks our ability to evaluate the conditions and limits of its applicability. However, one might think that such techniques are applicable here. After all, if we assume that the model has reasons for its outputs, and if we want to evaluate the first-order reasons for a particular output given an input, then there seems to be *prima facie* reason to believe that 'local' saliency methods are applicable here. The idea is that, given an input, we can use local explainability techniques to explain why the model made the output that it did.

There are two reasons why this approach fails. The first is that these techniques do not give us access to a DLM's 'reasons'. Rather, they give us local, linear estimates of how the model will behave given changes to values in the input space. From this, we must infer what the model is responding to. This is not at all the kind of thing scientists do when evaluating the claims of other experts. When it comes to evaluating the epistemic credentials of an expert's first-order reasons, we do not need to infer what those reasons are from estimates of how the expert would behave if presented with a different question. Second, recent work in interpretability (Adebayo et al., 2018) has shown that these approaches lack principles that determine the conditions under which the approaches themselves are reliable (Räz & Beisbart, 2022). Importantly, this is not to deny that these approaches can give us access to justified belief in the reliability of DLMs. Rather, that the access they give us is not of the same variety as we have when evaluating the reliability of experts.

5 Concluding discussion and new directions

Deep learning has become increasingly central to science, primarily due to its capacity to quickly, efficiently, and accurately predict and classify phenomena of scientific interest. This paper aimed to show that when scientists believe in the reliability of the predictions and classifications they get from DLMs, that belief cannot be modeled on the reliability of mere scientific instruments, nor can it be modeled on the reliability of other experts. The question of this paper is not whether scientists can be justified in trusting in the reliability of DLMs. I take it that they can. Instead, this paper has argued that the epistemic categories of justification for belief in the reliability of experts and instruments are distinct and that belief in the reliability of DLMs cannot be reduced to either.

One might conclude from the preceding arguments that there is no justification for belief in the reliability of deep learning models. This conclusion strikes me as overly pessimistic and fails to be sensitive to the astonishing scientific capabilities and breakthroughs that deep learning has recently enabled. In Sects. 3 and 4, I discussed techniques for understanding the behavior of deep learning models. While these tech-

niques are not yet robust, are post-hoc, and often fail to secure the kind of justification we need, they represent a novel class of approaches to securing justification that is still nascent but promising. A detailed treatment of recent advances in this area is well beyond the scope of this paper. Nevertheless, what these advances have in common is that they, in general, deploy the methods of scientific observation and experimentation that have traditionally been used to understand target systems of interest in the service of justification for belief in the reliability of novel methods that tell us about the world.[6]

As a result, the widespread use and reliance on deep learning models in science has opened up a qualitatively new epistemic category of reliability, and this represents an opportunity for genuinely novel philosophy of science. Recent work Räz (2022) linking explanation of deep learning models to statistical explanation (Salmon, 1971), as well as work that carefully demarcates the robustly justifiable role that deep learning can play in discovery (Duede, 2022), represent promising new epistemological avenues for the philosophy of science to explore.

Acknowledgements This manuscript benefited greatly from conversations with Kevin Davey, Tyler Millhouse, Jennifer Nagel, Wendy Parker, Tom Pashby, Anubav Vasudevan, Bill Wimsatt, participants of the *Theoretical Philosophy Workshop* at the University of Chicago, and the insightful feedback of two anonymous referees.

Funding This work was supported by the US *National Science Foundation* #2022023 NRT-HDR: AI-enabled Molecular Engineering of Materials and Systems (AIMEMS) for Sustainability.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest, including affiliation with or involvement in an organization or entity with a financial or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems, 31*, 9505–9515.

Ashby, W. R. (1961). *An introduction to cybernetics*. Chapman & Hall Ltd.

Baier, A. (1986). Trust and antitrust. *Ethics, 96*(2), 231–260.

Baird, D. (2004). *Thing knowledge: A philosophy of scientific instruments*. University of California Press.

Baird, D., & Faust, T. (1990). Scientific instruments, scientific progress and the cyclotron. *The British Journal for the Philosophy of Science, 41*(2), 147–175.

Baker, B., Lansdell, B., Kording, K. (2021). A philosophical understanding of representation for neuroscience. arXiv preprint. arXiv:2102.06592

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

Birch, J., Creel, K. A., Jha, A. K., & Plutynski, A. (2022). Clinical decisions using AI must consider patient values. *Nature Medicine, 28*(2), 229–232.

Boge, F. J. (2021). Two dimensions of opacity and the deep learning predicament. *Minds and Machines, 32*(1), 43–75.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen,

[6] Experimental techniques are used to calibrate some physically mediated instruments. However, DLMs are not physically mediated instruments. They are mathematical functions.

A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint. arXiv:2108.07258

Branch, B., Mirowski, P., & Mathewson, K. W. (2021). Collaborative storytelling with human actors and AI narrators. arXiv preprint. arXiv:2109.14728

Buckner, C. (2018). Empiricism without magic: Transformational abstraction in deep convolutional neural networks. *Synthese, 195*(12), 5339–5372.

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass, 14*(10), e12625.

Charbonneau, M. (2010). Extended thing knowledge. *Spontaneous Generations: A Journal for the History and Philosophy of Science, 4*(1), 116–128.

Chen, Y., Lin, Z., Zhao, X., Wang, G., & Yanfeng, G. (2014). Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote sensing, 7*(6), 2094–2107.

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science, 87*(4), 568–589.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., Mitani, A., Montanari, A., Nado, Z., Natarajan, V., Nielson, C., Osborne, T. F., Raman, R., Ramasamy, K., Sayres, R., Schrouff, J., Seneviratne, M., Sequeira, S., Suresh, H., Veitch, V., Vladymyrov, M., Wang, X., Webster, K., Yadlowsky, S., Yun, T., Zhai, X., & Sculley, D. (2020). Underspecification presents challenges for credibility in modern machine learning. arXiv preprint. arXiv:2011.03395

Duede, E. (2022). Deep learning opacity in scientific discovery. (Forthcoming at Philosophy of Science) arXiv preprint. arXiv:2206.00520

Elgin, C. Z. (2017). *True enough*. MIT Press.

Engelbart, D. C. (1962). *Augmenting human intellect: A conceptual framework*. Menlo Park.

Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., & Danks, D. (2021). Governing AI safety through independent audits. *Nature Machine Intelligence, 3*(7), 566–571.

Faulkner, P. (2007). On telling and trusting. *Mind, 116*(464), 875–902.

Fricker, E. (2006). Second-hand knowledge. *Philosophy and Phenomenological Research, 73*(3), 592–618.

Frigg, R. (2010). Fiction and scientific representation. In *Beyond mimesis and convention* (pp. 97–138). Springer.

Frigg, R., & Nguyen, J. (2016). The fiction view of models reloaded. *The Monist, 99*(3), 225–242.

Frigg, R., & Reiss, J. (2009). The philosophy of simulation: Hot new issues or same old stew? *Synthese, 169*(3), 593–613.

Frost-Arnold, K. (2013). Moral trust & scientific collaboration. *Studies in History and Philosophy of Science Part A, 44*(3), 301–310.

Galison, P. (1996). Computer simulations and the trading zone. In P. Galison & D. J. Stump (Eds.), *The disunity of science: Boundaries, contexts, and power* (pp. 118–157). Stanford University Press.

Galison, P. (1997). *Image and logic: A material culture of microphysics*. University of Chicago Press.

Gerken, M. (2015). The epistemic norms of intra-scientific testimony. *Philosophy of the Social Sciences, 45*(6), 568–595.

Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*, 3681–3688.

Giere, R. N. (2010). *Explaining science: A cognitive approach*. University of Chicago Press.

Goldberg, S. C. (2014). Interpersonal epistemic entitlements. *Philosophical Issues, 24*(1), 159–183.

Goldberg, S. C. (2020). Epistemically engineered environments. *Synthese, 197*(7), 2783–2802.

Goldberg, S. C. (2021). What epistemologists of testimony should learn from philosophers of science. *Synthese, 199*(5), 12541–12559.
Goldman, A. I. (1979). What is justified belief? In *Justification and knowledge* (pp. 1–23). Springer.
Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge University Press.
Hardin, R. (1996). *Trustworthiness. Ethics, 107*(1), 26–42.
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.
Harré, R. (2010). Equipment for an experiment. *Spontaneous Generations: A Journal for the History and Philosophy of Science, 4*(1), 30–38.
Hatherley, J. J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics, 46*(7), 478–481.
Hieronymi, P. (2008). The reasons of trust. *Australasian Journal of Philosophy, 86*(2), 213–236.
Hinchman, E. S. (2005). Telling as inviting to trust. *Philosophy and Phenomenological Research, 70*(3), 562–587.
Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy, 72*(1), 63–76.
Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. Oxford University Press.
Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese, 169*(3), 615–626.
Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124.
Jones, K. (1996). Trust as an affective attitude. *Ethics, 107*(1), 4–25.
Jones, K. (2012). *Trustworthiness. Ethics, 123*(1), 61–85.
Keren, A. (2014). Trust and belief: A preemptive reasons account. *Synthese, 191*(12), 2593–2615.
Khalifa, K. (2017). *Understanding, explanation, and scientific knowledge*. Cambridge University Press.
Lackey, J. (2010). *Learning from words: Testimony as a source of knowledge*. Oxford University Press.
Leavitt, M. L., & Morcos, A. (2020). Towards falsifiable interpretability research. arXiv preprint. arXiv:2010.12016
Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. *Available at SSRN 3403301*.
Lipton, Z. C. (2018). The mythos of model interpretability. *Queue, 16*(3), 31–57.
Meeker, K. (2004). Justification and the social nature of knowledge. *Philosophy and Phenomenological Research, 69*(1), 156–172.
Neyshabur, B., Tomioka, R., & Srebro, N. (2014). In search of the real inductive bias: On the role of implicit regularization in deep learning. arXiv preprint. arXiv:1412.6614
Nguyen, C. T. (2020). Trust as an unquestioning attitude. In *Oxford studies in epistemology*. Oxford: Oxford University Press.
Nickel, P. J. (2012). Trust and testimony. *Pacific Philosophical Quarterly, 93*(3), 301–316.
Nie, W., Zhang, Y., & Patel, A. (2018). A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In *International conference on machine learning* (pp. 3809–3818). PMLR.
Norton, S., & Suppe, F. (2001). Why atmospheric modeling is good science. In *Changing the atmosphere: Expert knowledge and environmental governance* (pp. 67–105).
Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science, 263*(5147), 641–646.
Parker, W. S. (2008). Computer simulation through an error-statistical lens. *Synthese, 163*(3), 371–384.
Parker, W. S. (2008). Franklin, Holmes, and the epistemology of computer simulation. *International Studies in the Philosophy of Science, 22*(2), 165–183.
Parker, W. S. (2020). Model evaluation: An adequacy-for-purpose view. *Philosophy of Science, 87*(3), 457–477.
Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. S., … Wellman, M. (2019). Machine behaviour. *Nature, 568*(7753), 477–486.
Räz, T. (2022). Understanding deep learning with statistical relevance. *Philosophy of Science, 89*(1), 20–41.
Räz, T., & Beisbart, C. (2022). The importance of understanding deep learning. *Erkenntnis*. https://doi.org/10.1007/s10670-022-00605-y

Rohrlich, F. (1990). Computer simulation in the physical sciences. In *PSA: Proceedings of the biennial meeting of the philosophy of science association* (Vol. 1990, pp. 507–518). Philosophy of Science Association.
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.
Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics, 26*(5), 2749–2767.
Salmon, W. C. (1971). *Statistical explanation and statistical relevance* (Vol. 69). University of Pittsburgh Press.
Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabi, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature, 577*(7792), 706–710.
Shapin, S., & Schaffer, S. (2011). *Leviathan and the air-pump*. Princeton University Press.
Sines, G., & Sakellarakis, Y. A. (1987). Lenses in antiquity. *American Journal of Archaeology, 91*, 191–196.
Smith, P. J., & Hoffman, R. R. (2017). *Cognitive systems engineering: The future for a changing world*. Crc Press.
Sourati, J., & Evans, J. (2021). Accelerating science with human versus alien artificial intelligences. arXiv preprint. arXiv:2104.05188
Stevens, R., Taylor, V., Nichols, J., Maccabe, A. B., Yelick, K., & Brown, D. (2020). AI for science. Technical report, Argonne National Lab.(ANL), Argonne.
Stinson, C. (2020). From implausible artificial neurons to idealized cognitive models: Rebooting philosophy of artificial intelligence. *Philosophy of Science, 87*(4), 590–611.
Sullivan, E. (2019). Understanding from machine learning models. *British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axz035
Wang, S., Kai, F., Luo, N., Cao, Y., Wu, F., Zhang, C., Heller, K. A, & You, L. (2019). Massive computational acceleration by using neural networks to emulate mechanism-based biological models. bioRxiv (p. 559559).
Weisberg, M. (2012). *Simulation and similarity: Using models to understand the world*. Oxford University Press.
Wilholt, T. (2020). Epistemic trust in science. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axs007
Winsberg, E. (2001). Simulations, models, and theories: Complex physical systems and their representations. *Philosophy of Science, 68*(S3), S442–S454.
Winsberg, E. (2003). Simulated experiments: Methodology for a virtual world. *Philosophy of Science, 70*(1), 105–125.
Winsberg, E. (2010). *Science in the age of computer simulation*. University of Chicago Press.
Zerilli, J. (2022). Explaining machine learning decisions. *Philosophy of Science, 89*(1), 1–19.
Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM, 64*(3), 107–115.
Zik, Y., & Hon, G. (2017). History of science and science combined: Solving a historical problem in optics—The case of Galileo and his telescope. *Archive for History of Exact Sciences, 71*(4), 337–344.