UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Why are human epistemic agents not displaced in machine learning scientific inquiries? A practice perspective on ML in science

Styger, Sahra A.; de Heer Kloots, M.; van der Wal, O.; Russo, F.

DOI
10.1007/978-3-032-03083-2_15

Publication date
2026

Document Version
Final published version

Published in
Philosophy of Science for Machine Learning

License
CC BY

Link to publication

Citation for published version (APA):
Styger, S. A., de Heer Kloots, M., van der Wal, O., & Russo, F. (2026). Why are human epistemic agents not displaced in machine learning scientific inquiries? A practice perspective on ML in science. In J. M. Durán, & G. Pozzi (Eds.), *Philosophy of Science for Machine Learning: Core Issues and New Perspectives* (pp. 315-337). (Synthese Library; Vol. 527). Springer. https://doi.org/10.1007/978-3-032-03083-2_15

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:11 Oct 2026

Chapter 15
Why Are Human Epistemic Agents Not Displaced in Machine Learning Scientific Inquiries? A Practice Perspective on ML in Science

Sahra A. Styger, Marianne de Heer Kloots, Oskar van der Wal, and Federica Russo

Abstract This chapter considers machine learning (ML) practices used in science. Because ML practices enjoy increasing degrees of automation at various stages of the process, the question whether human epistemic agents are displaced arises. We first point out that shifting focus from the ML outputs to the *practice* of designing and using ML models allows one to appreciate the role of different actors in this process, from the human designers and modelers to the algorithms themselves. We illustrate this point with a description of ML-based practices in neuroscience. We then go further with problematizing the role of human epistemic agents in ML and argue that they are not displaced.

Keywords Human epistemic agents · Artificial agents · ML practices · NLP · Cognitive neuroscience · Algorithmic bias · Human displacement · Human-in-the-loop

S. A. Styger (✉)
Department of Philosophy, University of Konstanz, Konstanz, Germany
e-mail: sahra.styger@uni-konstanz.de

M. de Heer Kloots · O. van der Wal
Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, The Netherlands
e-mail: m.l.s.deheerkloots@uva.nl; o.d.vanderwal@uva.nl

F. Russo
Utrecht University, Utrecht, The Netherlands
e-mail: f.russo@uu.nl

© The Author(s) 2026

J. M. Durán, G. Pozzi (eds.), *Philosophy of Science for Machine Learning*, Synthese Library 527, https://doi.org/10.1007/978-3-032-03083-2_15

15.1 Machine Learning and the Question of the Displacement of Scientists

Recent advancements in machine learning (ML) and in artificial intelligence (AI) have led to an increased societal prominence of artificial systems that display some degree of agency and of intelligent and autonomous behavior. To put it roughly, an agent is capable of acting. In that sense, computer agents or artificial agents are described to do things such as operate autonomously, perceive their environment, persist over a long period of time, adapt to change, and create and pursue goals (Russell & Norvig, 2021, 21–22). Then, ML in particular is an artificial *learning* agent, because ML improves performance after, e.g., observing some data, building a model based on the observed data, and using the model to solve a problem (Russell & Norvig, 2021, 669). According to Ezenkwu and Starkey (2019, 335), there is no unifying definition of machine autonomy in artificial agents to assess the degree of autonomy in, e.g., industrial robots or game playing agents. However, Ezenkwu and Starkey (2019, 336ff) propose two categories of attributes of machine autonomy: low-level attributes (incl. perception, actuation, learning, context-awareness, decision-making) and more advanced high-level attributes (incl. domain-independence, self-motivation, self-recovery, self-identification of goals). The high-level attributes in particular are the subject of numerous research projects in the field of autonomous systems with ML, which points to future progress in the degree of autonomy in artificial systems.

We, human epistemic agents, interact with such advanced artificial systems during everyday activities (think of smart devices for domestic use or chatbots used in customer care), and also in proper scientific settings (think of nested models used in climate science or the use of ML techniques in neuroscience). In this chapter, we focus on the use of ML models in scientific research. In numerous contexts, we are witnessing human scientists assisted by artificial counterparts. 'Intelligent' machines are able to take over many tasks where human expertise was deemed central, and this could be perceived as a threat to the unique role of humans in the process of scientific inquiry. With the rapid development of more advanced ML models, one therefore wonders whether humans may get epistemically displaced or even completely replaced by such intelligent machines. Humphreys (2004, 2009) has already expressed similar concerns about displacement of humans by computational science: computational methods used in science raise new questions for the philosophy of science, because computational science uses methods that "push humans away from the center of the epistemological enterprise" (Humphreys, 2009, 616) and of knowledge production. Or, more recently, the question pertains whether technologies that automatically induce scientific discoveries from data (e.g., AlphaFold discovering protein structures; see Jumper et al. (2021)) may with time become so advanced that scientists will be replaced in doing research.

It is worth noting that this sense of displacement is not new—think for instance of the Marxist critique of technology that followed the first industrial revolution—but with yet another wave of developments in AI and ML techniques in the fourth

industrial revolution, it is becoming as timely as ever. Recent philosophical works deal with critical reflections on automated science. For instance, Mieke Boon (2020) has pointed to the question of whether a future is conceivable in which ML algorithms may replace human scientists by taking over more and more epistemic tasks. Within the Philosophy of Information, Luciano Floridi (2016) has already discussed a similar phenomenon using the concept of 'in-betweenness'.

'In-betweenness' refers to the different forms of interactions between agents, qua users, and technologies. The general scheme involves a 'prompter' (i.e., whatever stimulates or suggests using a given technology), a user, and a technology in between them:

$$\text{user} \leftarrow \text{technology} \rightarrow \text{prompter}$$

The simplest type of relation is called *first-order technology*, in which nature acts as a prompter and technology is interposed between humans and nature as a reaction to said prompt:

$$\text{humans} \leftarrow \text{technology} \rightarrow \text{nature}$$

First-order technologies are, for instance, sunglasses to protect our eyes, an axe to split woods, or spectacles to improve our vision. Most of such first-order technologies are analogue technologies that help humans to cope with nature in simple ways. In science, we can think of a simple meter or an analogue telescope to be this kind of technology.

With *second-order technologies*, we alter this configuration, and now technologies are in-between humans and another technology:

$$\text{humans} \leftarrow \text{technology} \rightarrow \text{technology}$$

Thus, for instance, we use a screwdriver to operate on a screw, or a remote controller to turn on the TV. Many household appliances or scientific instruments before the digital revolution are of that type.

Finally, *third-order technologies* are those in which humans are (allegedly) out of this chain, according to the scheme:

$$\text{technology} \leftarrow \text{technology} \rightarrow \text{technology}$$

In this third scheme, technology is not just a prompter (the position on the right) and a mediator (middling position) *but is also an agent that reacts to a prompt* (the position on the left). Numerous digital technologies can interact with other technologies without humans being present. Prima facie, this happens in the Internet of Things, in deep-learning algorithms, or in 'nested' climate models. In science, the practice of using machine learning models to answer increasingly data-driven questions has been suggested as a paradigm-shift away from more traditional experiment-, theory- and simulation-driven approaches to scientific discovery (Hey

et al., 2009). Recent deep learning-powered breakthroughs in for example protein structure prediction (Jumper et al., 2021) and weather forecasting (Lam et al., 2022; Nguyen et al., 2023) can be seen as part of this new paradigm. It is this use of machine learning techniques in numerous scientific contexts that may lead to the impression that we are in Floridi's third-order configuration, in which humans in general and, specific to our context, scientists, seemingly do not appear anymore in the process that leads from data to the model to the scientific output.

In this chapter, we problematize the question of the displacement of humans, and specifically of human scientists by artificial agents in ML-based scientific inquiries. Our answer is that while it is undeniable that ML models are driving change in the way scientific inquiry is conducted, scientists are *not* ipso facto completely displaced. On the contrary, by adopting a *practice perspective* on our philosophical analysis of ML-based science, we show that scientists still have a distinct and fundamental role. Our point is partly descriptive, in that we rely on descriptions of ML practices, but is also normative in that it prescribes a certain role of human scientists in said practices. Briefly put, according to the practice perspective that we embrace, we enlarge the focus from the outputs of ML models as to include relevant and informative descriptions of all the actors involved in the generation of these outputs; methodologically, this is what allows us to identify the enduring role of human epistemic agents in the scientific process. Ultimately, in the chapter, we make two interrelated points. First, the question of 'displacement' does not call for a yes or no answer: The possible displacement comes in degrees, depending on which tasks are outsourced to artificial agents and by how much. Second, we argue that human scientists are not in fact completely displaced: The practice-based description we offer below shows that human scientists remain involved even when their tasks are largely outsourced to ML techniques. While some tasks may disappear, others may remain or arise; specifically, the creation, test, and interpretation of the outputs of ML techniques cannot be outsourced to ML themselves.

The chapter is organized as follows. In Sect. 15.2, we introduce the practice perspective that we use in the rest of the chapter. In Sect. 15.3, we describe machine learning techniques from this point of view, using natural language processing and ML in neuroscience as illustrations. We show that we need a more nuanced understanding of 'displacement of humans', meaning that scientists do not just 'vanish' completely. In Sect. 15.4, we continue the discussion of the potential displacement of humans by focusing on the outputs of ML models, and particularly by paying attention to questions of opacity and explainability and of bias. We conclude in Sect. 15.5 by returning to the concept of 'in-betweenness' and we argue that human scientists remain in the chain, in various capacities and roles, and foremost of the design choices that are needed in several moments of the practice of ML.

We write this contribution from the perspectives of philosophy of techno-science, philosophy of science in practice, and of empirical research in machine learning and neuroscience. We disclose our positionality to help the reader better understand not just the theoretical presuppositions of this chapter, but also its ambitions in terms of setting up a fruitful dialogue between philosophy and empirical research. Specifi-

cally, we embrace a 'practice perspective' and the emphasis on the intertwinement of science-technology and of humans-instruments, as is elaborated in the philosophy of techno-science (Russo, 2022). In this chapter, we specifically investigate how this approach helps shed light specifically on machine-learning practices, which is part of the research project "The Future of Creativity in Basic Research",[1] and how it can further shape empirical research ("How to bridge neurobiology and psycholinguistic theory by computational modelling",[2] "The biased reality of online media - Using stereotypes to make media manipulation visible"[3]). To do so, we combine expertise in philosophy of techno-science (Styger, Russo) and in machine learning (de Heer Kloots, van der Wal).

15.2 Machine Learning from a Practice Perspective

This chapter moves from a distinct approach in philosophy of science that gives prominence to the *practice of science*, in this case the practices of using machine learning techniques in scientific contexts. The methodology of a practice approach is the one broadly used in the circles of the philosophy of science in practice (Poliseli et al., 2022; Poliseli & Russo, 2022), and that has been specifically theorized by authors such as Chang (2014) in history and philosophy of science or as Pickering (1992) in STS (for a discussion, see the whole volume edited by Soler et al. (2014) and Russo (2022, Chap. 3).

In a practice approach, emphasis and attention is given to *how* science is carried out, over and above its finished products (e.g., established theories or successful explanations). Following Chang, it is important to note that in science we never deal with *one* well-defined practice, but rather with systems of practices, in which various elements of analysis play a role. Consider the following elements, (see Chang, 2014), especially Table 2 on p. 16):

- **Activity:** What is being done in the practice in question?
- **Aims:** What is the inherent purpose of this activity, and what external function does it serve?
- **Systematic context:** Does the activity constitute a part of a broader system of practices?
- **Agent:** Who is doing the activity?

[1] The Future of creativity in basic research: can artificial agents be authors of scientific discoveries?, funded by VW-Stiftung grant Az:97721 https://www.philosophie.uni-konstanz.de/ag-mueller/forschung/projekte/the-future-of-creativity/.

[2] Funded by the Netherlands Organization for Scientific Research (NWO), through a Gravitation Grant 024.001.006 to the Language in Interaction Consortium. https://www.dcc.ru.nl/languageininteraction/research-organization/big-questions/big-question-1/.

[3] Funded by the NWO as part of the Open Competition Digitalisation under project number 406.DI.19.059. https://www.nwo.nl/en/projects/406di19059.

- **The second person:** To/with whom?
- **Capabilities:** What must the agent be capable of, in order to carry out this activity?
- **Resources:** Which tools are necessary for this activity to be successful?
- **Freedom:** What kind of choices does the agent make?
- **Metaphysical principles:** What must we presume the world to be like, in order for this activity to be coherent?
- **Evaluation:** Who is judging the results, and by what criteria (in addition to coherence)?

A practice approach, at least in the version proposed by Russo (2022), does not prescribe to address *all* these aspects at once, but to specify, on any given occasion, the elements that are of interest or relevance. In our case, we are specifically interested in how human epistemic agents and artificial agents interact in scientific inquiries that employ ML-techniques, and we ask the question what their epistemic aims and purposes are. We emphasise this question because, following the 'in-betweenness' scheme, it looks like the application of ML techniques in science belongs to a third-order configuration of technologies, in which scientists are displaced and do not play a proper role in the scientific inquiry anymore. Is this really the case? As mentioned, we will show that scientists are *not* completely displaced. To answer this question, we embrace another important stronghold of a practice approach, namely its highly interdisciplinary endeavor, requiring an engagement between philosophers and scientists. The way in which such engagement happens is not predetermined and can take various forms.

With a practice approach, we wish to change the perspective from which ML models are typically analyzed. The usual narrative is about what an "algorithm does" and consequently, about what its output is. It is certainly true that technologies, and especially computational technologies in AI, have (degrees of) agency, as they have been described as intelligent agents (Russell & Norvig, 2021) or learning agents (Eva et al., 2023). However, it is important to consider how these artificial systems operate in an environment, which can be the space in which an algorithm works but also more broadly the "hybrid" network of humans and non-humans of socio-technical systems (Emery et al., 1960; Latour, 2005). In order to better understand what we mean by "environment", we borrow some intuitions from the work of Latour and Woolgar (1986), and notably the *Actor Network Theory*, that suggests analyzing any given case by mapping *all* the actors involved (humans and non-humans) that operate in an environment or, in the terminology of Chang, in a given context. That is certainly the specific computational framework for the case of ML, but also the physical, social, institutional setting in which ML practices are performed. The environment in which ML operates includes scientists as well as a whole range of infrastructure, from the institutional embedding to the material lab equipment, from the protocols of the experiment to the cultural and moral values of the human agents that design and use them. As scholars in Science and Technology Studies would put it, we are dealing with a *socio-technical system* and, from a

practice perspective, human epistemic agents are just as important as the algorithms and ML models in such systems.

In the following, we reconstruct salient aspects of ML from a practice perspective, introducing two *episodes* that belong to said practices. We use the term episode here, rather than 'case study' or 'example' to convey the idea that we are not dealing with special, isolated cases, and that while they are relevant examples of ML in practice, they also belong to a wider and longer narrative, or are like a *series* of episodes (Russo, 2022, Chap. 3).

15.3 Machine Learning in Practice

15.3.1 The Episode of Natural Language Processing

Machine learning as a field is only one of many research endeavors focused on developing artificial intelligence, i.e., computer systems that perform intelligent behaviors. The AI research community has historically been divided into so-called *symbolic* AI and *subsymbolic* (or *nonsymbolic*) AI. As described by Mitchell (2020), symbolic AI used logic-based systems to formalize descriptions of conscious thought processes, resulting in programs operating on manually defined symbols and rules. On the other hand, subsymbolic AI took more inspiration from neuroscience and unconscious thought processes, resulting in programs essentially consisting of mathematical operations on numbers that do not necessitate humanly interpretable reasoning steps. The past decade of developments in AI technologies has been dominated by the latter, subsymbolic paradigm, and specifically by the rise of *deep neural networks* (DNNs) — a class of machine learning systems that learns from large amounts of data, by adjusting weights in layered configurations of computing units. All AI application areas, including for example the image, audio, and text processing domains, have by now transitioned to largely deep learning-based systems, with large-scale pre-trained 'foundation models' at the core of most applications (Bommasani et al., 2022).

Machine learning refers to a whole set of techniques that allow models to learn from a given data set (called a training data set). ML models are designed to make accurate inferences on new data with a similar structure and distribution to the training data, i.e., to learn generalizable solutions. Typical tasks performed by ML models are categorization and recognition (e.g., of objects in images or of sentiments in texts) and generation (e.g., new text, following provided previous context).

ML approaches can be divided into different classes of learning problems that vary in the extent to which the prediction target requires expert design (Bishop, 2006; Jurafsky & Martin, 2023). Whereas supervised systems are trained on data with human-annotated labels (for example pairs of images and annotations of the pictured objects), unsupervised systems are rather tasked with clustering data points into groups (e.g., customer buying behaviors) without labeled classes being defined

beforehand. With the rise of deep representation learning approaches and internet-scale data, so-called self-supervised systems have gained traction, where the initial prediction target is contained within the available unannotated data (e.g., the next word in a sentence, given previous context). In the following, we go into more detail on the course of this transition and some of its consequences for scientific practice in the domain of natural language processing (NLP) specifically.

Natural language processing is an interdisciplinary field that is concerned with algorithms dealing with natural language related tasks. Examples of NLP domains are machine translation (e.g., Google translate), information retrieval (e.g., search engines), and chatbots (e.g., ChatGPT) (Jurafsky & Martin, 2023). The field of NLP has seen many changes in its research goals and epistemic activities. For instance, where 'Good Old-Fashioned AI' (GOFAI) focused on symbolic (explicit, rule-based) solutions, this was later abandoned in favor of statistical techniques that required fewer explicit rules to be hard-coded. This in turn, was largely replaced by subsymbolic (connectionist) techniques that relied on deep neural networks, which attempts at avoiding encoding any 'expert domain knowledge' in the instrument at all.

One trend, it seems, is that the development of computational methods is moving to that of increasingly *general* solutions (Sutton, 2019): where not so long ago, 'feature engineering' (the manually processing and selection of features from raw data) was a crucial part of the NLP modeling process, we now use self-supervised approaches to let these models learn such features from 'raw data'.[4]

At this point in time, language models based on deep neural networks—of which the Transformer (Vaswani et al., 2017) is currently the most popular one—have become the de facto NLP model. These models are trained to do a simple task—e.g., assign probabilities to sentences or predict the most likely next word in a given context (Jurafsky & Martin, 2023)—but have proven to be important building blocks for other tasks more 'downstream' (e.g., chatbots, like ChatGPT, are based on such language models). Some researchers refer to these language models as 'foundation models', since these are general purpose models that can be used in a broad range of applications (Bommasani et al., 2022). An even more recent development is that of 'prompt engineering', where the models are asked in natural language instructions (prompts) to perform specific tasks, without any explicit adaptation to a specific application (Liu et al., 2023). Interestingly, the adoption of these 'self-learning models' has given rise to a field of study (initially referred to as 'BERTology', now 'interpretability') that investigates the inner workings of transformer-based models: e.g., what features are found by the models (Rogers et al., 2021).

The practices in NLP have changed in other important ways, because of the technological changes it has relied on. The success of language models (and deep

[4] In a sense, one can argue that data is never a perfect and objective representation of the world. See for a discussion, for instance, Friedler et al. (2021). With this caveat, the data is less structured than it traditionally was (e.g., instead of syntactic and semantic features explicitly labeled, now the data may contain just textual data), which is why it is referred to as 'raw data'.

neural networks more generally) rests predominantly on the availability of (cheap) computation power (in the form of GPUs) and vast amounts of data ('big data'). Initially, abstractions of the NLP modeling process in the form of pretrained models (foundation models) and code libraries (e.g., those offered by HuggingFace, see Wolf et al. (2020)) have made it much easier for researchers to work with these language models and apply these to new contexts. However, now these models and datasets have become so large that it has become difficult to analyze, let alone train them without the right resources. State-of-the-art language models now comprise hundreds of billions of parameters, and training and analyzing models at that scale is substantially harder than 'smaller' models that fit on computer systems most researchers have access to. Limiting this access even further, many organizations only allow some access to these models via their own proprietary application programming interfaces (APIs). Most researchers are now confined to researching the models only 'downstream' in the modeling pipeline (i.e., studying the output of these models, a point raised in e.g., Talat et al., 2022). This asymmetry has consequences for the autonomy researchers and research institutions have in making epistemic design choices as well as the overall transparency of their research; a point we further discuss in Sects. 15.4 and 15.5.

The aims of researchers working in and with NLP are very diverse. These range from purely technical goals such as training and designing these models (including the collection of vast training datasets), evaluating and analyzing these models, to applying these to other fields. We also see how NLP has become an important factor in 'other sciences'. While NLP (and ML more generally) has clearly served as important inspirations and tools for other scientists, there has been critique on how it has constricted the epistemic practices of these researchers. For example, Gyllingberg et al. (2023) argue that the ML-based focus on fitting models to data draws attention away from the more fundamental scientific activities of formulating and analyzing the models yourself. In this light, we now turn to the epistemic use of ML generally, and NLP models more specifically, in cognitive neuroscience.

15.3.2 Machine Learning as a Modeling Practice: Episodes from Cognitive Neuroscience

The use of ML in cognitive neuroscience (see Chap. 20 by Crook and Kästner 2025) is particularly interesting to contextualize and conceptualize the use of ML as a *modeling practice*. Recalling Chang's element of 'activity', we here address the question of 'what is being done' when we model a phenomenon in science. By focusing on *modeling practices* that employ ML techniques, we can make explicit the epistemic activities, the epistemic aims and the epistemic agents involved in this techno-scientific practice. This is a first necessary step on our way to answer the question if human researchers are being displaced in scientific practice.

In the following episode from cognitive neuroscience, we contextualize the practice of modelling brain activity with ML-based systems. Which epistemic aims and agents are involved in studying the brain with ML? Neuroscientists have long employed statistical regression models in building predictive models of brain activity, which learn from data what stimulus features are most important for predicting activation in (particular areas of) the human brain. However, recent work in this area makes increasing use of features extracted from deep learning systems as the input to such predictive models (a configuration that prima facie fits third-order technologies mentioned above), rather than manually defined predictors designed by scientists. While brain activity of a participant watching an image can be predicted based on annotated features of the scientists' choice (e.g., image brightness, colors, contents), higher predictive accuracy is achieved when using the internal activations of a deep neural network processing the raw image pixels instead, even though such features are not easily interpretable for humans (for a discussion, see also Chap. 12 by Kieval 2025) and Chap. 13 by Freiesleben 2025). Hence, it might appear that the human scientist's role in the practice of modelling brain activity (choosing the relevant predictors) has been replaced by deep learning technologies. However, humans still play a crucial role in the *design* of experiments using these technologies, and thus in scientific progress, which underlines that scientists are not displaced. Here, the early design choices made by scientists on how to study brain activity by modelling brain activity with ML should therefore not be mistaken for displacement. It is also worth noting, as we also mentioned earlier, that human epistemic agents still have roles to play at different stages of the process, a point in line with 'computational reliabilism' as elaborated by Durán (2023).

Cognitive neuroscience is the scientific field that studies the neurobiological bases of mental processes such as language and visual perception, and it is one area of science where ML-based approaches have made a transformative impact. In the past decade, a new strand of research in this field has been developed that makes use of artificial intelligence technologies as scientific models of cognitive tasks (see i.a., (Kriegeskorte & Douglas, 2018; Cichy & Kaiser, 2019; Ma & Peters, 2020)). The general 'cognitive computational neuroscience', 'neuroconnectionist', or 'Neuro-AI' approach embraced in this line of work is to directly compare deep learning systems (for example an image classification or language model) with human or animal subjects in a neuroscientific experiment. This comparison involves the internal states generated inside the deep neural network while processing a stimulus on one side, and the experimental subjects' brain activity when processing the same stimulus on the other side (for example, observing an image or reading a piece of text). DNN models are then evaluated on their ability to capture neurally relevant stimulus properties, reflected in how well brain activity aligns with, or is predictable from, the models' internal states. Such DNN-based predictive models of brain activity based on internal states of AI systems first took off with computer vision systems and neural activity in visual perception (Yamins & DiCarlo, 2016), but have now also produced impressive results with NLP systems and neural activity related to language comprehension (Schrimpf et al., 2021; Caucheteux & King, 2022).

The epistemic aim of predicting neural activity from observed stimuli is not new, although it contrasts with more traditional research practices in neuroscience and psychology based on controlled lab-based experiments. Such studies generally aim to demonstrate relevant contrasts between two or more theory-based but usually quite artificial experimental conditions. In other words, psychological theories are tested by contrasting very specific subcomponents of the stimulus space against each other with the goal to confirm or reject hypotheses. One reason this ML-based approach to studying the brain is embraced is that it allows for more naturalistic experimental paradigms: Studying brain activity through predictive models arguably allows for theorizing about more naturalistic and ecologically valid human behaviors (Yarkoni & Westfall, 2017; Hasson et al., 2020).

However, the goal of predicting brain activity recorded in more naturalistic settings in itself does not require the use of DNN-based features as input to the prediction model. An example of predictive modelling in the language domain can be found in the work of Wehbe et al. (2014), who built a generative model of brain activity recorded during story reading. In their model, brain activity is predicted based on descriptive features of the text being read (e.g., the length of words or their grammatical category). Yet, recent work using models based on internal states of NLP systems achieve higher neural predictivity in language regions than encoding models based on lexical and grammatical feature annotations (Kumar et al., 2023). Hence, a second major reason for scientists' design choice to employ AI technologies as instruments for neuroscience is empirical: Predictive models using DNN-based features achieve much higher accuracy than encoding models using theory-based hand-engineered features.

In shifting from hand-engineered to DNN-based features, some interpretability and control over encoding model design is lost in exchange for predictive accuracy: The basis of encoding models are now features learned by a deep learning system to achieve a technological goal, rather than those formulated by scientific theories of particular cognitive phenomena. As mentioned above, this is an instance of 'outsourcing tasks' in the sense of bestowing 'agency' to ML techniques by means of epistemically using ML agents for our scientific goals. This trade-off is welcomed by researchers who emphasize that nature does not owe us easy answers, so to speak: There is no reason to assume that the inherent complexity of brain functioning can be captured in human linguistic terminology. For example, Doerig et al. (2023, 432) write: "(...) human-interpretable labels for neural activity are limited by the imagination of researchers, or simply by language. But natural mechanisms are not necessarily bounded within these constraints: neural selectivity can often rely on more complex features that only imperfectly map onto human-interpretable categories".

Nevertheless, the NeuroAI approach also receives criticism. Often, these critiques are grounded in a disagreement about the value of predictive models and emphasize the importance of explanation beyond prediction in scientific progress. Accurate prediction of neural activity is deemed useless when models are not grounded in explanatory cognitive theories; in this way, the practice of predicting neural activity using DNN features without such a theory has been likened to per-

fectly predicting the time on analog clocks with the digits displayed on digital clock, which is not an explanation of either clock's internal working (Guest & Martin, 2023). Neural activity may be well predicted by completely unrelated processes (Meijer, 2021), or by models that do not show human-like patterns in their output behavior (Bowers et al., 2022). A general point of criticism is that DNN-based predictions of brain activity are not explanations of brain functioning precisely because the internal states of DNN models are not directly interpretable by human scientists. Hence, the reason for high predictive accuracies remains uninterpretable as well. Here, the displacement of theory-based experimental control by DNN-discovered features seems to cross a line where the kind of scientific understanding that some neuroscientists strive for is out of reach.

However, the choice to prioritize predictive accuracy over human interpretability or explainability does not mean that human scientists are displaced (for a discussion, see Chap. 7 by Páez, 2025). Important choices in the brain activity prediction pipeline are still human-made: Human scientists choose stimuli, construct the setting in which brain activity is recorded, and make choices about what model architecture to use, what data to train it on, exactly what model-generated output to use for prediction. These choices allow scientists to use and investigate brain activity prediction models epistemically by seeking explanations for why particular modelling choices lead to better predictions than others. The choice of using deep learning technologies in cognitive modelling confines the possible space of scientific inquiry to those behaviors and cognitive phenomena that deep learning models can capture, rather than those that are most accessible to human imagination. But this also opens up new epistemic aims and activities, directed at improving and better understanding the modelling pipeline—an area of research where neuroscientists and artificial intelligence researchers are aligned and can mutually benefit from new discoveries (Ivanova et al., 2021; Kar et al., 2022).

15.4 Design Choices at the Core of ML Practice and the Case of Algorithmic Bias

As explained earlier in Sect. 15.2, a practice perspective gives emphasis to *how* science is carried out, over and above its finished products. It is however a good moment to consider the finished products of ML techniques, because it is precisely their opacity (see Chap. 1 by Beisbart (2025) and Chap. 2 by Formanek 2025) that has given rise to philosophical and empirical discussions about trust, reliability, and explainability (Humphreys, 2009; Durán & Jongsma, 2021; Langer et al., 2021; Durán & Formanek, 2018). Specifically, we ask the question: How should we consider results and evidence generated by ML, if we know little about *how* they are generated?

Addressing this question leads us to discuss algorithmic bias or data bias. These are well-known problems, not only for computer scientists, as many different forms

of biases affect our research and ML practices. From a practice perspective, it is obvious that 'we' (those who do science) are responsible for many forms of biases ending up in our ML models, which means that our scientific results may be flawed as a consequence. However, as discussed in Sect. 15.3, even if researchers are able to understand their own positionality, the *asymmetry* in access to these models (i.e., researchers cannot train all models themselves and rely on other organizations to provide pre-trained models) makes it close to impossible to study all sources of biases in the ML models they use. Hence, the responsibility of human epistemic agents to remain actively in the loop to check and control our ML practices, for example, by applying fairness principles to the optimizations, by understanding how marginalized groups can be mis- or underrepresented in the results, and by generally being aware of the potential sources of bias.

Using the episode from cognitive neuroscience, it is still possible for scientists to manipulate models in relevant ways to successfully investigate research questions, even when model features are initially not transparently interpretable by humans. Still, the lack of interpretable reasoning steps in ML-based systems underscores the importance of investigating biases and finding ways to ensure e.g., reliability, trust, and explainability. We connect here to questions of explainability and of transparency that are widely discussed in the literature, as well as elsewhere (see Chap. 4 by Durán, 2025, Chap. 5 by Alvarado, 2025, Chap. 6 by Zednik and Verreault-Julien, 2025, Chap. 8 by Buijsman, 2025, and Chap. 9 by Räz 2025). Our intention is not to contribute to the conceptualization and/or operationalization and implementation of these concepts *directly*, but rather to reiterate a point made earlier: humans are only *allegedly* displaced, it is for us mainly and foremost a question of *design* that is at stake. We instead pose the question of transparency or explainability in the following terms: how much are design choices involved in requiring that certain standards of explainability and of transparency are met and ensured? This is why we think we should (also) focus on our choices about how much explainability or transparency is desirable or required, and depending on the context, and not (only) on what they are. Because ultimately it is the human epistemic agent such as the scientist engaging in ML-based scientific research, who wants and needs their ML models to be trustworthy, transparent, and reliable. ML models and their results should be transparent, reliable and explainable, both from a technological and a political policy-making point of view.

Furthermore, if we can establish a prominent role of humans via design choices, it is also important to note that ML tools (and any other technical artefacts) are not merely executors of some scientific protocol, in the sense that we might also obtain unexpected and difficult-to-explain outputs, in part because we don't have full understanding of the inner working of ML, and in part because novel outputs may emerge from machine-machine or human-machine interactions. Collectively, our considerations about the role of humans in the practice of ML and the conceptualization of (digital) technology as not being inert (i.e., there is a dynamic influence between epistemic practices and technologies) leads to abandoning a form of opposition between *us* (humans) and *them* (the technologies), and embrace a view according to which human and artificial agents are *in a partnership* in the process

of knowledge production. We will specify this partnership in more detail with the idea of humans remaining in the loop in the next section. It also follows that in this partnership the terms are constantly negotiated, and so it must be a continuous question we ask about our role and that of the technologies, what we are prepared (not) to accept from the results generated by technologies (and ML specifically in this case). This argument has been made reflecting on techno-scientific practices in general (Russo, 2022), and we have particularized here some of those considerations for the case of ML.

We continue our reflections on the 'outputs' of ML in general and of NLP in particular by looking at the debates on bias. ML is an area in which questions of bias are becoming relevant and urgent, primarily because the community is increasingly acknowledging cases of harmful biases in ML outputs, such as in recommendations for health systems, where the implications of algorithmic bias are being extensively discussed (Panch et al., 2019; Aquino et al., 2023).

While the study of algorithmic bias is not new (see e.g., Friedman & Nissenbaum, 1996)), recent developments in ML models, especially those based on neural networks, and their increased adoption in real-world applications, has led to growing concerns about the social biases and harms that these models may propagate and even amplify (Zhao et al., 2017; Lloyd, 2018; Hooker et al., 2020; Hall et al., 2022).

To take the example of natural language processing (as discussed in Sect. 15.3.1), we see a trend of increasingly large language models (LLM), both in parameter size and the amount of textual training data, which poses many challenges for researchers to effectively study the problem of bias e.g., doing it qualitatively/manually becomes impossible. More so, because these biases may be introduced at different stages in the development cycle of LLMs (e.g., in the choice of training data, objective function, or task definition, see (Sun et al., 2019; Hovy & Prabhumoye, 2021). On top of that, researchers face many conceptual problems, because of the socio-technical nature of algorithmic bias: Rather than it being a purely statistical phenomenon, defining bias relies on the normative questions and socio-technical context at hand (e.g., linguistic, cultural, legal; see e.g., (Blodgett et al., 2020; Talat et al., 2022).

In fact, while it is tempting to view ML models as simply encoding pre-existing biases from the data, the adoption of ML models in real-world applications may be seen as part of socio-technical systems themselves, which introduce new biases and harms, and even reshape norms and values as society adapts to these changes (see e.g., the discussion by Kidd and Birhane, 2023). For instance, the use of applications like ChatGPT by medical practitioners may propagate harmful race-based medicine (Omiye et al., 2023). The problem of bias, therefore, requires an interdisciplinary approach to tackle these challenges. This also means that a purely black-box (e.g., evaluating the model's behavior through prompting) or mechanistic understanding (e.g., studying what kind of representations are encoded in the model's parameters) of the model is not sufficient for tackling the questions of bias and harms.

Besides technical work on detecting and mitigating bias, we need other approaches, based on ethnographic and social science methods, and including as well cultural analyses. Studies like these would allow approaching ML models

as socio-technical systems, and treat them as part of a larger context: Who designs these systems? What are the social harms when using these? What are the power structures at hand? This is in line with a growing sentiment in the field of algorithmic bias, see e.g., (Margaret Mitchell et al., 2019; Blodgett et al., 2020; Bender et al., 2021; Talat et al., 2022). We believe this approach is equally important in understanding the biases of these models, which means that scientists do not only have a responsibility to work towards the transparency of their epistemic practices using ML models as we discussed in Sect. 15.3, but also to engage with a broader research and stakeholder community outside of their own field.

In sum, in our view, scientists have an even more important role in the scientific process now, since ML models are increasingly used in real-world systems affecting actual people. In combination with the increasing opacity of ML systems, this means that we have a high epistemic and moral responsibility and must play an active role in these practices.

15.5 The Place of Scientists in ML-Based Inquiries and the Missing Link in the 'in-betweenness' Chain

We build here on the previous section with some philosophical considerations about the possible displacements of humans in general and scientists in particular. Our arguments are partly descriptive and partly normative. On the descriptive side, detailed and informative reconstructions of ML practices show that humans are not entirely displaced. For one thing, this is not a question requiring a yes or no answer: Rather, the role of the human epistemic agent is changing rather than diminished. In certain tasks, humans are partly displaced (e.g., in feature engineering as discussed in Sect. 15.3), but other responsibilities and tasks have emerged as a consequence such as the interpretation of models (explainability) and ensuring the 'fairness' and reliability of the scientific practices. On the normative side, our point is that mere presence of humans is not enough to conclude that they are not displaced, because we should require that ML is practiced with a high-level of self-reflection related to design choices, as we now argue.

In the practice of machine learning, there are many translation steps that need to be performed by the researchers. For a certain task, data must be collected (often annotated as well), an objective function defined, a model designed and trained, and finally its performance evaluated. Often, researchers also need to fine-tune a pre-trained model to adapt it for a specific domain or task it was not originally designed for. At each of these steps, the human epistemic agents have to make numerous decisions about what is needed for their epistemic goals.

While in the early days of NLP with ML, researchers had to design the whole 'pipeline', they now often rely on 'pre-trained' language models. As also discussed in Sect. 15.3.1, others with often lots of resources train a model on an unsupervised language modeling task, which researchers can then use as the basis for their own

specific use-case. NLP researchers are more and more relying on these and other 'abstractions', allowing them to focus on the particular research question at hand. However, this also means that researchers accept the design decisions others have made by proxy (including normative ones).

Yet, at the same time, these abstractions also serve as bottlenecks, and the engineering of most low-level computational aspects of the ML pipeline (e.g., writing GPU drivers, understanding the computational hardware; see also Anthony et al. (2024)) becomes important to facilitate the creation of large ML models. This is a skill that is not very prevalent, especially not among researchers. Together with the larger demand of (computational) resources, we find that institutions have come to play a central role. This is not to say that researchers have no choice in the face of technological determinism, as many institutions committed to 'open science' have sprung up to counter the initially closed-source tech-corporation dominated field (see e.g., BigScience,[5] EleutherAI,[6] Aya[7]). Hence, researchers have a choice to engage with the organizations that align with their own principles and choose their ML models accordingly. Institutions are now playing a crucial role in ML research, and while smaller labs and individual researchers cannot build these larger ML systems themselves anymore, they still have a say in which institutions and their models to engage with.

It is now useful to return to the idea of in-betweenness introduced in Sect. 15.1. We make two interrelated points. First, in the third order relations (technology-technology-technology), the absence of humans is not real but only apparent—this follows from applying a practice approach descriptively; second, the presence or absence of humans in ML practices can also be a *choice* made at design level, which is a normative consideration.

Descriptively, humans may be displaced in the practice of ML, for instance in the use of a ML algorithm for specific purposes. Not quite: somebody *has* designed this algorithm and is using it. At times, these human scientists are not very visible because they are part of large research consortia or institutions. A key feature of ML is the degree of autonomy and agency we wish to give such algorithms is in fact a *design choice*. So, not only are we still in the loop, but we can decide and control *how much* we want to be in the loop—all these design choices require a high level of self-reflection. To reconsider the schematization of third order relations, the situation would look like this (Fig. 15.1).

Our proposed schematization of the third order relations explicitly shows that human epistemic agents or scientists are still connected in various ways to technological artifacts, even to technologies that can be said to have certain aspects of agency and that are more and more autonomous from us.

By using episodes from NLP in cognitive neuroscience, we have been able to show that we need a more nuanced understanding of 'human displacement'.

[5] https://bigscience.huggingface.co/.

[6] https://www.eleuther.ai/.

[7] https://cohere.com/research/aya.

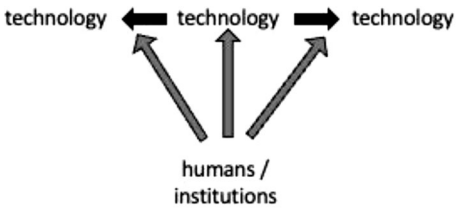

Fig. 15.1 Third-order technology relations including humans and institutions

Scientists do not simply 'disappear' when we engage more intensively with ML-based investigations in science. On the contrary, the NeuroAI episode illustrates how the epistemic use of ML technologies can be seen as a deliberate design decision by researchers that opens up new avenues for scientific inquiry. It is a modeling strategy and a deliberative decision by the scientists in charge, whose goal is to predict human brain activity. This can be described as a different approach to the problem of studying the human brain and demonstrates the potential epistemic benefits of using ML in science.

Also, when we look at the outputs of ML in ML-based scientific inquiries, we were able to emphasize where the place of humans is: Since there is social bias and scientific outputs can be influenced by it, humans are responsible for actively staying in the loop to check and control, e.g., by implementing fairness principles such as trust, reliability and explainability and investigating the bias of these models.

There is another important way in which human epistemic agents are not displaced, but play a key role and "new" places are created for them. Research and the use of ML techniques take place in institutional contexts where there are often protocols and guidelines for the correct, ethical and fair conduct of research. This is not so much about individuals, but rather about epistemic agents qua scientific communities, qua roles in institutions and organizations, and about the institutions and organizations themselves.

15.6 Concluding Remarks

In this chapter, we were tasked to explore the question whether the use of ML techniques displaces human epistemic agents in the process of scientific inquiry, thus leaving little or no roles to humans. We agree that ML does indeed change the mode of scientific inquiry. But the point is: This is not a change that we passively go along with. Rather, it is a change that we initiate and that we choose. Since we are designers and users of ML, we must ask ourselves the question: How many aspects of agency and autonomy do we want to grant to artificial systems? Our answers to this question will, in turn, determine the extent to which we are willing to be displaced in ML-based scientific practices. In other words, our displacement is not automatic or inevitable, but largely a choice. We need to stay in control over what we want ML to do for us, while accepting that we cannot control everything. This is

another reason to design responsibly and to guide the entire process of developing, implementing, using and evaluating ML. Of course, this is all an iterative process, as human epistemic agents must remain in the loop of the ML-based epistemic enterprise.

Acknowledgements We are very grateful to the editors for the opportunity to contribute to this volume and for their comments on an earlier draft. Research for this paper has been supported by VW-Stiftung, 2021–2024, support by Volkswagen Stiftung grant Az:97721 (S.A. Styger), NWO Gravitation Grant 024.001.006 to the Language in Interaction Consortium (M.de Heer Kloots), NWO, Open Competition Digitalisation under project number 406.DI.19.059 (O. vd Wal).

References

Alvarado, R. (2025). Challenges for computational reliabilism: Epistemic warrants, endogeneity and error-based opacity in machine learning. In J. M. Durán & G. Pozzi (Eds.), *Philosophy of science for machine learning: Core issues and new perspectives*. Synthese Library, Springer.

Anthony, Q., Hatef, J., Narayanan, D., Biderman, S., Bekman, S., Yin, J., Shafi, A., Subramoni, H., & Panda, D. (2024). The case for co-designing model architectures with hardware. arXiv. https://doi.org/10.48550/arXiv.2401.14489.

Aquino, Y. S. J., Carter, S. M., Houssami, N., Braunack-Mayer, A., Win, K. T., Degeling, C., Wang, L., & Rogers, W. A. (2023). Practical, epistemic and normative implications of algorithmic bias in healthcare artificial intelligence: A qualitative study of multidisciplinary expert perspectives. *Journal of Medical Ethics, 51*(6), 420–428. https://doi.org/10.1136/jme-2022-108850

Beisbart, C. (2025). In which ways is machine learning opaque? In J. M. Durán & G. Pozzi (Eds.), *Philosophy of science for machine learning: Core issues and new perspectives*. Synthese Library, Springer.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (FAccT '21) (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bishop, C. M. (2006). *Pattern recognition and machine learning. Information science and statistics*. Springer. https://link.springer.com/book/9780387310732

Blodgett, S. L., Barocas, S., Hal Daumé, I. I. I., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 5454–5476). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., et al. (2022). On the opportunities and risks of foundation models. arXiv. https://doi.org/10.48550/arXiv.2108.07258.

Boon, M. (2020). How scientists are brought back into science—The error of empiricism. In M. Bertolaso & F. Sterpetti (Eds.), *A critical reflection on automated science* (Vol. 1, pp. 43–65). Springer International Publishing. https://doi.org/10.1007/978-3-030-25001-0_4

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., et al. (2022). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences, 46*, 1–74. https://doi.org/10.1017/S0140525X22002813

Buijsman, S. (2025). Machine learning models as mathematics. In J. M. Durán & G. Pozzi (Eds.), *Philosophy of science for machine learning: Core issues and new perspectives*. Synthese Library, Springer.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology, 5*(1), 1–10. https://doi.org/10.1038/s42003-022-03036-1

Chang, H. (2014). Epistemic activities and systems of practice: Units of analysis in philosophy of science after the practice turn. In L. Soler, S. Zwart, M. Lynch, & V. Israel-Jost (Eds.), *Science after the practice turn in the philosophy, history, and the social studies of science* (Routledge studies in the philosophy of science) (pp. 67–79). Routledge, Taylor & Francis Group.

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences, 23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Crook, B., & Kästner, L. (2025). Don't fear the bogeyman: On why there is no prediction-understanding trade-off for deep-learning in neuroscience. In J. M. Durán & G. Pozzi (Eds.), *Philosophy of science for machine learning: Core issues and new perspectives*. Synthese Library, Springer.

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., et al. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience, 24*(7), 431–450. https://doi.org/10.1038/s41583-023-00705-w

Durán, J. M. (2023). Machine learning, justification, and computational reliabilism.

Durán, J. M. (2025). Beyond transparency: Computational reliabilism as an externalist epistemology for algorithms. In J. M. Durán & G. Pozzi (Eds.), *Philosophy of science for machine learning: Core issues and new perspectives*. Synthese Library, Springer.

Durán, J. M., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines, 28*(4), 645–666. https://doi.org/10.1007/s11023-018-9481-6

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics, 47*(5), 329–335. https://doi.org/10.1136/medethics-2020-106820

Emery, F. E., Trist, E. L., Churchman, C. W., & Verhulst, M. (1960). Socio-technical systems. In C. W. Churman & M. Verhulst (Eds.), *Management science models and techniques* (Vol. 2, pp. 83–97). Pergamon.

Eva, B., Ried, K., Müller, T., & Briegel, H. J. (2023). How a minimal learning agent can infer the existence of unobserved variables in a complex environment. *Minds and Machines, 33*(1), 185–219. https://doi.org/10.1007/s11023-022-09619-5

Ezenkwu, C. P., & Starkey, A. (2019). Machine autonomy: Definition, approaches, challenges and research gaps. In K. Arai, R. Bhatia, & S. Kapoor (Eds.), *Intelligent computing* (Advances in intelligent systems and computing) (pp. 335–358). Springer International Publishing. https://doi.org/10.1007/978-3-030-22871-2_24

Floridi, L. (2016). *The 4th revolution: How the Infosphere is reshaping human reality*. Oxford University Press.

Formanek, N. (2025). How I stopped worrying and learned to love opacity. In J. M. Durán & G. Pozzi (Eds.), *Philosophy of science for machine learning: Core issues and new perspectives*. Synthese Library, Springer.

Freiesleben, T. (2025). Artificial neural nets and the representation of human concepts. In J. M. Durán & G. Pozzi (Eds.), *Philosophy of science for machine learning: Core issues and new perspectives*. Synthese Library, Springer.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM, 64*(4), 136–143. https://doi.org/10.1145/3433949

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems, 14*(3), 330–347. https://doi.org/10.1145/230538.230561

Guest, O., & Martin, A. E. (2023). On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior, 6*(2), 213–227. https://doi.org/10.1007/s42113-022-00166-x

Gyllingberg, L., Birhane, A., & Sumpter, D. J. T. (2023). The lost art of mathematical modelling. arXiv. https://doi.org/10.48550/arXiv.2301.08559.

Hall, M., van der Maaten, L., Gustafson, L., Jones, M., & Adcock, A. (2022). A systematic study of bias amplification. arXiv. https://doi.org/10.48550/arXiv.2201.11706.

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron, 105*(3), 416–434. https://doi.org/10.1016/j.neuron.2019.12.002

Hey, A. J. G., Tansley, S., & Tolle, K. M. (Eds.). (2009). *The fourth paradigm: Data-intensive scientific discovery* (Vol. 1). Microsoft Research Redmond. https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/

Hooker, S., Moorosi, N., Clark, G., Bengio, S., & Denton, E. (2020). Characterising bias in compressed models. arXiv. https://doi.org/10.48550/arXiv.2010.03058.

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass, 15*(8), e12432. https://doi.org/10.1111/lnc3.12432

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. https://doi.org/10.1007/s11229-008-9435-2

Ivanova, A. A., Hewitt, J., & Zaslavsky, N. (2021). Probing artificial neural networks: Insights from neuroscience. arXiv. https://doi.org/10.48550/arXiv.2104.08197.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing.* (3rd ed.) https://web.stanford.edu/~jurafsky/slp3/

Kar, K., Kornblith, S., & Fedorenko, E. (2022). Interpretability of artificial neural network models in artificial intelligence versus neuroscience. *Nature Machine Intelligence, 4*(12), 1065–1067. https://doi.org/10.1038/s42256-022-00592-3

Kidd, C., & Birhane, A. (2023). How AI can distort human beliefs. *Science, 380*(6651), 1222–1223. https://doi.org/10.1126/science.adi0248

Kieval, P. H. (2025). Representation learning without representationalism. A non-representationalist account of deep learning models in scientific practice. In J. M. Durán & G. Pozzi (Eds.), *Philosophy of science for machine learning: Core issues and new perspectives*. Synthese Library, Springer.

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience, 21*(9), 1148–1160. https://doi.org/10.1038/s41593-018-0210-5

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., & Nastase, S. A. (2023). Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. bioRxiv. https://doi.org/10.1101/2022.06.08.495348.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Pritzel, A., & Ravuri, S., et al. (2022). GraphCast: Learning skillful medium-range global weather forecasting. arXiv. https://doi.org/10.48550/arXiv.2212.12794.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (XAI)?—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence, 296*(July), 103473. https://doi.org/10.1016/j.artint.2021.103473

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

Latour, B., & Woolgar, S. (1986). In J. Salk (Ed.), *Laboratory life: The construction of scientific facts*. Princeton University Press. https://press.princeton.edu/books/paperback/9780691028323/laboratory-life

Liu, P., Yuan, W., Jinlan, F., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), 195:1–195:35. https://doi.org/10.1145/3560815

Lloyd, K. (2018). Bias amplification in artificial intelligence systems. arXiv. https://doi.org/10.48550/arXiv.1809.07842.

Ma, W. J., & Peters, B. (2020). A neural network walks into a lab: Towards using deep nets as models for human behavior. arXiv. https://doi.org/10.48550/arXiv.2005.02181.

Meijer, G. (2021). Neurons in the mouse brain correlate with Cryptocurrency Price: A cautionary tale. PsyArXiv. https://doi.org/10.31234/osf.io/fa4wz.

Mitchell, M. (2020). *Artificial intelligence: A guide for thinking humans*. Penguin. https://www.penguin.co.uk/books/294649/artificial-intelligence-by-mitchell-melanie/9780241404836

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (FAT* '19) (pp. 220–229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., & Grover, A. (2023). ClimaX: A foundation model for weather and climate. arXiv. https://doi.org/10.48550/arXiv.2301.10343.

Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., & Daneshjou, R. (2023). Large language models propagate race-based medicine. *npj Digital Medicine, 6*(1), 1–4. https://doi.org/10.1038/s41746-023-00939-z

Páez, A. (2025). Axe the X in XAI: A plea for understandable AI. In J. M. Durán & G. Pozzi (Eds.), *Philosophy of science for machine learning: Core issues and new perspectives*. Synthese Library, Springer.

Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: Implications for health systems. *Journal of Global Health, 9*(2), 010318. https://doi.org/10.7189/jogh.09.020318

Pickering, A. (Ed.). (1992). *Science as practice and culture*. University of Chicago Press.

Poliseli, L., Coutinho, J. G. E., Viana, B., Russo, F., & El-Hani, C. N. (2022). Philosophy of science in practice in ecological model building. *Biology & Philosophy, 37*(4), 21. https://doi.org/10.1007/s10539-022-09851-4

Poliseli, L., & Russo, F. (2022). Philosophy of science in practice and weak scientism together apart. In M. M. Mizrahi (Ed.), *For and against scientism: Science, methodology, and the future of philosophy* (Ch.6. Collective studies in knowledge and society). Rowman & Littlefield.

Räz, T. (2025). From explanations to interpretability and Back. In J. M. Durán & G. Pozzi (Eds.), *Philosophy of science for machine learning: Core issues and new perspectives*. Synthese Library, Springer.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics, 8*(January), 842–866. https://doi.org/10.1162/tacl_a_00349

Russell, S., & Norvig, P. (2021). Artificial intelligence, global edition. https://elibrary.pearson.de/book/99.150005/9781292401171

Russo, F. (2022). *Techno-scientific practices: An informational approach*. Rowman & Littlefield.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative Modeling converges on predictive processing. *Proceedings of the National Academy of Sciences, 118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Soler, L., Zwart, S., Lynch, M., & Israel-Jost, V. (2014). *Science after the practice turn in the philosophy, history, and social studies of science. Routledge studies in the philosophy of science*. Routledge, Taylor & Francis Group.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1630–1640). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1159

Sutton, R. (2019). The bitter lesson. *Incomplete Ideas* (blog). 13 March 2019. http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Talat, Z., Névéol, A., Biderman, S., Clinciu, M., Dey, M., Longpre, S., Luccioni, S., et al. (2022). You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5—Workshop on challenges & perspectives in creating large language models* (pp. 26–41). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.bigscience-1.3

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story Reading subprocesses. *Edited by Kevin Paterson. PLoS ONE, 9*(11), e112575. https://doi.org/10.1371/journal.pone.0112575

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience, 19*(3), 356–365. https://doi.org/10.1038/nn.4244

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zednik, C., & Verreault-Julien. (2025). Can explainable AI contribute to justification? In J. M. Durán & G. Pozzi (Eds.), *Philosophy of science for machine learning: Core issues and new perspectives*. Synthese Library, Springer.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on empirical methods in natural language processing* (pp. 2979–2989). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1323

Sahra A. Styger has been an academic staff member and PhD candidate at the Department of Philosophy at the University of Konstanz, Germany, since 2021. As a focus for her PhD thesis, she studied epistemic agents and scientific products in ML-based science. Her research specifically addresses the foundations of ML-based science from a philosophy of science and epistemology perspective.

Marianne de Heer Kloots is a PhD candidate at the Institute for Logic, Language and Computation at the University of Amsterdam, the Netherlands, since 2021. Her research focuses on interpreting deep learning models for audio and text processing, and using them in models of human speech and language comprehension.

Oskar van der Wal is a PhD candidate at the Institute for Logic, Language and Computation at the University of Amsterdam, the Netherlands, since 2020. His research broadly focuses on understanding why and how language models exhibit social biases using interpretability tools. He is particularly interested in how we can reliably measure and mitigate bias, and the question of how to ground current evaluation practices in real-world harm. The views expressed in this work do not represent any undisclosed current or future affiliations.

Federica Russo is Professor of Philosophy and Ethics of Techno-Science and Westerdijk Chair at the Freudenthal Institute, Utrecht University. Her research concerns epistemological, methodological, and normative aspects they arise in the health and social sciences, with special attention to policy contexts and to the highly technologized character of these fields. She is the author of *Techno-Scientific Practices. An Informational Approach* (RLI, 2022).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.