The Epistemic Alignment Problem of Machine Learning*

Florian J. Boge
Institute for Philosophy and Political Science, TU Dortmund University, Dortmund, Germany
• Lamarr Institute for Machine Learning and Artificial Intelligence
• Emmy Noether Group *UDNN: Scientific Understanding and Deep Neural Networks*

florian-johannes.boge@tu-dortmund.de

Abstract: The value alignment problem of Machine Learning is the problem of properly aligning the objectives we put into ML systems with human values. I argue that deployments of Machine Learning in research contexts create an analogous, Epistemic Alignment Problem. Given a distinction from epistemology between epistemically final and instrumental values, this problem can be seen to have two levels: In level one, an ML system is consciously misaligned with an epistemically final value to prioritize an instrumental one. In level two, it is inadvertently misaligned with an epistemically final value. I argue that only level two should truly worry us.

Keywords: epistemic values • machine learning • alignment problem

1 Introduction

The *value alignment problem* (VAP) of Machine Learning (ML) was introduced by Russell (2019) as the problem of aligning the objectives we put into ML systems with human values. Here,

> values are natural or non-natural facts about what is good or bad, and about what kinds of things ought to be promoted. This normative sense of value differs significantly from the notion of value applied to goods or commodities in market contexts. (Gabriel, 2020, 422)

Thus, the VAP is a problem of ML *ethics*. But in pursuing science with ML, scientists too instruct machines with objectives. Hence, is there an analogous problem for the philosophy of *science*?

Since the days of Rudner (1953), we know that moral and social values have a say in the practice of science. Others, such as Levi (1962, 49), however, urged that scientists only need to focus on "other desiderata such

*This is a preprint of a paper accepted for presentation in the 30th Biennial Meeting of the Philosophy of Science Association (PSA2026).

as simplicity, explanatory power, etc."; desiderata that later came to be known as *epistemic values*. Whether we believe that social and moral values are inescapably intertwined with scientific practice (Douglas, 2009; Longino, 1990) or that they can be marginalized by focusing on suitably hedged scientific claims (Betz, 2013): The question of whether usages of ML in science are aligned with *epistemic* values is a different one from the one whether they are aligned with *moral* values (also Lacey, 2005, 89 ff.). In this paper, I will address what I call the *Epistemic Alignment Problem* (EAP) of ML: The problem of aligning the objectives we put into the ML systems deployed in research contexts with the epistemic values we uphold in these contexts.

As I shall argue, this problem has two levels, only one of which is indeed an epistemically serious problem: In *level one* epistemic misalignment, an ML system is consciously misaligned with epistemically *final* values – values that constitutes candidate *aims* of science – but still aligned with *other* epistemic values traditionally in the focus of philosophers of science. Level one misalignment does pose problems, but these are not epistemically daunting. In *level two* misalignment, however, an ML system is *inadvertently* misaligned with epistemically final values. Level two misalignment, I argue, is what we should be worried about. I will support this reasoning by two case-studies illustrating the differences.

The structure is this: In section 2, I consider the VAP in more detail. In section 3, I revisit the debate on epistemic values and borrow epistemology's distinction between epistemically final and instrumental values. In section 4, I apply this distinction to define the EAP and to explain its two levels. I there also show why level two is epistemically worse than level one.

2 A Closer Look at the Value Alignment Problem

Russell and Norvig (2021, 38–9) define the VAP as follows:

Value Alignment Problem (VAP): The objectives put into ML systems must be aligned with human values.

Accordingly, value *misalignment* may be defined as follows:

Value Misalignment (VM): The objectives put into ML systems are not aligned with human values.

As we saw, 'values' is to be primarily interpreted in terms of moral values here; but how are 'objectives

put into a ML systems'? An early description of the VAP was given by Russell in an interview with *Quanta Magazine*, where he explained it as follows:

> You build a system that's extremely good at optimizing some utility function, but the utility function isn't quite right. (as cited in Wolchover, 2015, 9)

Similarly, Russell and Norvig (2021, 1830–1) write:

> We run into trouble when a utility function fails to capture background societal norms about acceptable behavior.

Thus, 'objectives put into an ML system' are conceptualized as specified by a utility function that the ML system optimizes. In most ML applications, this is expressed "as a negative: to minimize a loss function rather than maximize a utility function." (Russell and Norvig, 2021, 1235) Hence, without loss of generality, we may think of the objective we put into a machine as being specified by a *loss function*, quantifying the distance between some property of an ML system and a desired property of that system. For example, if we train a system with deep reinforcement learning, the loss function will specify the expected maximum difference of some quality function Q at a state s brought about by some action a relative to a configuration of a neural network to the Q at a sought-for reference state and corresponding neural-net configuration, offset by some primitive reward, r (Mnih and Kavukcuoglu, 2017). But how can we know whether r and Q appropriately represent what we want from the machine, especially in the long run?

Two things are noteworthy here: First, a misalignment with human values may also come about in different ways than a loss function being set-up inappropriately; say, through such things as architectural constraints that make it impossible for the machine to achieve the objective set forth, or because the distribution of the data trained on is biased in such a way that the ML system inherits this bias. Second, the VAP was primarily conceptualize by Russell as a problem about *future*, catastrophic societal risks. However, of course there also current issues arsing from VM.

A case exemplifying both these features – present-day misalignment, and not specifically due to an inappropriate loss function – is the recent case of the recidivism prediction software COMPAS that inherited racial biases from training on past court decisions (Biddle, 2022). For the purposes of this paper, it will be

convenient to focus on such present-day misalignment. However, it will also be helpful to give specific consideration to the loss function in the cases considered below, as this make the misalignment at stake obvious. Thus, while the loss function is not the *only* source of misalignment, it is still an important one.

3 Epistemic Values Reconsidered

3.1 Why Value Epistemic Values?

To properly set up the epistemic alignment problem, it is necessary to look into the topic of epistemic values. A classic reference is Kuhn (1977), where he listed only five values: accuracy, consistency, scope, simplicity, and fruitfulness. This list has been extended by such things as testability, non-*ad hoc*-ness, symmetry, visualizability, and conservativeness (Schindler, 2018, 5), or explanatory and modal power (Lacey, 2005, 59).

Why should such values be upheld? In some cases, a direct connection to *truth* might be assumed, depending on the interpretation of the term at stake. Assuming that truth is what we aim for in science, this would make the valuability of epistemic values understandable. For instance, in measurement theory, 'accuracy' is distinguished from 'precision', and sometimes defined by reference to the closeness of measurements to the *true* value, whereas the latter means a narrow spread between repeated measurement-values with the same method (Tal, 2011, 1085). But there are readings of 'accuracy' that do not presuppose a direct connection to truth (*ibid.*, 1084–5). As the example shows, it is unclear whether there is an intimate connection between epistemic values and truth (also Laudan, 2004).

In the context of ML, this is exacerbated: Values such as simplicity, scope, and non-*ad hoc*-ness are rejected exactly in the service of truth. The Nobel prize-winning model AlphaFold2, e.g., is ostensibly complex, hyper-specialized to protein-prediction, and based on a fit to data – a form of *ad hoc*-ness. Nevertheless, it has given us access to accurate models of protein structures. Nobody would bemoan the loss of simplicity, scope, and non-*ad hoc*-ness as indicating that the predictions are unlikely to be true: A better story is needed.

A first attempt to make sense of the foregoing might run as follows: Douglas (2009) suggests that not all so-called epistemic values should indeed count as epistemic. In a parlance introduced by Lacey (2005), they might rather be called 'cognitive'. Douglas (2009, 93) defines cognitive values as "those aspects of scientific

work that help one think through the evidential and inferential aspects of one's theories and data." They are, in other words, mere aids to (good) thinking. In contrast, epistemic values are "basic criteria that any scientific work must meet" (Douglas, 2009, 93); they are "about the ultimate goal of research" (Douglas, 2009, 33).

What are examples of cognitive and epistemic values, respectively? As cognitive values, Douglas (2009, 93) lists the usual features, such as simplicity, scope, fruitfulness, or external consistency (i.e., consistency with other theories). Under epistemic values, she only lists internal consistency and predictive competence (i.e., the ability to provide accurate predictions). This is an ostensibly sparse list, but can we at least make sense of cases like AlphaFold2 by appeal to the distinction between cognitive and epistemic values?

It is certainly right that a simpler model with a greater scope, i.e., one which also delivers predictions and maybe even explanations for phenomena related to protein structure, would be preferable to a complex, opaque model like AlphaFold2. Thus, we might still find these features cognitively valuable. And it is also right that AlphaFold2's value depends entirely on its predictive capabilities, and that it would be rejected if it was set up in such a way as to deliver contradictory information on certain protein structures. Hence, Douglas' distinction clearly gets something right. However, the distinction is incapable of explaining why we happily sacrifice certain cognitive values if this gets us faster to accurate predictions – and thus maybe closer to the *truth*. After all, AlphaFold2's predictive accuracy was considered worthy of a Nobel prize.

3.2 Epistemically Final and Instrumental Values

The situation can be better understood by appeal to a different distinction, one from epistemology, between epistemically *instrumental* and epistemically *final* values.

Thus, Pritchard (2010, 11; orig. emph.) suggests to call something "a fundamental epistemic good" if its "epistemic value is at least sometimes not simply instrumental value relative to a further *epistemic* good." If an epistemic good is fundamental, then its value is epistemically *final* – if it is not fundamental, then its value is epistemically *instrumental* (*ibid.*, 12–13).[1]

What are candidate final epistemic values? Douglas (2009, 33) held epistemic values properly-so-called to be "about the ultimate goal of research", but 'being about something' is not the same as 'being something'.

[1] Whether epistemic value is irreducible, or whether we endorse even epistemically final values for the sake of furthering practical ends (or *moral* goods) is an open question.

Thus, if our theories were merely internally consistent and predictively competent, we would not be satisfied. However, Douglas' suggestion to turn to "the ultimate goal of research" for determining epistemic value is helpful in any case: If we want to understand final epistemic value in science, we should turn to conceptions of the *aim* of science. Discussions about science's aim have been intimately intertwined with discussions about progress: That whichever defines the aim of science determines when science makes progress (Rowbottom, 2023, 32). Candidate aims of science are knowledge (Bird, 2007), truth (Cevolani and Tambolo, 2013; Niiniluoto, 2014), and understanding (de Regt, 2017; Dellsén, 2021; Elgin, 2017).

I will only consider these as candidates for epistemcially *final* values in the context of science.[2] All other so-called cognitive or epistemic values will be considered epistemically *instrumental*: if a theory is internally inconsistent or predictively incompetent, then it seems ill-posed to guide us to *truth* or *knowledge*.[3] Similarly, a violation of many of Douglas' cognitive values stands obviously in the way of achieving *understanding*: If a theory is too complex, it will be incapable of promoting understanding, given human cognitive limitations (de Regt and Gijsbers, 2016; Elgin, 2017); if it is narrow in scope, it will not extend our grasp of how things in one domain hang together (Elgin, 2017; Strevens, 2013); if it is not externally consistent, it won't extend our grasp of how they hang together with things in other domains; if it is not explanatorily powerful, it falls short of providing the main vehicle for scientific understanding (de Regt, 2017; Strevens, 2013).

Note that the instrumentality of cognitive values is in general not deterministic: The less complex theory or model will be more *likely* to promote understanding. But there is also a sense in which ostensibly complex theories like the Standard Model of particle physics have improved our understanding of reality more than Newton's simpler laws.

The relations between candidate epistemically final values are notoriously complex: Knowledge implies truth, but it is unclear whether it reduces to truth + X (Williamson, 2002). The 'swamping problem' (Kvanvig, 2003) says that is also not unequivocally clear why knowledge should have added value over truth. Understanding is generally assumed to involve truth, but an understanding-promoting theory or model may tolerate – even require – many falsehoods (de Regt and Gijsbers, 2016; Elgin, 2017; Strevens, 2013). Finally, some

[2]There are also conceptions of progress that prioritize problem-solving (Kuhn, 1970) or justification (Stegenga, 2024), but I will bracket these here.

[3]Thought neither consistency nor predictive competence need be *necessary* for truth (Laudan, 1981; Priest, 2006).

view understanding as a special kind of knowledge (Grimm, 2006; Kelp, 2015; Khalifa, 2017), others disagree (Hills, 2016; Wilkenfeld, 2016).

Given the state of these debates, it seems reasonable that we might value knowledge, truth and understanding all for their own sake. But how can we choose which one to prioritize? I suggest that this is a matter of context. For example, if one is in a stage of research where new, likely true hypotheses about a range of phenomena are to be discovered (thought they still fall short of constituting *knowledge*), then this stage finds a natural endpoint once a plausible, candidate *true* hypothesis has been found. Thus, relative to that research-context, truth is epistemically final. If one's research is instead experimental and exploratory (Steinle, 2016), one might seek to discover novel *phenomena*, and the research is concluded once such a phenomenon has been firmly established. Relative to that context, *knowledge* will be epistemically final (Bird, 2023, 17).[4] In the following, I will hence treat epistemic finality as thus relativized to a research context. Below, I will speak of epistemically final values being *prioritized by the context*.

4 The Epistemic Alignment Problem

4.1 Epistemic (Mis-)Alignment

Epistemic misalignment broadly occurs when the objectives put into an ML system in a research context do not align with the epistemic values endorsed in this research context. In the previous section, I argued that this only becomes interesting when there is misalignment with epistemically *final* values. Furthermore, I argued that the research context selects the epistemically final values to prioritize. Thus, the epistemic alignment problem may be defined as follows:

Epistemic Alignment Problem (EAP): The objectives put into an ML system deployed in a research context must be aligned with the epistemically final values prioritized by that research context.

For example, if the context is aimed at the discovery of novel phenomena, then the objectives put into an ML system must be aligned with the production of knowledge (Bird, 2023, 17). If, in contrast, the context is

[4]Defenders of monistic conceptions of science's aim could counter that some contexts are subordinate to others. For instance, a defender of knowledge as the aim of science might emphasize the context of justification over the context of discovery; a defender of understanding might argue that only the explanation of novel phenomena is final. I acknowledge these possibilities, but leave them aside, as they merely amount to an *extension* of my ideas.

aimed at establishing understanding, this would be insufficient.

As we saw, the objectives put into an ML model can be misaligned with epistemically instrumental values, such as simplicity, while still being aligned with some candidate epistemically final values, such as truth. Furthermore, they can be aligned with some epistemically final values while being misaligned with others: In the case of AlphaFold2, we saw that it increases our knowledge of protein structures. But researchers bemoan that it "has not led us to a deeper *mechanistic understanding* of exactly how a protein sequence folds" (Nussinov et al., 2022, 6372; emph. added). The first aspect has been handled here by a reference to the final / instrumental distinction, and by noting that epistemically instrumental values merely stand in a likelihood-increasing relation to epistemically final values, not a deterministic one. The second may be handled by reference to the contextual nature of epistemically final values, but this will require further attention below.

We can now see that (at least) two types of misalignment may arise under these circumstances: In the first one, the epistemically final value of the context is consciously neglected and an *instrumental* value prioritized. This may (but need not) be accompanied by the achievement of an epistemically final value *not* prioritized by the context. In the second type, the objectives put into the machine are *inadvertently* not aligned with the epistemically final value prioritized by the context. Since type two is, as we shall see, in a clear sense worse than the first one, I shall also speak of two *levels* of misalignment:

Level One Epistemic Misalignment (EM_1): The objectives put into an ML system deployed in a research context are consciously misaligned with the epistemically final value prioritized by that research context, and instead aligned only with an epistemically *instrumental* value.

Level Two Epistemic Misalignment (EM_2): The objectives put into an ML system deployed in a research context are *inadvertently* misaligned with the epistemically final value prioritized by that research context.

I shall illustrate these by two case studies, and in the course explain why EM_2 is worse.

4.2 Level One Misalignment

An example of EM_1 is the case of AlphaFold2. I acknowledged that AlphaFold2 may bring about knowledge, but with the revitalization of epistemic finality to research contexts, we can make sense of verdicts from

the structural biology-community bemoaning the want of a "deeper mechanistic understanding" (Nussinov et al., 2022, 6372). Thus, consider how the 'protein folding problem' is usually considered to contain three sub-problems (Schuster, 2026, 14): (i) Which forces determine the folded structure? (ii) How is the folded structure achieved so quickly? (iii) How can the folded structure be determined from the amino acid sequence? Arguably, (i) and (ii) inquire about rendering protein folding *understandable*, but only (iii) is solved by AlphaFold2 (also Schuster, 2026, 19).

If the folding problem really contains sub-problems (i)–(iii), then research into protein folding prioritizes (mechanistic) understanding. However, in setting up AlphaFold2, this relevance of mechanistic information is marginalized: AlphaFold2's first part, the 'evoformer', only looks at co-evolutionary information and arranges a geometric map (the 'pair representation'), building on evolutionary plausibility. The second part, the 'structure module', then uses this information to create a 3D-structure (Jumper et al., 2021a). In no stage does mechanistic information enter, or is explicitly enforced.

This is vivid also at the level of the loss function: The overall loss function is a superposition of different loss terms (Jumper et al., 2021b, 32 ff.). The main component is the so called 'frame aligned point error' which computes the mean distance between atoms in predicted and ground truth structures, relative to different frames. However, the other terms similarly prioritize geometric properties, or the ability to retrodict the evolutionary information from an intermediate stage. Nowhere is mechanistic information enforced.

This is a clear case of EM_1 if we take mechanistic understanding to be final in structural protein-biology, given parts (ii) and (iii) of the folding problem: The loss function prioritizes predictive accuracy, an epistemically instrumental value, but in no way enforces elements that bring about understanding; the epistemically final value in that context.

Nevertheless, EM_1 need not be terribly worrisome here for three reasons: First, as I already noted, the prioritization of predictive accuracy has at least in this case brought about other candidate epistemically final values, such as knowledge. Knowledge is not prioritized by the research context, but is still valuable in itself. Second, given that part of the folding problem, (iii), is fully defined as a prediction problem, one could argue that there is a narrower context in which knowledge is prioritized; AlphaFold2 would be well-aligned with

the epistemically final value prioritized by that narrower context.

Third, there is work by Roney and Ovchinnikov (2022) that suggests that AlphaFold2 might have tacitly learned an energy function: Roney and Ovchinnikov (2022) fed AlphaFold2 with one amino acid sequence at a time, next to a decoy protein-structure as template. They then investigated the correlation between AlphaFold2's confidence and the closeness of the decoy to a real structure. This correlation was very high but no co-evolution information could be used on a single sequence. They hence concluded that AlphaFold2 must have learned an energy function, "a function that has an optimum around the native structure and generally correlates with the probability that a protein sequence will adopt a given conformation" (Roney and Ovchinnikov, 2022, 2). If this is correct, AlphaFold2 has learned mechanistically relevant information while tasked to solve the prediction problem.

Of course, *extracting* that information from AlphaFold2 may be a formidable challenge (Boge, 2021, 2024; Buckner, 2020). It would also fall short of providing an outright mechanism. Nevertheless, without AlphaFold2, there might be no clue into what direction research into mechanistic models should go. Furthermore, research on extracting the contents of ML systems with symbolic regression could pave a way towards extracting mechanistic information (Wetzel, 2025). Hence, while EM_1 *is* a problem, it is not epistemically daunting.

4.3 Level Two Misalignment

EM_2 is exemplified by a different case that has interested several philosophers (Boge and De Regt, 2025; Koberinski, forth.): Particle physicists have tried to leverage ML models called 'autoencoders' to detect anomalous data with unsupervised learning, thus possibly finding signs of unknown physics. An autoencoder is a neural network that becomes slimmer towards the middle, and then builds up to the size of its input again. In this way, an autoencoder can be trained to reconstruct images, based on a compressed, latent representation in the slimmest layer. If such an autoencoder is trained on known images but then fails to reconstruct particular images correctly, this is an indication that these images have anomalous features that might escape human attention (Finke et al., 2021).

Autoencoders have thus been trained on images of simulated particle jets; conical streams of particles

produced as a result of the decay of other particles like quarks and gluons. Several results were very promising (Farina et al., 2020; Heimel et al., 2019), but a study by Finke et al. (2021) varied the kind of data taken as regular data and those taken as an anomaly-model. As it turned out, the simulated data normally used as the anomaly-model were in general more complex, in the sense of containing more bright pixels and having more structure. When Finke et al. (2021) exchanged the two types of data, the autoencoder could not detect anomalies anymore, as it now learned to reconstruct arbitrarily complex images.[5]

Finke et al. (2021) traced the issue back to the loss function, which was originally given by the mean square error-loss: Closer inspection revealed that this loss incentivizes the machine to focus on reconstructing only bright pixels, as this will decrease the mean square error faster (Finke et al., 2021, 7).

This is a case of EM_2, as the objective is *inadvertently* misaligned with the epistemically final value prioritized by the context: *Knowledge* of hitherto unknown phenomena. This is a major problem as an ML system set up in this fashion may create the illusion of achieving this knowledge, while in reality merely achieving the reconstruction of more or less complex images. We should be worried about EM_2 as it may take time to realize what objective the machine is *really* following – as was the case in the example.

5 Conclusions

I argued that when ML systems are deployed in research contexts, scientist face an *epistemic alignment problem*: The objectives put into the machine must be aligned with the epistemically final values prioritized by the research context. Misalignment may occur on two different levels: On the first level, researchers consciously misalign the ML system with an epistemically final value and prioritize an epistemically instrumental one instead. This can be problematic, but need not be daunting: Other epistemically final values can sometimes be achieved, prioritized by a narrower research context. Furthermore, a trained, successful model can sometimes be interpreted in such a way as to gives clues for reaching the epistemically final value originally prioritized.

The second level, however, is daunting: The ML system is here inadvertently misaligned with the epistemically final value prioritized by the context but successful deployment in limited studies may create the illusion

[5]These results were exacerbated by the use of different architectures in other studies (Buss et al., 2023; Dillon et al., 2023; Fraser et al., 2022).

that it is not. Furthermore, it may be difficult to detect this level of misalignment, whence researchers could be set on a wrong track for long time-scales. A recommendation that could be based on this is to consider the objectives put into a machine as carefully in scientific contexts as this should be done in contexts with high societal stakes.

Acknowledgments The research for this paper was generously funded by the German Research Foundation (DFG), through the project *UDNN: Scientific Understanding and Deep Neural Networks* (grant 508844757).

References

Betz, G. (2013). In defence of the value free ideal. *European Journal for Philosophy of Science*, 3(2):207–220.

Biddle, J. B. (2022). On predicting recidivism: Epistemic risk, tradeoffs, and values in machine learning. *Canadian Journal of Philosophy*, 52(3):321–341.

Bird, A. (2007). What is scientific progress? *Noûs*, 41(1):64–89.

Bird, A. (2023). The epistemic approach: Scientific progress as the accumulation of knowledge. In Shan, Y., editor, *New Philosophical Perspectives on Scientific Progress*, pages 13–26. Routledge.

Boge, F. J. (2021). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*, pages 1–33. https://doi.org/10.1007/s11023-021-09569-4.

Boge, F. J. (2024). Functional concept proxies and the actually smart hans problem: What's special about deep neural networks in science. *Synthese*, 203(1):16.

Boge, F. J. and De Regt, H. W. (2025). Machine learning discoveries and scientific understanding in particle physics: Problems and prospects. In *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*, pages 403–434. Springer.

Buckner, C. (2020). Understanding adversarial examples requires a theory of artefacts for deep learning. *Nature Machine Intelligence*, 2(12):731–736.

Buss, T., Dillon, B. M., Finke, T., Krämer, M., Morandini, A., Mück, A., Oleksiyuk, I., and Plehn, T. (2023). What's anomalous in lhc jets? *SciPost Physics*, 15(4):168.

Cevolani, G. and Tambolo, L. (2013). Progress as approximation to the truth: A defence of the verisimilitudinarian approach. *Erkenntnis*, 78(4):921–935.

de Regt, H. (2017). *Understanding Scientific Understanding*. Oxford University Press.

de Regt, H. W. and Gijsbers, V. (2016). How false theories can yield genuine understanding. In Baumberger, C., Grimm, S., and Ammon, S., editors, *Explaining understanding*, pages 50–75. Routledge.

Dellsén, F. (2021). Understanding scientific progress: The noetic account. *Synthese*, 199(3):11249–11278.

Dillon, B. M., Favaro, L., Plehn, T., Sorrenson, P., and Krämer, M. (2023). A normalized autoencoder for lhc triggers. *SciPost Physics Core*, 6(4):074.

Douglas, H. (2009). *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press.

Elgin, C. Z. (2017). *True Enough*. Cambridge MA, London: MIT Press.

Farina, M., Nakai, Y., and Shih, D. (2020). Searching for new physics with deep autoencoders. *Physical Review D*, 101(7):075021.

Finke, T., Krämer, M., Morandini, A., Mück, A., and Oleksiyuk, I. (2021). Autoencoders for unsupervised anomaly detection in high energy physics. *Journal of High-Energy Physics*. https://doi.org/10.1007/JHEP06(2021)161.

Fraser, K., Homiller, S., Mishra, R. K., Ostdiek, B., and Schwartz, M. D. (2022). Challenges for unsupervised anomaly detection in particle physics. *Journal of High Energy Physics*, 2022(3):1–31.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437.

Grimm, S. R. (2006). Is understanding a species of knowledge? *British Journal for the Philosophy of Science*, 57(3):515–535.

Heimel, T., Kasieczka, G., Plehn, T., and Thompson, J. (2019). Qcd or what? *SciPost Physics*, 6(3):030.

Hills, A. (2016). Understanding why. *Noûs*, 50(4):661–688.

Jumper, J. et al. (2021a). Highly accurate protein structure prediction with alphafold. *Nature*. https://doi.org/10.1038/s41586-021-03819-2.

Jumper, J. et al. (2021b). Supplementary information for: Highly accurate protein structure prediction with alphafold. *Nature Portfolio*. https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-021-03819-2/MediaObjects/41586_2021_3819_MOESM1_ESM.pdf.

Kelp, C. (2015). Understanding phenomena. *Synthese*, 192(12):3799–3816.

Khalifa, K. (2017). *Understanding, Explanation, and Scientific Knowledge*. Cambridge University Press.

Koberinski, A. (forth.). Searching high and low: precision measurement, machine learning, and experimental discovery in particle physics. *Synthese*. preprint: https://philsci-archive.pitt.edu/27809/.

Kuhn, T. S. (1970). *The structure of scientific revolutions*. University of Chicago Press, 2nd edition.

Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. In Kuhn, T. S., editor, *The Essential Tension*, pages 320–39. University of Chicago Press.

Kvanvig, J. L. (2003). *The Value of Knowledge and the Pursuit of Understanding*. Cambridge Studies in Philosophy. Cambridge University Press.

Lacey, H. (2005). *Is science value free?* Routledge.

Laudan, L. (1981). A confutation of convergent realism. *Philosophy of science*, 48(1):19–49.

Laudan, L. (2004). The epistemic, the cognitive, and the social. In Machamer, P. K. and Wolters, G., editors, *Science, Values, and Objectivity*, pages 14–23. University of Pittsburgh Press.

Levi, I. (1962). On the seriousness of mistakes. *Philosophy of Science*, 29(1):47–65.

Longino, H. (1990). *Science as Social Knowledge*. Princeton University Press.

Mnih, V. and Kavukcuoglu, K. (2017). Methods and apparatus for reinforcement learning. US Patent 9,679,258.

Niiniluoto, I. (2014). Scientific progress as increasing verisimilitude. *Studies in History and Philosophy of Science Part A*, 46:73–77.

Nussinov, R., Zhang, M., Liu, Y., and Jang, H. (2022). Alphafold, artificial intelligence (ai), and allostery. *The Journal of Physical Chemistry B*, 126(34):6372–6383.

Priest, G. (2006). *In Contradiction*. Oxford: Oxford University Press, 2nd expanded edition.

Pritchard, D. (2010). Knowledge and understanding. In Pritchard, D., Millar, A., and Haddock, A., editors, *The Nature and Value of Knowledge*, pages 3–90. Oxford: Oxford University Press.

Roney, J. P. and Ovchinnikov, S. (2022). State-of-the-art estimation of protein model accuracy using alphafold. *Phys. Rev. Lett.*, 129:238101.

Rowbottom, D. P. (2023). *Scientific Progress*. Elements in the Philosophy of Science. Cambridge University Press.

Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science*, 20(1):1–6.

Russell, S. (2019). *Human compatible: AI and the problem of control*. Penguin Uk.

Russell, S. and Norvig, P. (2021). *Artificial Intelligence: A Modern Approach (4th ed.)*. Prentice Hall series in artificial intelligence. Prentice Hall.

Schindler, S. (2018). *Theoretical virtues in science: Uncovering reality through theory*. Cambridge University Press.

Schuster, A. (2026). Understanding protein folding with machine learning models? the case of alphafold2. *Synthese*, 207(2):66.

Stegenga, J. (2024). Justifying scientific progress. *Philosophy of Science*, 91:543–560.

Steinle, F. (2016). *Exploratory Experiments: Ampère, Faraday, and the Origins of Electrodynamics*. University of Pittsburgh Press.

Strevens, M. (2013). No understanding without explanation. *Studies in history and philosophy of science Part A*, 44(3):510–515.

Tal, E. (2011). How accurate is the standard second? *Philosophy of Science*, 78(5):1082–1096.

Wetzel, S. J. (2025). Closed-form interpretation of neural network classifiers with symbolic gradients. *Machine Learning: Science and Technology*, 6(1):015035.

Wilkenfeld, D. A. (2016). Understanding without believing. In Baumberger, C., Grimm, S., and Ammon, S., editors, *Explaining understanding*, pages 318–334. Routledge.

Williamson, T. (2002). *Knowledge and its Limits*. Oxford University Press.

Wolchover, N. (2015). Concerns of an artificial intelligence pioneer. *Quanta Magazine*, April 21. https://www.quantamagazine.org/artificial-intelligence-aligned-with-human-values-qa-with-stuart-russell-20150421/.