

Understanding Artificial Neural Networks: Mysterianism about Known Mechanism is Mysticism

Olivia Guest^{1,2}, Nancy Abigail Nuñez Hernández³, and Mark Blokpoel^{1,2}

¹Department of Cognitive Science and Artificial Intelligence, Radboud University, The Netherlands

²Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, The Netherlands

³Facultad de Estudios Superiores Acatlán-Universidad Nacional Autónoma de México, Mexico

Mysterianism is the idea that human cognition, mind, cannot be understood. Taking this concept and applying it to *known* mechanism — such that claims are made that we do not know how engineered systems, such as artificial neural networks (ANNs), work, or that they constitute black boxes that we can only open with difficulty — is inappropriate at best and malicious at worst. We do know the mechanistic structure of such models because we designed and built them. We also do know their functional role (what they are for) as well as the mathematical function they are asked to approximate (map inputs to target outputs). Because mysterianist beliefs about known systems, such as ANNs, are often expressed, scientists need to sit up and take notice. We provide an error theory as to what is going on to help unpick this metatheoretical blunder. Ultimately, the problem is that ‘understanding’ is not a technical term in these cases: the word is co-opted for a specific narrative to sell ‘artificial intelligence’ through mystification. All computational systems, from pendulums to databases, will behave in ways we cannot predict or control — this is not a unique property of ANNs — and experts do indeed grasp the computational properties of these systems nonetheless.

Keywords: artificial neural networks; mechanism; black box; epistemology; artificial intelligence; understanding

Is it just hyperbole and advertising hype or, for all our pride in science and technology, do the wonders and terrors of sorcery still haunt our souls?

William A Stahl (1995, p. 234)

Cognitive science seeks inter alia mechanisms that explain cognition (Boden 2006; Cao and Raja 2024; Egan 2017; Krickel 2023; Martin and Doumas 2017; Miłkowski 2016; Von Eckardt and Poland 2004; Wright and Bechtel 2007). Under the aegis

 Olivia Guest

Acknowledgements: The authors would like to thank the *Computational Cognitive Science* group at the Donders Centre for Cognition for their invaluable feedback on this work, especially Natalia Scharfenberg. Thanks also go to Jed Brown, Mark Dingemanse, Dagmar Monett, and Katia Schwerzmann for comments; and to Iris van Rooij for extensive discussions, feedback, and analyses.

CRedit: Conceptualization: OG; Formal analysis: OG; Investigation: OG; Project administration: OG; Visualization: OG, MB; Writing (original draft): OG, NANH; Writing (review & editing): OG, NANH, MB.

Correspondence: Olivia Guest, Donders Institute for Brain, Cognition and Behaviour, Radboud University, The Netherlands. E-mail: olivia.guest@donders.ru.nl

of psychology and cognitive neuroscience, computationalist accounts of cognition often reign supreme. These accounts — models, theories, frameworks (Guest and Martin 2021) — have at their core the assumption that cognition can be understood as a form of computation (cf. van Rooij, Guest, et al. 2024). Furthermore, in the contemporary landscape of our fields, the mainstream view essentially acquiesces to, or vociferously promotes, connectionism for all types of modelling and theorising endeavours (Guest and Martin 2025b, 2023). All engineered models, connectionist and other computational models, comprise known mechanisms that we ourselves set into motion to capture some phenomenon or system (Guest 2023; Guest and Martin 2025a; Guest, Scharfenberg, et al. 2025).

In contrast to all this are the pronouncements of colleagues, technology industry employees, and journalists that proclaim that somehow artificial neural networks (ANNs), especially large language models (LLMs), and in general all artificial intelligence (AI) are *outside our current expert understanding*, often calling them a *black box* (see Box 1). If so, then what are we as scientists to do given that furthering understanding is our purview? And to add to that our supposed lack of understanding is presented to policy-makers and funders as rationale for research that is incompatible with what we know to be the case. Relatedly, contemporary

media discourse reflects past hype cycles, such as that of the personal computer in the 1980s (Stahl 1995), weaving in a modern belief about lack of understanding mechanisms that make up ANNs. *Mysterianism* is the belief that cognition cannot be understood in principle (Flanagan 1991). This may well be the case, especially if by ‘understanding’ we mean formally and/or computationally explicated (van Rooij, Guest, et al. 2024). However, applying mysterianist belief to *known* mechanisms, such as ANNs, seems like a very different claim altogether: a category error.

These mystifying or obfuscatory metaphors about mechanism (Christin 2020; Petrick 2020) bring to mind: “Any sufficiently advanced technology is indistinguishable from magic.” (Clarke 1973, p. 21; also noticed by Penn 2022) For example, these systems are described as: “powerful digital minds that no one — not even their creators — can understand, predict, or reliably control.” (by a mix of CEOs and founders from Elon Musk and Steve Wozniak, to academic experts in AI, like Yoshua Bengio and Stuart Russell, in *Pause Giant AI Experiments: An Open Letter*, Bengio et al. 2023). Others repeat this same mantra, with OpenAI’s Bills et al. (2023) claiming that:

we do not understand how [LLMs] work. [...] Our explanations also do not explain what causes behavior at a mechanistic level, which could cause our understanding to generalize incorrectly. To predict rare or out-of-distribution model behaviors, it seems possible that we will need a more mechanistic understanding of models. (emphasis added, n.p.)

Even critical scholars repeat this, such as Jobin and Katzenbach (2023):

we are living in a time when the infrastructures and institutions of our everyday lives are being (re)built at the hands of techniques that already elude popular and professional understanding. (emphasis added, p. 48)

Journalists like Perrigo (2024), also follow suit:

Today’s artificial intelligence is often described as a “black box.” [T]he inner workings of the AI models remain opaque, and efforts to peer inside them to check exactly what is happening haven’t progressed very far. Nobody truly understands what [the internals] mean, or how they work. (emphasis added, n.p.)

Relatedly, the media and the public are then — perhaps rightly — alarmed and follow up with questions such as: “Without a thorough (*mechanistic*) understanding of the early antecedents of these powerful future LLMs, can we realistically expect to manage these AIs[...]?” (emphasis added,

Box 1: Terminology for unpicking labels and conceptual precursors to the reasoning behind ‘ANNs are a black box or unknowable, even if opened’ depicted in Figure 1.

ANNs, LLMs, & AI

In modern parlance, in both scholarly and popular science and technology discussions, these three acronyms are often used interchangeably. Definitions of AI and LLM are hard to pin down — as AI is a marketing term for making research sound flashy, neither ‘intelligence’ nor ‘artificial’ are coherent scientific concepts, and the ‘large’ in LLM is not used with any given cut-off (Guest 2026; Guest, Suarez, et al. 2025). ANN is an umbrella term for models of the data or of cognition that involve matrix multiplications followed by squashing functions (Guest and Martin 2025b, 2023).

BLACK BOX

A phrase from cybernetics, used inter alia in engineering and computer science, that denotes a system or sub-system that we (aim to) understand only through examining its input-output mappings (Ashby 1956; Glanville 2009). This pure functionalist view on engineered systems does not preclude peeking inside for the ground truth after performing the black box analysis. On the contrary, for non-engineered systems there is no peeking inside — such a manoeuvre is not possible — as no specifications or latent human-designed methods were used to create the system (Guest, Scharfenberg, et al. 2025).

LEIBNIZ’ MILL

The Mill Argument, Leibniz’ Mill, or Leibniz’ Gap, is an argument against the idea that mechanical components suffice to understand cognition. Leibniz (1714/1989) asks us to imagine stepping inside a machine that produces cognition in the same way we can a mill, an exemplary machine. In so doing we observe the interactions of the mill’s cogwheels, levers, and other components. By looking at the mechanical constituents of the machine, Leibniz explains, there will never be anything we will come across that explains cognition (Cummins 2000, 2010; Duncan 2012; Hattab 2011; Kulstad and Carlin 2020; Rozemond 2014, 2019).

Sullivan 2023)¹ Behind these confusions and questions lies a core category error: that we do not understand ANNs, LLMs, and AI generally (see Box 1).

As an academic community, we should not only *not* mystify AI, or any other technology, we should actively *demystify* it to our students, colleagues and the general public, especially if it is our expertise (e.g. Suchman 1987; Suchman 2019). We also have the duty to explain, disrupt, and halt these marketing and discursive trends (e.g. Matten 2026; Tully et al. 2025). These two duties are part of our role as academics anyway, but in the current climate there is a heightened sense of urgency when the technology industry drives and even leads modelling work (e.g. the authors of Bills et al. 2023 all work for OpenAI). Especially urgent is the need to address and subvert the misuse of and purposeful damage to scientific or engineering terminology, like ‘mechanistic understanding’ and ‘black box’ (see Box 1).² As has been said before:

Many people [...] are AI illiterate—understandably, because of the misleading ways its loudest champions describe the technology, and troublingly, because that illiteracy makes them vulnerable to one of the most concerning near-term AI threats: the possibility that they will enter into corrosive relationships (intellectual, spiritual, romantic) with machines that only seem like they have ideas or emotions. (Harper 2025, n.p.; also Guest 2026; Guest, Suarez, et al. 2025; Montell 2021; Suarez et al. 2025)

As we shall see, three statements appear to hang together in disequilibrium, risking our reasoning, our metatheoretical calculus over our science, and eroding how well models can mediate between theory and observation (Elgin 2017; Guest 2023; Guest and Martin 2023; Morgan and Morrison 1999):

Statement 1: ANNs are inspired by the human organism, typically the brain or cognition. Or worse: ANNs constitute a type of mind. (See α in Figure 1.)

While largely uncontroversial to claim this (e.g. Pickering 2014), it nonetheless requires extensive scientific understanding of, and consensus on, human cognition, brain, and behaviour.³ We have nothing of the sort. Besides, if we did, and used these principles in engineering AI systems, then we would not have the problem of not understanding them. This is discussed in depth in Guest and Martin (2025b, 2023).

Statement 2: ANNs constitute a black box that we do not mechanistically understand. Or worse: we cannot ever understand ANNs because they are more intelligent than us. (See Appendix: *ANNs models are a black box or unknowable even if opened* and β in Figure 1.)

Given the ANNs are engineered systems, this is baseless marketing and misuse of terminology. This paper is dedicated to correcting and explaining this flawed assertion and its related reasoning errors.

Statement 3: We understand neither human nor artificial cognition, and therefore they are similar. Or worse: we cannot ever understand either because both are equivalently hard. (See Appendix: *We understand neither human nor artificial cognition, and therefore they are similar* and α , β , and γ in Figure 1.)

This is faulty reasoning (e.g. false analogy, argument from ignorance, discussed in Guest and Martin 2025b, 2023; “the equal opacity argument”, discussed in Peters 2023), as well as going against the previous two points.

These three statements are both problematic in their own right and are incoherent with each other. They exist in *reflective disequilibrium*: “collectively [they do not] constitute an interwoven tapestry of commitments that we can on reflection endorse.” (Elgin 2017, p. 4). The latent assumptions for 1–3 are shown in Figure 1. For 1 and 3 above, they are addressed in past work, such as Guest and Martin (2025b, 2023) and van Rooij, Guest, et al. (2024). Herein, we will address the properties, drivers, and repercussions of statement 2: *ANNs constitute a black box that we do not mechanistically understand* (which is nonetheless interwoven with 1 & 3). How did we get here, and how is this assertion wrong?

¹Regardless, we can attempt to manage things we do not understand, mechanistically or otherwise. For example, we can legislate — a form of attempting to manage something — against harming other people without having a deep understanding, mechanistic or otherwise of why or how people harm each other.

²Importantly, as academics, we can also completely reject all the work by companies as explicitly non-science — since it does not arise without considerable conflicts of interest that are not dispelled merely by naming and uncovering them. Thus, identifying the work correctly as marketing and hype, pumped out to serve their ends and to line their pockets with profit, is not enough (cf. Brause et al. 2023; Chuan 2023; Nguyen and Hekman 2024; Sanguinetti and Palomo 2024; Tsimpoukis 2025). Because even in this case, the damage to the ecosystem of scientific knowledge continues unabated and the harms especially to the most vulnerable amongst us are unimpeded. Academia is meant to protect and use our own expertise exactly in such cases.

³For many examples see Pickering (2014), such as: “On 13 December 1948, the Daily Herald carried a front-page article entitled ‘The Clicking Brain Is Cleverer Than Man’s,’ featuring a machine called the homeostat built by W. Ross Ashby.” (p. 1)

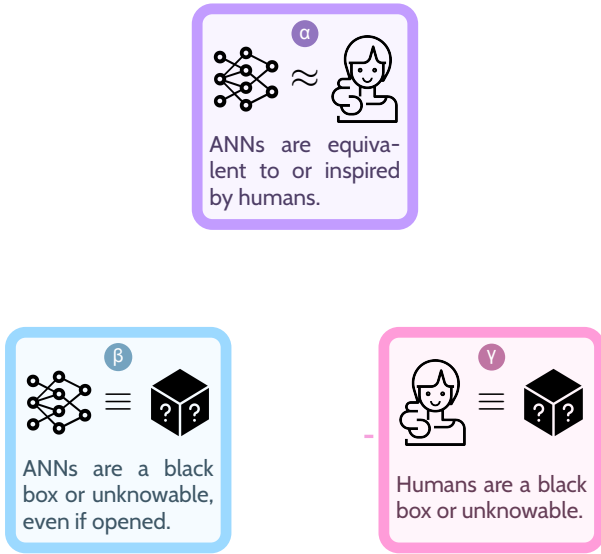


Figure 1

Three co-occurring (latent) assertions in the context of contemporary reasoning on AI and cognition. While they may appear to hang together in “reflective equilibrium” (Elgin 2017), their metatheoretical examination brings their contradictions to light. Starting from the top, and moving clockwise: α) represents the typical assertion of so-called cognitive, neural, or otherwise biological plausibility or inspiration presented for ANNs; β) is the claim we tackle head on here – both in terms of analysing the constituent terms and in how it relates to the other two assumptions; and γ) exemplifies a type of mysterianism, which while not always explicitly expressed, plays a background role and importantly comes into conflict with the other two claims in important ways.

2 (Mis)understanding (the lack of) understanding

And then this guy arrives, with a big black box of deep learning with 100 million parameters in it that he'd trained, and he totally disrupted the whole field.

V (as quoted in Cardon et al. 2018, p. 3)

Before all the 2020s AI hype, there was the 2010s hype that existed in a seemingly different landscape. It allowed for sensible counter from journalists, such as that “the most extreme promises of AI are based on a flawed premise: that we *understand* human intelligence” (emphasis added, Conditt 2016). Counters from technology industry players, such as Microsoft’s co-founder, were also mainstream:

Creating this kind of advanced software requires a prior *scientific understanding of the foundations of human cognition*, and we are just scraping

the surface of this. [T]he difficulty of building human-level software goes deeper than computationally modeling the structural connections and biology of each of our neurons. [And e]specially for the cognitive neuroscience of humans, we are not close to the requisite level of *functional knowledge*. [A]s we learn more and more about the actual complexity of how the brain functions, the main thing we find is that the problem is actually getting harder. (emphasis added, Allen and Greaves 2011; also see Siegel 2019)

It is *still* true that we do not know how to design human-like systems because we do not understand cognition itself as well as due to many other reasons (Elish and boyd 2018; Guest, Scharfenberg, et al. 2025; Rich, Blokpoel, et al. 2020; Rich, de Haan, et al. 2021; van Rooij, Guest, et al. 2024).

In the 2020s remarkably perhaps, not only have journalists and technology sector-affiliated writers predominantly moved to spreading exaggerated claims about AI, but also our colleagues increasingly shift to uncritical views. Many started this shift from the 2010s, and mainstream acceptance since then has only expanded (Guest and Martin 2025b; Guest, Suarez, et al. 2025). An important shift, the one we care about herein, is the unproblematised assertion that: “No one yet knows how ChatGPT and its artificial intelligence cousins will transform the world, and one reason is that *no one really knows what goes on inside them*.” (emphasis added Musser 2023, n.p.) And this public face of not understanding is widespread: “Dario Amodei [CEO of Anthropic, an AI company] stood before the U.S. Senate in 2023 and said [...] that *even the people building artificial intelligence don’t understand how it works*.” (emphasis added, Tarita 2025, n.p.) Melanie Mitchell claims “I don’t know how [LLMs a]re doing it or if they could do it more generally the way humans do—but they’ve challenged my views” (Musser 2023, n.p.). Yoshua Bengio has been on a similar track for at least a decade, claiming that:

it’s exactly because *we can’t mathematically pick apart a decision made by deep learning software* that it works so well. [Furthermore,] Bengio argues that trusting a computer is no different, or more dangerous, than trusting another person. ‘You don’t understand, in fine detail, the person in front of you, but you trust them’ (emphasis added; Pearson 2016, n.p.; also recall Figure 1 above).

And returning to Amodei (2025), he claims:

this *lack of understanding* is essentially unprecedented in the history of technology. [And o]ur *inability to understand models’ internal mechanisms* means that we cannot meaningfully pre-

dict such behaviors, and therefore struggle to rule them out (emphasis added, n.p.).

In a footnote he contrasts what he sees as what we do understand versus what we do not:

In the case of AI systems, we can set the basic architecture (usually some variant of the Transformer), the broad type of data they receive, and the high-level algorithm used to train them, but *the model's actual cognitive mechanisms* emerge organically from these ingredients, and our *understanding of them is poor*. (emphasis added, Amodei 2025, n.p.; also see Bricken et al. 2023)

Journalists often accept these premises and conclusions: “The people who develop AI are increasingly having problems explaining *how* it works and determining *why* it has the outputs it has.” (emphasis added, Xiang 2022, n.p.) Also:

The wildest, scariest, indisputable truth about AI's large language models is that *the companies building them don't know exactly why or how they work*. Sit with that for a moment. The most powerful companies, racing to build the most powerful superhuman intelligence capabilities [...] *don't know why their machines do what they do*.” (emphasis added, VandeHei and Allen 2025, n.p.; also e.g. Pedreschi et al. 2019)

The dramatic irony here is not that we have to sit with the idea that companies do not know how these models work, but that companies' representatives, including affiliated scientists, are willing to go on record saying they do not know how these models work (Salvaggio 2025).⁴ And this has been brewing for at least a decade, with non-controversial pieces in *The Atlantic*, for example, titled “Not Even the People Who Write Algorithms Really Know How They Work” (LaFrance 2015; also see: Bucher 2018, 2025; Guest and Martin 2025b)

This discourse is paradigmatic, representing a typical series of mainstreamed views on understanding models with relevant practitioners and proponents, suggesting that: “While there are some types of AI that humans can comprehend, there are others that, because of their complexity and high dimensionality, are *beyond the ken of human intelligence*.” (emphasis added, Calvello 2023, n.p.; recall Figure 1) As well as that: “While the potential of ANNs is clear, they are still surrounded by an air of mystery and intrigue, leading to a *lack of understanding of their inner workings*.” (emphasis added Maier et al. 2023, p. 14) Some journalists, on the other hand, are more aware of the impending and present issues with the erosion of trust in experts by claiming we ourselves have no idea how these models work, which in turn reinforces harm to the public understanding of AI: “If people understand what large language models are and are

not; what they can and cannot do; what work, interactions, and parts of life they should—and should not—replace, they may be spared its worst consequences.” (Harper 2025, n.p.) As Bryan Pfaffenberger (1988) says: “Technology, in short, is a mystifying force of the first order[,] suspending us in webs of significance that we ourselves create.” (p. 250)

In the remainder of this section, we will disentangle mechanistic understanding from other types of scientific understanding and from non-scientific, user, or lay understanding — the main focus being the difference between mechanistic ‘how’ understanding versus any other relevant type. To do this, first we will discuss understanding generally from contemporary philosophy of science; and second, present two case studies on the confines of mechanistic understanding in science and statistics. These case studies aim to highlight what it means to have a mechanistic understanding in typical scientific parlance and how this is violated in the AI context.

The concept of understanding has been widely discussed in philosophy since ancient times, but the inquiry into understanding has emerged with distinctive force and interest in more recent debates in epistemology and philosophy of science. Herein, we touch on the relevant set of prominent accounts that fundamentally treat understanding as a cognitive phenomenon.

Catherine Elgin's work on understanding has played a key role in the philosophical discussion on understanding. While other epistemologists and philosophers of science focused their investigations on knowledge as the main cognitive achievement of science, Elgin (2002) argues that, in many cases, cognitive progress happens when we advance understanding, which is not reducible to the mere accumulation of new justified true beliefs. As a matter of fact, Elgin claims: “Not being restricted to facts, understanding is more comprehensive than knowledge ever hoped to be.” (Elgin 1993, p. 14) According to her, understanding

involves the ability to profit from cognitive labors, to draw out the implications of findings, to integrate them into theory, to utilize them in practice. Understanding a particular fact or finding, concept or value, technique or law is largely a matter of knowing where it fits and how it functions in the matrix of commitments that constitute science. And neither knowing

⁴Notably, there is a confusion between why and how questions (Prigogine and Stengers 2018). Let us assume the why questions collapse into how questions, as they always do in science and related disciplines like mathematics, e.g. in a maths problem ‘why is $x = 5$?’ is answered step-by-step by demonstrating *how* x takes on that value. If why questions do not collapse into how questions and they remain qualitatively separate then the answers will never come from science anyway. So this possible distinction could become moot, thus boiling back down to just a how question for our scope here as companies claim to want to use scientific research techniques.

where nor knowing how reduces to the knowing that that traditional epistemology explicates. (Elgin 1993, p. 15)

Thus, understanding encompasses and enables a number of cognitive skills and achievements that cannot be reduced to propositional knowledge.

From a different angle, Emily Sullivan (2018) also emphasizes the role that cognitive abilities play in understanding by adopting Allison Hills' (2016) characterization of the abilities needed for understanding, which states that

if you understand why p (and q is why p), then you believe that p and that q is why p and in the right sort of circumstances you can successfully:

- i) follow some explanation of why p given by someone else.
- ii) explain why p in your own words.
- iii) draw the conclusion that p (or probably that p) from the information that q .
- iv) draw the conclusion that p' (or probably p') from the information that q' (where p' and q' are similar but not identical to p and q).
- v) given the information that p , give the right explanation, q .
- vi) given the information that p' , give the right explanation q' . (Hills 2016, p. 663)

According to Sullivan, such cognitive abilities are constitutive of understanding-why, and are the same kind of cognitive abilities that we find in ordinary cases of knowledge-that (Sullivan 2018).

Differences as well as overlap exist between Elgin's and Sullivan's approaches to understanding. On the one hand, unlike Elgin, Sullivan is not opposed to relating understanding to knowledge; on the other hand, both conceive understanding in terms of cognition. While Elgin's account of understanding claims that representations in science and other disciplines advance our understanding despite being false, Finnur Dellsén, Tina Firing, and their colleagues (2024) argue that the mental representations involved in understanding must match the world. According to them, "an agent S understands X to the extent that S accurately and comprehensibly represents the network of dependence relations in which X stands, or fails to stand, to other things." (Dellsén et al. 2024, p. 676) However, they claim that their account of understanding is epistemically undemanding insofar as it does not require beliefs about X nor S 's possession of justification towards X or its network of dependence relations. It's worth noting that Dellsén et al. (2024) also links understanding to cognition; they view it as a cognitive achievement that constitutes progress in a field of knowledge.

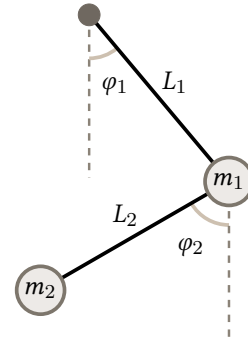


Figure 2

A schematic of the double pendulum, which consists of a single pivot point on which two masses m_1 and m_2 hang. The masses are connected by two weightless rods L_1 and L_2 with a joint in the middle. After a certain energy threshold is applied to the pendulum it displays chaotic motion (Stachowiak and Okada 2006).

In line with the aforementioned account, the idea that understanding is a cognitive relationship constituted by mental representations that encode the right kind of dependence relations is also assumed by Tania Lombrozo and Daniel Wilkenfeld (2019), who distinguish between mechanistic and functional understanding. According to their view, "mechanistic understanding relies on an appreciation of parts, processes, and proximate causal mechanisms[. While f]unctional understanding, by contrast, relies on an appreciation of functions, goals, and purpose." (Lombrozo and Wilkenfeld 2019, p. 209) For Lombrozo and Wilkenfeld, mechanistic and functional understanding are distinct insofar as they involve different objects, and they involve different epistemic relationships. Hence, confounding one with the other may lead us astray from grasping the phenomenon. As we shall see below, these two — mechanistic and functional perspectives on understanding a system — do become confused or otherwise entangled. But first as promised, two case studies.

The Double Pendulum: A pendulum comprises a mass hung such that it can swing from a pivot point under the influence of gravity (Birx 2009). The double pendulum is made of two simple pendulums joined end-to-end, as depicted in Figure 2, which nonetheless exhibits "bewildering complexity" (Richter and Scholz 1984; also see Korsch et al. 2008; Ohlhoff and Richter 2000; Stachowiak and Okada 2006).⁵ Despite the chaotic nature of the movements after a certain energy threshold, aspects of the double pendulum, such as

⁵In fact even a simple pendulum can be chaotic: "If the initial push is just enough to bring it into a vertical position with zero velocity, the direction in which it will fall, and therefore the nature of its motion, are indeterminate." (Prigogine and Stengers 2018, p. 73)

the locations of the masses, $m_1 = (x_1, y_1)$ and $m_2 = (x_2, y_2)$ are nonetheless described by simple equations (Levien and Tan 1993; Neumann 2002):

$$\begin{aligned} x_1 &= L_1 \sin \varphi_1 & x_2 &= x_1 + L_2 \sin \varphi_2 \\ y_1 &= -L_1 \cos \varphi_1 & y_2 &= y_1 - L_2 \cos \varphi_2 \end{aligned} \quad (1)$$

where there are two weightless rods of lengths L_1, L_2 , which carry masses m_1 and m_2 , and are connected by a joint as shown in Figure 2; and φ_i are the angles of each m_i (where $\varphi_i = 0$ means the respective L_i is vertically downwards, and anti-clockwise means positive values; Korsch et al. 2008; Neumann 2002; Puzyrov et al. 2022).

The double pendulum model fulfils many of the requirements that are described above as desiderata for ‘mechanistic understanding’ in the case of AI models. We know the mechanisms, and we can calculate the location of m_1 and m_2 (see Figure 2; as well as many other aspects of the pendulum we do not go into here) because of that. Hence, analytic and computational models exist of this apparatus (e.g. in Python, Neumann 2002; and in JavaScript, Nolte 2020). And yet, it is chaotic, and indeed unpredictable in the real world, since we do not know, and in fact cannot know, the initial conditions (Shinbrot et al. 1992). This is because “tiny errors in measuring the current state of the atmosphere (or any other chaotic system) would be amplified rapidly making it impossible to forecast.” (Adam and Peck 2005, p. 7)

Do we understand double pendulums? If no, what is there left to understand? ‘No’ here makes assertions about experts not understanding ANNs banal, since run-of-the-mill double pendulums are also not understood. It calls into question any and all of our understanding of physics and related systems. If yes, how is this example different to an ANN? Under makeism, inter alia the idea that through building we understand something (van Rooij, Guest, et al. 2024), and coupled with the fact clearly we *can* build both pendulums and ANNs: we can claim we do indeed understand. What is missing in our understanding of the ANN, that is present in our understanding here? Both are fully modelled on a computer.

Logistic regression: A familiar subtype of the generalised linear model is logistic regression, which uses the logistic function to predict the probability of the outcome or dependent variable \hat{Y} given the input or independent variables (Kleinbaum and Klein 2002; Poston Jr et al. 2023; Stoltzfus 2011; for an extensive history see: Cramer 2002). Logistic regression, much like most ANNs, is a kind of ‘supervised machine learning’ — it requires labelled input-output pairs and internalises the relationships between them to statistically predict out-of-sample data (cf. Nusinovi et al. 2020). The output prediction has a range between 0 and 1, which represents the likelihood of a particular input being in either one or the other category, such as yes/no or some other binary classification decision. We use the logistic function

to calculate this probability, where in the case of a single independent variable X looks like this:

$$p(\hat{Y}) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 X))} \quad (2)$$

where β_0 is the intercept (called the bias in an ANN), β_1 is the regression coefficient or slope (called the connection weight in an ANN), and $\exp z$ is the exponential function (Cramer 2002; Kleinbaum and Klein 2002; in an ANN other functions can also be used such as rectified linear unit, ReLU, activation function, El Ghaoui et al. 2021; Fukushima 1969; Goyal et al. 2020; Householder 1941; Krizhevsky et al. 2017).

The similarities between ANNs and logistic regression run deep because the probability above also captures the basics of a single unit in an ANN firing given input. In other words, logistic regression is a special case of an ANN with a single unit. And so violations of logistic regression’s assumptions could affect ANNs as well (Stoltzfus 2011). It also means that interpretations of what the coefficient means for the outcome variable apply equally to ANNs and logistic regression (Cramer 2002). The links between statistics and ANNs go back about half a century if not longer (e.g. Oja 1989; Sanger 1989a; also Schmidhuber 2015).

Do we understand logistic regression? If no, what is there left to understand? As before with the pendulum, answering ‘no’ makes assertions about experts not understanding ANNs banal, since the simpler and special case of an ANN, logistic regression, is also not understood. Also it calls into question any and all of our understanding of statistics and of mathematical systems. If yes, how is this case different to an ANN? If we do indeed understand how a single unit in an ANN works, what is the difference when there are many hundreds or thousands? Does understanding a single unit not imply mechanistic understanding of more than a single unit? Is ‘mechanistic understanding’ not exactly this form of understanding? Additionally, and again under makeism, clearly we *can* build both a single unit as well as a full-blown thousand- or million-unit ANN, so — as before — this adds weight on the side of understanding. What is missing in our understanding of the ANN that is present in our understanding of logistic regression? Both are unpredictable if we cannot infer the unknown parameters (the beta coefficients) without the ground truth dataset. And both are fully modelled on a computer if otherwise — just like the complex pendulum before.

The pendulum and regression cases highlight issues with the claim that we do not understand ANNs, AI, or any engineered system in the general case (recall b in Figure 1). And so typically to engineer something, some basic mechanistic understanding is required even if that understanding is not (any more) scientific. But in the specific case of computational models of basic physical systems or simple statistics, as with the pendulum and regression above, we have complete understanding of the equations which govern these models,

since we built them and since they have been created for the sole purpose to be understood. In contrast to this mainstream even prosaic understanding of ‘understanding’, is the case of the contemporary ANN model. As touched on above, it is claimed to be a black box — opaque, unknown — by critics and proponents alike. We turn to this below.

3 Shutting up the Black Box

The shared myth-making around the black box of AI cultivates useful confusion about what can and cannot be known. It helps to cultivate the systems as more mysterious, even sublime, than they really are. It leads to a cottage industry of thinkers framing a series of design tricks to organize text output as an “other” intelligence[.]

Eryk Salvaggio (2025, n.p.)

Typically, using a ‘black box’ in engineering or science implies a strictly functional understanding⁶ of an engineered system — one of relating inputs to outputs via a so-called black box: an unknown part of the system (see Box 1; for a history of the term, see Petrick 2020; also see Nizami 2020). The black box not only need not be further mechanistically analysed or understood, but also the functional input-output mapping suffices *because* mechanistic understanding is often impossible (also for formal reasons; Ashby 1956; Beer 1959; Bunge 1963; cf. Guidotti et al. 2018; Hassija et al. 2024; Pasquale 2015), such as in cases where we have limited access for whatever reason to the ground truth of the system (Guest, Scharfenberg, et al. 2025).

We have already touched on the use of ‘black box’ with respect to ANNs, especially LLMs, and AI generally (viz. Buckner 2021; Cassauwers 2020; Chirimuuta 2021a; Durán and Jongsma 2021; Peters 2023; Tangermann 2024; Zednik 2021). For example, there is the more nuanced: “Although large neural networks, in particular LLMs, are frequently referred to as ‘black boxes’, there has been significant progress in understanding the internals of these systems.” (Griffiths et al. 2025, p. 13) Such statements coexist in the field with warnings from

Tommi Jaakkola, a professor at MIT who works on applications of machine learning[, and who says:] “Whether it’s an investment decision, a medical decision, or maybe a military decision, you don’t want to just rely on a ‘black box’ method.” (Knight 2017, n.p.)

As well as with IBM’s statements:

A black box AI is an AI system whose internal workings are a mystery[.]

AI developers broadly know how data moves through each layer of the network, and they

have a general sense of what the models do with the data they ingest. But they don’t know all the specifics. For example, they might not know what it means when a certain combination of neurons activates, or exactly how the model finds and combines vector embeddings to respond to a prompt.

Even open-source AI models that share their underlying code are ultimately black boxes because users still cannot interpret what happens within each layer of the model when it’s active. (Kosinski 2024, n.p.)

This tells a story that matches the other retellings (Amodei 2025; Bills et al. 2023; Perrigo 2024; Sullivan 2023): no matter what and with even 100% access to the full codebase we have no understanding. ANNs are a completely opaque black box.

In contrast, what has been happening for a while is a form of terminological blurring (Montell 2021): “the metaphor of the black box itself constitutes such interference. In our attempts to open the black box, what frequently gets ignored is the question of whether the metaphor of the black box holds at all.” (Bucher 2018, p. 47) Petrick (2020) explains:

Cyberneticians defined the black box as a system where only the inputs and outputs are known, with the inner workings unknown or unknowable[:] a way to break down and analyze systems too large to understand in other ways. A complex system could be simplified, made calculable and replicable, through an understanding of its inputs and outputs. (p. 576)

Ashby (1956) gives examples of how this method can be generalised from engineering: to the clinical researcher trying to understand neuropsychological patients “by means of tests given and speech observed, to deduce something of the mechanisms that are involved” (p. 86). He also includes the “psychologist who is studying a rat in a maze [in order] to deduce something about the neuronc mechanism” (Ashby 1956, p. 86) — even going so far as to claim the black box can apply “everywhere”. As Medina (2011) explains “cybernetic thinking influenced the trajectory of operations research, computer engineering, control engineering, complex systems, psychology, and neuroscience.” (p. 11) This is notable because in the 2020s, about 70 years later, this technique is

⁶Although many violate this idea of a strict distinction: “The core idea is that functional analyses are sketches of mechanisms, in which some structural aspects of a mechanistic explanation are omitted. Once the missing aspects are filled in, a functional analysis turns into a full-blown mechanistic explanation.” (Piccinini and Craver 2011, p. 284) Either way of course, if functional understanding leads to “full-blown mechanistic”, that poses no issues and in fact strengthens our position herein.

described as travelling in the exact opposite direction from psychology to the study of the so-called modern black box ANNs (e.g. Pellert et al. 2024; cf. Petrick 2020). And in a twist of irony the black box is now used in an obfuscatory way to avoid understanding as opposed to furthering and enhancing it (Bucher 2025; Christin 2020): “the black box metaphor became a black box” (Petrick 2020, p. 576).

Black boxing is a powerful method, no doubt, but has several limitations, which the original proponents go to serious pains to explain (e.g. Ashby 1956; Glanville 2009). Importantly, it cannot find the actual internals because of “an infinity of possible internal mechanisms. [...] It has been shown by Shannon that any given behaviour can be produced by an indefinitely large number of possible networks.” (Ashby 1956, p. 93; Guest, Blokpoel, et al. 2026) Notwithstanding, in a real black box, a physical engineered device with interconnected wiring and visible mechanisms, both peeking inside is ultimately possible, as is reading documentation, asking engineering experts or even the original designers, and corporate espionage (as in cases described in Guest, Scharfenberg, et al. 2025). Although, all this is outside the black box method! In biological systems as well as corporate black boxes access to the ground truth is often not allowed, in the latter case, or actually provably impossible in the former (e.g. Schierwagen 2012).

What does presage something of the contemporary in the original discussions of the black box, however, are the potential slippages into the false analogy we warn about (recall Figure 1; also: Guest and Martin 2025b, 2023):

It should be noticed that as soon as some of a system’s variables become unobservable, the “system” represented by the remainder may develop remarkable, even miraculous, properties. A commonplace illustration is given by conjuring, which achieves (apparently) the miraculous, simply because not all the significant variables are observable. It is possible that some of the brain’s “miraculous” properties—of showing “foresight”, “intelligence”, etc.—are miraculous only because we have not so far been able to observe the events in all the significant variables. (Ashby 1956, p. 114)

Contrast the above with the following, which dispels the flawed logic of equivocating between human and machine (Figure 1):

While opacity is a distinguishing feature of many other areas of science and technology, the myths surrounding computing may stem less from the fact that it is an opaque esoteric subject and more from the way in which it can be seen to blur the boundary between people and machines (Turkle 1984). To be sure, most people do not

understand the workings of a television set or how to program their video cassette recorders properly, but then they do not usually believe that these machines can have intelligence. The public myths about computing and AI are also no doubt due to the ways in which computers are often depicted in the mass media — e.g. as an abstract source of wisdom, or as a mechanical brain. (Bloomfield 1987, p. 72)

Returning to the contemporary discourses on the black box, and as already evinced by statements above (for full list see Appendix: *ANNs models are a black box or unknowable even if opened*), we see how they become ever more problematic; deviating in intent from original propositions and definitions completely. Such as Natasha Alechina explaining: “If AI systems operate in a black box, we have no insight into how to fix the system if there’s a mistake” (Jiménez 2025, n.p.) Or IBM:

If a black box model does make the wrong decisions or consistently produces inaccurate or harmful outputs, it can be hard to adjust the model to correct this behavior. Without knowing exactly what happens inside the model, users cannot pinpoint exactly where it is going wrong. (Kosinski 2024, n.p.)

On the contrary, we *do* have insight and *can* explain — and scientifically so — a lot of what a system is doing wrong (or right) using a non-mechanistic understanding. And we can even repair them. IT technicians do not require computer science or hardware-related degrees to infer that one may need a new power supply or motherboard; exactly because of the modularity and multiple realisation of such engineered systems (viz. Chirimuuta 2018, 2021b; Egan 2017; Figdor 2010; Guest, Blokpoel, et al. 2026; Guest and Martin 2021, 2023; Guest, Scharfenberg, et al. 2025; Hardcastle 1995, 1996, 2019; Litch 1997; Polger and Shapiro 2016; Ross 2020; Tsouna 2023). And remember, in a true black box situation we *also* cannot know for certain “where it is going wrong” for a given system by definition. In stark contrast to the technology industry’s claims, much of computer science and software engineering — such as theoretical computer science, as well as the development of specification languages — define very complex systems in terms of functional descriptions (see examples in Cooper and Guest 2014; Guest and Martin 2021). Furthermore, a human (non-expert) user of any number of machines can know (where) they are going wrong (e.g. software bugs) without looking at the source code or any inner workings (e.g. Spectre is a hardware bug diagnosed using software and fixed by redesigning hardware; Kahn et al. 2018; Kocher et al. 2019). And so, “‘AI models are black boxes’ [...] sounds like a truism, and could yet not be further from the truth.” (Offert and Dhaliwal 2025, p. 5)

In conclusion, and as Cummins (2000) also explains:

It is possible to understand how a mechanism works, and hence to be in a position to explain its behavior and capacities—the effects it exhibits—without being able to predict or control its behavior. This is true generally of stochastic or chaotic systems. It is also true of systems whose relevant initial states are unknowable or simply unknown. In possession of a machine table for a Turing machine, I can explain all of its capacities, but, lacking knowledge of its initial state, I may be unable to predict its behavior (Moore 1956). [...]

So, systems can be well-understood yet unpredictable. (pp. 119–120)

Ultimately:

- a) “Everything is a black box because we can never have complete knowledge of how anything operates; all we can easily observe are those inputs and outputs that we can affect and perceive.” (Petrick 2020, p. 588; also Ashby 1956) And “the so-called ‘black box problem’ that traditionally arises whenever a[n AI] system’s behaviour cannot be explained by appealing to its initial construction and programming” (Wadden 2022, p. 764) is not unique to contemporary AI. We cannot predict the “behaviour” of the factual double pendulum, but that detracts little from our understanding. ‘Black box’ is thus open to *full deflation*.
- b) Or black boxing something in the scientific and engineering sense indeed *provides us with understanding* (e.g. Ashby 1956; and functional analysis can even give us mechanistic understanding according to Piccinini and Craver 2011).
- c) Or *both*, and so AI of any type being a black box is deeply unproblematic, frankly pedestrian, in the scientific sense of understanding it mechanistically and functionally (cf. Amore 2020; Creel 2020; Longo 2025).

Therefore, closely inspecting ‘black box’ in this context indicates understanding ANNs, LLMs, AI and related concepts is a given or otherwise non-exceptional and unproblematic. In the following section, we explain how despite all this, mystification persists to the point of full blown mysterianism.

4 Grist to the Leibniz Mill

It must be confessed, moreover, that perception, and that which depends on it, are inexplicable by mechanical causes, that is, by figures and motions. And, supposing that there were a mechanism so constructed as to think, feel and have perception, we might enter it as into a mill. And this granted, we should only find on visiting it, pieces which push one against another, but never anything by which to explain a perception.

Gottfried Leibniz, sect. 17 (1714/1989)

We have seen that AI’s creators claim that these systems are somehow inherently or predominantly unknowable even though their mechanisms are not only known but directly implemented on a computer as a function of known mathematical formalisms (for about half a century or longer, e.g. Bobrowski 1978; Rumelhart, Durbin, et al. 1995; Rumelhart, Hinton, et al. 1986; Rumelhart and McClelland 1986; Sanger 1989b; Williams and Zipser 1989, 1995; see Schmidhuber 2015 for a historical perspective). We will tackle this head-on with an error theory as a corrective and an explanation. Finally, we can address: *how is an ANN a black box to its own creators?* Our error theory to the issues with proclaimed lack of understanding due to the opaque nature of models rendering them black boxes comprises three related parts:

- 1) a **confusion between explanans** — that which does the explaining or modelling, such as a theory or model (natural language, formalisms, mathematics) — **and explanandum** — that which is being modelled or needs to be explained (data, observations);
- 2) a **misunderstanding of** what scientific **understanding** (mechanistic, functional, or otherwise) and theorising are in comparison to data and statistical models thereof; and finally and perhaps most importantly at this juncture,
- 3) if we grant Leibniz’ Mill argument (see Box 1) then yes, indeed, **no understanding** (of the high calibre sought) **can ever be obtained** by looking at mechanisms.

Once the two interlocking misapprehensions (1 and 2 above) of scientific modelling and theorising are seen in the right light, it becomes clear that insisting such models are black boxes that we do not understand (recall Figure 1) is a pragmatic frame, but not one that is reasonable. This is pragmatic insofar as it allows for the maintaining of ANNs as a theory or model of brain, cognition, or behaviour with no real scientific legwork to go with it, in contrast to classical connectionism (pre-2010 ANN modelling work; Guest and Martin 2025b) and other norms of cognitive scientific practice (Blokpoel 2018; Guest 2023; Guest and Martin 2021, 2023; van Rooij and Baggio 2021; van Rooij and Guest 2025a). And if

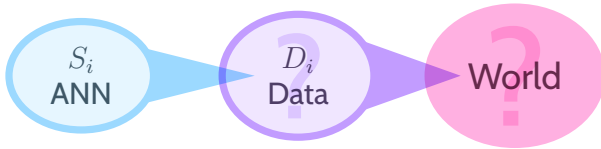


Figure 3

Not all systems, S_i relate to their relevant data, D_i and world phenomena this way, but ANN models (as many inferential statistical models) do when used without scientific understanding of the data-generating process or without verbal or formal theory to match (Guest and Martin 2025b, 2021, 2023). As we move from left to right above, we move from full understanding of engineered systems, S_i to less or even possibly none of the data, D_i to the world which can be thought of as the set of phenomena to be understood.

indeed the third part about Leibniz' Mill (3 above) is at play, then no scientific theory under mechanical materialism will ever provide understanding of the type required (Cummins 2000; Leibniz 1989; Rozemond 2014, 2019; Sullivan 2022). We discuss each of the above three points in turn below.

This extract from Rudin and Radin (2019) is informative because it appears to describe a sensible state of affairs until one realises that “understand” below has to mean memorise:

these black box models are created directly from data by an algorithm, meaning that *humans, even those who design them, cannot understand how variables are being combined to make predictions*. Even if one has a list of the input variables, black box predictive models can be such complicated functions of the variables that *no human can understand* how the variables are jointly related to each other to reach a final prediction. (emphasis added; p. 3)

Would anybody say *nobody* understands the dictionary just because nobody can memorise the whole book? And vice versa would anybody say somebody who can (only) memorise the dictionary has therefore understood it? To add on more contradictions, usually ‘understanding’ something is not memorising it — rote learning is very valuable, but not the cornerstone of understanding. Yet we often see claims that the ANN models — which typically can only memorise, even if lossily — do not provide us with understanding (recall Figure 1; Ahmed et al. 2026; Zhang et al. 2017). Remember that, according to those experts and critics, this lack of understanding embodied by ANN models is why *we* need to understand them!

In addition, often non-connectionist systems, such as expert systems based on formal logic, are described as interpretable or explainable. This is because they display through

their design the processes they carry out using terms and expressions humans can more easily interpret. To turn an ANN into such an interpretable system is to design and build (from scratch) such a system using easily understood internal mechanisms (Blokpoel 2018; Cooper and Guest 2014; Naur 1985).




To address the first point above, the **confusion between explanans and explanandum**, let us leave contemporary AI to the side for a moment (cf. Guest 2026). Imagine two systems S_1 and S_2 comprising two databases with ‘identical’ content. One is pre-digitisation and paper-based, wherein every record is a physical piece of paper and is stored in a filing cabinet, D_1 . The other, D_2 is a modern relational database, which can be queried, meaning it can be accessed using a formal database query language, like SQL (Structured Query Language; Pratt and Last 2015). D_2 is the digitised version of the contents of D_1 . Let us also assume one does not know — as one never does, hence why such systems exist — every record in these two databases by heart. Everything found out by us about S_1 and S_2 is a property of the data stored in each D_i ; and not caused by, in an authorial sense, the SQL query in the case of S_2 . Importantly, we (can) know how the SQL queries’ mechanisms work because humans engineered the implementation and we specifically wrote the queries. One may think that the systems have properties other than as derived from the data. But these properties are known a priori; they stem from the nature of the query language or its use by us, since of course similar or even identical results from the database can be obtained from dramatically different SQL statements with different time and space complexities.

Remember that an important difference between S_1 and S_2 is the substrate, one is paper-based, the other is digital (see Table 1). S_1 , of course, has no built-in automated search. So any methodical search in S_1 is manual and has to be first planned or designed if we are to be efficient — we need to figure out how to deal with these piles of papers. In the case of S_2 , it comes with a formal language which allows for structured queries by design. Importantly, in both of these cases, the *data* in the database is the ‘mysterious’ one. It is a thing in itself, while our manual searching is separate to D_1 as is the SQL, which is a formal system above D_2 . And what is more is that SQL is not a single static way to access glimpses of the data, but a nigh on Turing-complete language which is written as we see fit to query S_2 with the goal of returning relevant parts of D_2 . Is understanding a specific database *about* the data inside, which is unique to D_i or is understanding a database *about* understanding SQL? Most likely, understanding a specific database means understanding the contents stored in D_i .

An ANN is no different (see Figure 3 and compare and contrast the rows of Table 1). It is a formal system, S_3 with known or knowable equations and mechanisms (recall Equation 2), defined with respect to some dataset D_3 . In contrast,

Table 1

In contrast to scientific, and especially cognitive scientific models, the systems below do not embody an understanding of the world, nor the phenomena within it, nor the observations from the world represented in data. What they do is in fact exclusively house data in a post-processed state, and in a lossless and formally query-able (first two rows) or non-formally query-able and often stochastic and lossy (last row) way.⁷

ASPECTS OF SYSTEMS THAT OPERATE OVER DATA			
System	Properties of S_i	Data Involved in D_i	Related World Phenomena
 Paper Database	By no means formal; there is not only always a human-in-the-loop, but also more margin of error in certain common situations, e.g. a person can forget to return a file to the cabinet after reading, or a file can be damaged beyond repair because backups are either non-existent or time-consuming. Full-blown cognition is used.	Pre-processed data is that which, say in a doctor's office, patients fill in paper forms handed to them by the receptionist. The office staff then read the paper forms and decide if any changes must be made, text rewritten, or similar and afterwards place the now post-processed form inside the relevant filing cabinet.	The world causes the previous cell, say the patients' characteristics about themselves; these are a function of their knowledge. As depicted in Figure 3, the world is sampled to create D_i , in our case the paper forms, but the full picture is not in any way understood fully by looking only at D_i . We want the process behind it.
 Relational Database	A fully formal system with the ability to apply a structured query language to extract subsets of the database. Such constraints address some of the issues above, e.g. that certain forms of damage (although by no means all) are mitigated as querying is through automated copies.	As above, pre-processed data is input and subject to checks, e.g. validating phone numbers or email addresses using regular expressions, or using drop-down menus for selecting the country. Post-processed data is stored losslessly by the relational database management system.	Even though digitalisation allows for the automation of figures and descriptive and inferential statistics, understanding this data is not possible only through such analyses. In science, this is why we need theory and models to understand phenomena represented by the data.
 Artificial Neural Network	While a fully formal system (recall Equation 2), what this system does not afford us is the ability to query it using a formal language like SQL. In the generative or language model subtype (wherein so-called prompts are used) repeatable and lossless interfacing is also ruled out. ⁷	Data for modern ANNs is incredibly post-processed, with labels added to photographs and so-called reinforcement learning from human feedback for contemporary language models, which is human input as a function of the model's output (Guest 2026; Guest and Martin 2025b).	Contemporary ANN systems contain the world directly through people who intervene because full automation is impossible. But even setting this aside, as with the above cells for paper and digital databases, it remains the case that we do not understand how D_i came about.

we do not know what processes or mechanisms generated the dataset: that data generation is the subject under study. If the statistics extracted by the ANN much like the results of the infinite possible queries we could write in SQL, give us unexpected patterns in the output data, then that is because those patterns might indeed exist in the data. The fact it is surprising is because we do not (yet) understand the mechanisms responsible for generating the data and not because something unknown in the system *other than* D_3 is generating the patterns (*ceteris paribus*, i.e. assuming no catastrophic hardware or software bugs). What this means is that S_3 , like any S_i above, queries lossily⁷ or otherwise D_i , but it does not produce subsets of D_i in a causal or authorial

sense (viz. Guest, Suarez, et al. 2025; Guest and van Rooij 2025; van Rooij and Guest 2025b).

Understanding this first point, also sheds light on the second, on why “[b]y late 2015, technology companies that were once seen as being at the forefront of Big Data began rebranding their efforts as ‘AI.’” (Elish and boyd 2018, p. 60) How can there be mystique maintained, mystifierism upheld, magic performed dextrously, if we call it what it is? ‘Big

⁷Not all ANNs internalise data in a lossy way (cf. Liu et al. 2026), nor are all ANNs stochastic, and neither are all generative models language models — many such confusions are untangled in Guest, Suarez, et al. (2025).

data’ — it is just a lot of data. Data is exactly what makes an ANN, just like any statistical model, function for a certain use case.

LeCun would say: “I have a 10-layer neural network and it does the same thing”. Then we would say, “Are you sure? What’s new?” Because once you have a neural network, even though it might have 10 layers this time, it doesn’t work any better than the last one. (Cardon et al. 2018, p. XXVII)

More data is what changed — that is “what’s new” — from nine thousand data points in 1988 to ImageNet’s 14 million in 2018 (Cardon et al. 2018) to essentially the totality of internet-accessible human data in 2025.⁸

And so to say we do *not* understand ANNs is a mischaracterisation; it constitutes a **misunderstanding of understanding**. It is the data-generating process of the real world we do not understand. Not any mechanism in the ANN itself; not any model operating over the data. To address this second misunderstanding, we need to grapple with how *no* inferential statistical model from logistic regression to ANNs by virtue of just being applied to the data provides explanations. A scientific explanation of a dataset is the product of more than just a model being fit to the data. In the simpler case of logistic regression, Equation 2 does not explain how $p(\hat{Y})$ is predicted by X other than in the statistical sense and by definition. In the scientific sense — where we are on the hunt for a theory — a mere application of a statistical test is no more or less explanatory than the data itself. It provides a view on the data, such as how different variables relate to each other and if those relationships are statistically meaningful, but it does not reveal anything we did not know or at least suspect before. And that is the point of statistics, to confirm or not a hypothesis we may have about the dataset (Guest and Martin 2021). Theories are a different beast altogether.

So when scientists, engineers, or others insist we do not understand an ANN, we must ask what a mechanistic understanding is if not like in Equation 2 a formal mathematical description of the components. If they claim it is a functional understanding that we lack, then we must remember a black box model as-is is a functional perspective on a system. What we do not understand is the data compressed or stored, lossily or losslessly, inside the model. Once we admit this, we can start to build and use theories and models effectively and conscientiously in science (Blokpoel 2018; Guest 2023; Hardcastle 1996; Morgan and Morrison 1999; van Rooij and Baggio 2021). Theories need to be understandable; so if proponents understand nothing of their theory, they are in a pickle as it does no explaining. Models also by their nature should provide a sensible mediation between theory, verbal and formal, and observations (Guest and Martin 2025b, 2021, 2023; Morgan and Morrison 1999).⁹ If the important differences

between model, theory, and observation are blurred, then these need to be disentangled first and foremost. Only then can we move to discussing what understanding, if any, is provided by our formal accounts.

Finally, we can tackle what turns this all on its head, that it is indeed true that in one special sense of understanding we can appear to open the black box and continue not to possess any: if machines *are* minds (recall Figure 1). In this case **no understanding can ever be obtained**. Cummins (2000) explains this:

So, even if we are convinced that the mind is the brain, or a process going on in the brain, physical observation of the brain seems to give us data in the wrong vocabulary: synapses rather than thoughts. When we look at a brain, even a living brain, we do not see thoughts. [...] If you had a psychology camera and took a snapshot of a living brain, you would [...] see beliefs, desires, intentions, and their canonical relations. But to build a psychology camera, you would need to somehow bridge Leibniz’s Gap by correlating observed brain properties, events, and processes with beliefs, desires, and intentions, and this, at least for now, is beyond us. (p. 128)

Bridging the Leibnizian gap is the general case of so-called Marrian bridging laws. All this is impossible in principle, as mentioned above, in both engineered and biological systems (Arkoudas 2008; Ashby 1956; Guest and Martin 2025b; Guest, Scharfenberg, et al. 2025; Hardcastle 2019; Krickel 2024; Serban 2015).

If machines have a psychology (recall Figure 1), then we have sealed the deal already on understanding them (Sullivan 2022). To wit, the bar is thus so high such that neither functional (recall the black box must be opened) nor mechanistic (once we open the box, the formalisms inside are not informative) understanding are enough. The obfuscated explanandum is the human data modelled by the system (recall Figure 3). Ilya Sutskever implicitly notes this: “You could even go as far as to say that data is the fossil fuel of AI. It was like, created somehow. And now we use it.”¹⁰ So in a sense, yes, the system *is* a mind in this roundabout way: a

⁸Although we cannot know for certain as many of the industry’s ANN models as their source code is closed (Dingemanse 2025; Hao 2025; Jackson 2024; Liesenfeld and Dingemanse 2024; Liesenfeld, Lopez, et al. 2023; Maffulli 2023; Maris 2025; Mirowski 2023; Nolan 2025; Solaiman 2023; Thorne 2009; Widder et al. 2024).

⁹Relatedly, Fischer et al. (2025) discuss how understanding or how so-called explanations produced by AI systems, like ANNs, exist in a complex relationship to non-expert users.

¹⁰Credit to Alayo Tripp for locating this useful quote: <https://bsky.app/profile/phonotactician.bsky.social/post/3ldhe2bw7ls2w>

huge set of data generated by human cognition (Guest 2026). Relatedly, makeism — the idea that through building we come to understanding (van Rooij, Guest, et al. 2024) — collapses under its own weight since contemporary ANN-based AI has indeed been built, but many proponents believe this building has brought no understanding (see [APPENDIX: QUOTES ABOUT ANNs, LLMs, & AI](#)). And ironically, this conclusion with respect to it bringing us no deep theoretical scientific understanding is in fact right, but for misplaced reasons.

5 A Clash of Doctrines is an Opportunity

Whitehead wrote that a “clash of doctrines is not a disaster, it is an opportunity.” [T]he convergence of different problems and points of view may break open the compartments and stir up scientific culture. These turning points have consequences that go beyond their scientific context and influence the intellectual scene as a whole. Inversely, global problems often have been sources of inspiration to science.

Prigogine and Stengers (2018, p. 213)

In the past sections, we explained that: understanding of ANNs meets basic scientific criteria, or if not then no science does; and that in this context: the use of ‘black box’ is not only mistaken and mischievous, but also malicious or minimally marketing. What we do not understand is the data, and the processes which we believe generate it in the world — in other words, reality and its reflections and distortions in datasets (Elgin 2017; Vallor 2024). Mysterianism, therefore, about known mechanism is not only unwarranted, but anti-scientific. Besides, what chance do we have to understand anything, if we cannot understand models of our own making?

Rodney Brooks’s statement that the “the best model of the world is the world itself” (Dreyfus 2007, p. 1140; also see Cardon et al. 2018) appears to be the default view in contemporary AI. And while such a view may be entirely appropriate for engineering robots, it does not promote scientific understanding. It rests on merging theory and model, on the one hand, and the world and phenomena it contains, on the other hand. Such thoughtless mergers are the main confusions we have unpicked and tackled above, and which are captured in Figure 1. And confusions of this nature can be seen as spreading from cybernetics to connectionism and to mainstream psychology and cognitive science and back in various waves (viz. Ashby 1956; Guest and Martin 2025b; Petrick 2020). Chirimuuta (2025) also touches on this from a philosophical and historical overview through discussing the “*fallacy of misplaced concreteness*[, which] is the mistake of taking the abstractions of science for concrete reality, confusing the model with the target, the map with the territory (Whitehead 1925/1967, pp. 51–55).” (p. 246; also see ‘formal realism’, Chirimuuta 2024; and makeism, van Rooij, Guest, et al. 2024)

When proponents say ANN models’ behaviours are not predictable, we cannot but scoff since the output of the model is by definition statistical prediction. ANNs, as turbo-charged statistical models (recall their formal relation to logistic regression) can only but provide correlations. Recalling the mathematical model of the double pendulum: it escapes correlation (statistical prediction) with a physical pendulum very quickly, because both are chaotic systems. Nobody intellectually honest, however, would argue that because the double pendulum equations cannot numerically match the measurements from a physical such pendulum once in motion, that therefore we understand nothing about it.

Relatedly, if “[t]ransparency about these models is the most important thing to ensure safety [i.e. predictability]” (Melanie Mitchell, quoted in Musser 2023) then we already know what is in the black box: a lossy reflection of the data. Since we understand nothing of the data — both because science is an unfinished unending journey and because the datasets are hidden from researchers⁸ — we can easily conclude in principle that ANNs are therefore unsafe and unpredictable. Like the chaotic motions of the double pendulum, AI is not predictable in this sense, and in context not something that can be controlled. This remains the case until and unless we have better models of the data and better theories in general. Scientific models, and especially cognitive models, by definition cannot merely be statistical models of the data, but must constitute a qualitative capturing of the part of the world we want to understand.

Most importantly of all, the absconding of scientific duty by those who purport to use ANNs to replace human participants becomes blindingly obvious as such a dereliction in light of our exposition above. This is because the use of such models *other than* to statistically capture the data — as so-called AI surrogates (cf. Crockett and Messeri 2025; Dillion et al. 2023; Guest and van Rooij 2025) or equally badly as theories or black boxes to be understood (cf. Guest and Martin 2025b; van Rooij and Guest 2025a; van Rooij, Guest, et al. 2024) — will detract from actual understanding of the phenomena, such as the human participants’ cognitive capacities. To add insult to injury, not only do such false frames as examined herein waste scientists time, they also constitute red herrings since such examinations of models to understand them are fool’s errands. As per the titular expression, *mysterianism about known mechanism is mysticism* and by no means a valid scientific assumption to build on.

Ultimately, if we accept Leibniz’ gap (recall Box 1) we are indeed forced to come full circle and admit we do not understand how ANNs work. But not because we do not, functionally nor mechanistically, understand how they work, but because the referent shifted. It is the dataset (and the phenomena that we suspect generate it) that we do not understand and on which contemporary ANN-based AI so infamously depends. And so we can and should admit that slapping a

statistical test, whether a single logistic regression or a billion, onto a dataset will never bring understanding of the dataset and the world. Only science can do that. Laborious, slow and steady, deep and thoughtful scientific theorising (Stengers and Muecke 2018). To pretend otherwise, and insist on a flawed meaning of black boxing, is to shirk a responsibility that is inherently uniquely human (Chiodo 2022; Christin 2020; Devezer 2024; Rich, de Haan, et al. 2021).

As cognitive scientists, we can use this opportunity to subvert the doctrine of contemporary AI, correlationism (Guest 2026), modern connectionism (viz. Guest and Martin 2025b), and mysterianism about mechanism which aim to reduce science to the absurd. Such as when we slip and feel beguiled: “We are ten times more fascinated by clockwork imitations than by real human beings performing the same task.” (McCorduck 2004, p. 3) Instead, we can feel confident that this is actually a deep appreciation, even if misplaced, for the human cognitive labour, our endeavour of computational modelling, that goes into these machines (Guest 2026).

Our models, technologies, machines, formalisms, are wondrous reflections of cognition. But importantly, such models neither stand in for the theories nor for the phenomena between which they mediate (Morgan and Morrison 1999). In a world where the technology industry and even our colleagues reject theory building, through ignorance, mockery, and semantic shell games, we can subvert their feigned lack of understanding to our advantage. Scientific theorising has more not less value in times of scarce deep thinking, thoughtless technosolutionism, and obfuscation of the cognitive.

References

- Adam, Shaffique and Walt Peck (2005). *Double Pendulum: A Bridge between Regular Dynamics and Chaos*. [Online; accessed 2025-12-30]. URL: <https://www.chess.cornell.edu/sites/default/files/inline-files/ll-dp-manual.pdf>.
- Ahmed, Ahmed et al. (2026). *Extracting books from production language models*. arXiv: 2601.02671 [cs.CL]. URL: <https://arxiv.org/abs/2601.02671>.
- Allen, Paul G. and Mark Greaves (2011). *Paul Allen: The Singularity Isn't Near*. [Online; accessed 2025-08-15]. URL: <https://www.technologyreview.com/2011/10/12/190773/paul-allen-the-singularity-isnt-near/>.
- Amodei, Dario (2025). *The Urgency of Interpretability*. URL: <https://www.darioamodei.com/post/the-urgency-of-interpretability>.
- Amoore, Louise (2020). *Cloud ethics: Algorithms and the attributes of ourselves and others*. Duke University Press.
- Arkoudas, Konstantine (2008). ‘Computation, hypercomputation, and physical science’. In: *Journal of Applied Logic* 6.4, pp. 461–475.
- Ashby, William Ross (1956). ‘An introduction to cybernetics’. In: Bailey, Mark (2023). ‘Scientific American’. In: *Scientific American*. URL: <https://www.scientificamerican.com/article/how-can-we-trust-ai-if-we-dont-know-how-it-works/>.
- Beer, David (2023). ‘Why humans will never understand AI’. In: *BBC*. URL: <https://www.bbc.com/future/article/20230405-why-ai-is-becoming-impossible-for-humans-to-understand>.
- Beer, Stafford (1959). ‘What has cybernetics to do with operational research?’ In: *Journal of the Operational Research Society* 10.1, pp. 1–21.
- Bengio, Yoshua et al. (2023). *Pause Giant AI Experiments: An Open Letter - Future of Life Institute*. URL: <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.
- Beres, Derek (2017). *Even AI Creators Don't Understand How Complex AI Works*. URL: <https://bigthink.com/the-future/black-box-ai/>.
- Bills, Steven et al. (2023). *Language models can explain neurons in language models*. URL: <https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html>.
- Birx, H James (2009). ‘Encyclopedia of Time: Science’. In: *Philosophy, Theology, & Culture* 3.
- Blokpoel, Mark (2018). ‘Sculpting Computational-Level Models’. In: *Topics in cognitive science* 10.3, pp. 641–648.
- Bloomfield, Brian P. (1987). ‘The culture of artificial intelligence’. In: *The question of artificial intelligence*. Routledge, pp. 59–105.
- Bobrowski, L (1978). ‘Learning processes in multilayer threshold nets’. In: *Biological Cybernetics* 31.1, pp. 1–6.
- Boden, Margaret (2006). *Mind As Machine: A History of Cognitive Science Two-Volume Set*. Oxford University Press, USA.
- Brause, Saba Rebecca et al. (2023). ‘Media representations of artificial intelligence: surveying the field’. In: *Handbook of critical studies of artificial intelligence*, pp. 277–288.
- Bricken, Trenton et al. (2023). *Towards Monosemanticity: Decomposing Language Models With Dictionary Learning*. URL: <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- Bucher, Taina (2018). *If...Then: Algorithmic Power and Politics*. Oxford Studies in Digital Politics. Oxford University Press, USA.
- (2025). ‘Beyond the hype: Reframing AI through algorithms and culture’. In: *Journal of Communication* 75.1, pp. 81–84.
- Buckner, Cameron (2021). ‘Black Boxes or Unflattering Mirrors? Comparative Bias in the Science of Machine Behaviour’. In: *The British Journal for the Philosophy of Science* 74.3, pp. 681–712. URL: https://www.journals.uchicago.edu/doi/10.1086/714960#_i1.
- Bunge, Mario (1963). ‘A general black box theory’. In: *Philosophy of Science* 30.4, pp. 346–358.

- Calvello, Angelo (2023). 'We Will Never Fully Understand How AI Works — But That Shouldn't Stop You From Using It'. In: *Institutional Investor*. URL: <https://www.institutionalinvestor.com/article/2bstr1aauiex25yto45c/opinion/we-will-never-fully-understand-how-ai-works-but-that-shouldnt-stop-you-from-using-it>.
- Cao, Carla E and Vicente Raja (2024). 'Mechanisms after the end of New Mechanism'. In: *Cognitive Neuroscience* 15.3-4, pp. 100–101.
- Cardon, Dominique, Jean-Philippe Cointet, and Antoine Mazières (2018). 'La revanche des neurones'. In: *Réseaux* 211.5, pp. 173–220.
- Cassauwers, Tom (2020). 'Opening the 'black box' of artificial intelligence'. In: *Horizon Magazine*. URL: <https://projects.research-and-innovation.ec.europa.eu/en/horizon-magazine/opening-black-box-artificial-intelligence>.
- Castelvecchi, Davide (2016). 'Can we open the black box of AI?'. In: *Nature* 538.7623, pp. 20–23. URL: <https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731>.
- Chalmers, David, Anil Seth, and Brian Greene (2024). *What Creates Consciousness?* | World Science Festival. Youtube. URL: <https://www.youtube.com/watch?v=06-iq-0yJNM>.
- Chiodo, Simona (2022). 'Human Autonomy, Technological Automation'. In: *AI and Society* 37.1, pp. 39–48.
- Chirumuuta, Mazviita (2018). 'Marr, Mayr, and MR: What functionalism should now be about'. In: *Philosophical Psychology* 31.3, pp. 403–418.
- (2021a). 'Prediction versus understanding in computationally enhanced neuroscience'. In: *Synthese* 199.1, pp. 767–790.
- (2021b). 'Your brain is like a computer: Function, analogy, simplification'. In: *Neural mechanisms: New challenges in the philosophy of neuroscience*. Ed. by Fabrizio Calzavarini and Marco Viola. Springer, pp. 235–261.
- (2024). *The brain abstracted: Simplification in the history and philosophy of neuroscience*. MIT Press.
- (2025). 'Why are we still suffering from the blind spot?' In: *Phenomenology and the Cognitive Sciences*.
- Christin, Angèle (2020). 'The ethnographer and the algorithm: Beyond the black box'. In: *Theory and Society* 49.5, pp. 897–918.
- Chuan, Ching-Hua (2023). 'A critical review of news framing of artificial intelligence'. In: *Handbook of critical studies of artificial intelligence*, pp. 266–276.
- Clarke, Arthur C. (1973). *Profiles of the future: an inquiry into the limits of the possible*. New York, Harper & Row.
- Condit, Jessica (2016). *We don't understand AI because we don't understand intelligence*. URL: <https://www.engadget.com/2016-08-15-technological-singularity-problems-brain-mind.html>.
- Cooper, Richard P and Olivia Guest (2014). 'Implementations are not specifications: Specification, replication and experimentation in computational cognitive modeling'. In: *Cognitive Systems Research* 27, pp. 42–49.
- Cramer, Jan Salomon (2002). *The origins of logistic regression*. Tech. rep. Tinbergen Institute discussion paper.
- Creel, Kathleen A. (2020). 'Transparency in Complex Computational Systems'. In: *Philosophy of Science* 87.4, pp. 568–589.
- Crockett, MJ and Lisa Messeri (2025). 'AI Surrogates and illusions of generalizability in cognitive science'. In: *Trends in Cognitive Sciences*.
- Cummins, Robert (2000). '"How does it work?" versus "What are the laws?": Two conceptions of psychological explanation'. In: *Explanation and cognition*. MIT press.
- (2010). 'What is it like to be a computer?' In: First. Oxford University Press, USA.
- Curry, Rachel (2024). *Sam Altman Says OpenAI Doesn't Fully Understand How GPT Works Despite Rapid Progress*. URL: <https://observer.com/2024/05/sam-altman-openai-gpt-ai-for-good-conference/>.
- Dellsén, Finnur et al. (2024). 'What is philosophical progress?' In: *Philosophy and Phenomenological Research* 109.2, pp. 663–693. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/phpr.13067>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/phpr.13067>.
- Devezer, Berna (2024). 'There are no shortcuts to theory.' In: *Behavioral & Brain Sciences* 47.
- Dillion, Danica et al. (2023). 'Can AI language models replace human participants?' In: *Trends in Cognitive Sciences* 27.7, pp. 597–600.
- Dingemanse, Mark (2025). 'Lumo: the least open 'open' model we've seen'. In: *European Open Source AI Index*. URL: <https://osai-index.eu/news/lumo-proton-least-open>.
- Dreyfus, Hubert L. (2007). 'Why Heideggerian AI failed and how fixing it would require making it more Heideggerian'. In: *Artificial Intelligence* 171.18. Special Review Issue, pp. 1137–1160. URL: <https://www.sciencedirect.com/science/article/pii/S0004370207001452>.
- Duncan, Stewart (2012). 'Leibniz's Mill Arguments Against Materialism'. In: *The Philosophical Quarterly* 62.247, pp. 250–272.
- Dupré, Maggie Harrison (2024). *Scientists Have a Dirty Secret: Nobody Knows How AI Actually Works*. URL: <https://futurism.com/the-byte/nobody-knows-how-ai-works>.
- Durán, Juan Manuel and Karin Rolanda Jongsma (2021). 'Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI'. In: *Journal of medical ethics* 47.5, pp. 329–335.
- Egan, Frances (2017). 'Function-theoretic explanation'. In: *Explanation and integration in mind and brain science*, pp. 145–163.

- El Ghaoui, Laurent et al. (2021). 'Implicit Deep Learning'. In: *SIAM Journal on Mathematics of Data Science* 3.3, pp. 930–958.
- Elgin, Catherine Z. (1993). 'Understanding: Art and Science'. In: *Synthese* 95.1, pp. 13–28. URL: <http://www.jstor.org/stable/20117763> (visited on 04/04/2026).
- (2002). 'Art in the Advancement of Understanding'. In: *American Philosophical Quarterly* 39.1, pp. 1–12. URL: <http://www.jstor.org/stable/20010054> (visited on 02/19/2026).
- (2017). *True enough*. MIT press.
- Elish, Madeleine Clare and danah boyd (2018). 'Situating methods in the magic of Big Data and AI'. In: *Communication monographs* 85.1, pp. 57–80.
- Engelsman, Martijn (2018). *Have neural networks gone beyond our understanding?* [Online; accessed 2026-01-13]. URL: <https://www.tudelft.nl/en/delft-outlook/articles/have-neural-networks-gone-beyond-our-understanding>.
- Figdor, Carrie (2010). 'Neuroscience and the multiple realization of cognitive functions'. In: *Philosophy of Science* 77.3, pp. 419–456.
- Fischer, Simon W. S. et al. (2025). *A Taxonomy of Questions for Critical Reflection in Machine-Assisted Decision-Making*. arXiv: 2504.12830 [cs.HC]. URL: <https://arxiv.org/abs/2504.12830>.
- Flanagan, Owen (1991). 'The science of the mind'. In: MIT press. Chap. Consciousness.
- Fukushima, Kunihiro (1969). 'Visual feature extraction by a multilayered network of analog threshold elements'. In: *IEEE Transactions on Systems Science and Cybernetics* 5.4, pp. 322–333.
- Ghosh, Pallab (2025). 'The people who think AI might become conscious'. In: *BBC*. URL: <https://www.bbc.com/news/articles/c0k3700zljjo>.
- Glanville, Ranulph (2009). 'Putting the black box in place: Its status'. In: *Cybernetics and Human Knowing* 16.1-2, pp. 153–167.
- Goyal, Mohit, Rajan Goyal, and Brejesh Lall (2020). *Learning Activation Functions: A new paradigm for understanding Neural Networks*. arXiv: 1906.09529 [cs.LG]. URL: <https://arxiv.org/abs/1906.09529>.
- Griffiths, Thomas L. et al. (2025). *Whither symbols in the era of advanced neural networks?* arXiv: 2508.05776 [cs.AI]. URL: <https://arxiv.org/abs/2508.05776>.
- Guest, Olivia (2023). 'What makes a good theory, and how do we make a theory good?' In: URL: psyarxiv.com/8fxds.
- (2026). 'What Does 'Human-Centred AI Mean?' In: *Behavioral Sciences* 16.4.
- Guest, Olivia, Mark Blokpoel, and Iris van Rooij (2026). 'What the Func? Multiple Realizability Need not be Vague'. In: URL: <https://doi.org/10.5281/zenodo.19388964>.
- Guest, Olivia and Andrea Martin (2025a). *Are Neurocognitive Representations 'Small Cakes'?* URL: <https://philsci-archive.pitt.edu/24834/>.
- Guest, Olivia and Andrea E Martin (2025b). 'A metatheory of classical and modern connectionism.' In: *Psychological Review*.
- (2021). 'How Computational Modeling Can Force Theory Building in Psychological Science'. In: *Perspectives on Psychological Science* 0.0. PMID: 33482070, p. 1745691620970585. eprint: <https://doi.org/10.1177/1745691620970585>. URL: <https://doi.org/10.1177/1745691620970585>.
- (2023). 'On Logical Inference over Brains, Behaviour, and Artificial Neural Networks'. In: *Computational Brain & Behavior*. URL: <http://dx.doi.org/10.1007/s42113-022-00166-x>.
- Guest, Olivia, Natalia Scharfenberg, and Iris van Rooij (2025). 'Modern Alchemy: Neurocognitive Reverse Engineering'. In: .
- Guest, Olivia, Marcela Suarez, et al. (2025). 'Against the Uncritical Adoption of 'AI' Technologies in Academia'. In: *Manuscript in Preparation*.
- Guest, Olivia and Iris van Rooij (2025). 'Critical Artificial Intelligence Literacy for Psychologists'. In: *PsyArXiv*.
- Guidotti, Riccardo et al. (2018). *A Survey Of Methods For Explaining Black Box Models*. arXiv: 1802.01933 [cs.CY]. URL: <https://arxiv.org/abs/1802.01933>.
- Hao, Karen (2025). *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI*. Penguin Press.
- Hardcastle, Valerie Gray (1995). 'Computationalism'. In: *Synthese* 105, pp. 303–317.
- (1996). *How to build a theory in cognitive science*. State University of New York Press.
- (2019). 'The interface between psychology and neuroscience'. In: *The Routledge companion to philosophy of psychology*. Routledge, pp. 164–179.
- Harper, Tyler Austin (2025). *Artificial Intelligence Is Not Intelligent*. URL: <https://www.theatlantic.com/culture/archive/2025/06/artificial-intelligence-illiteracy/683021/>.
- Hassija, Vikas et al. (2024). 'Interpreting black-box models: a review on explainable artificial intelligence'. In: *Cognitive Computation* 16.1, pp. 45–74.
- Hattab, Helen (2011). 'The Mechanical Philosophy'. In: *The Oxford Handbook of Philosophy in Early Modern Europe*. Oxford University Press. eprint: https://academic.oup.com/book/0/chapter/293052160/chapter-ag-pdf/44513539/book_34545_section_293052160.ag.pdf. URL: <https://doi.org/10.1093/oxfordhb/9780199556137.003.0005>.
- Hills, Alison (2016). 'Understanding Why'. In: *Noûs* 50.4, pp. 661–688. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/nous.12092>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/nous.12092>.

- Householder, Alston S (1941). 'A theory of steady-state activity in nerve-fiber networks: I. Definitions and preliminary lemmas'. In: *The bulletin of mathematical biophysics* 3.2, pp. 63–69.
- Hutson, Matthew (2024). 'How does ChatGPT 'think'? Psychology and neuroscience crack open AI large language models'. In: *Nature* 629.8014, pp. 986–988.
- Jackson, Sarah (2024). *Sam Altman Explains OpenAI's Shift to Closed AI Models*. URL: <https://www.businessinsider.com/sam-altman-why-openai-closed-source-ai-models-2024-11?international=true&r=US&IR=T>.
- Jiménez, Marta (2025). *Unboxing the black box of AI*. URL: <https://www.uu.nl/en/organisation/in-depth/unboxing-the-black-box-of-ai>.
- Jobin, Anna and Christian Katzenbach (2023). 'The becoming of AI: a critical perspective on the contingent formation of AI'. In: *Handbook of critical studies of artificial intelligence*. Edward Elgar Publishing, pp. 43–55.
- Kahn, Jeremy, Alex Webb, and Mara Bernath (2018). 'How a 22-Year-Old Discovered the Worst Chip Flaws in History'. In: *Bloomberg*.
- Kaplan, Jared et al. (2020). 'Scaling laws for neural language models'. In: *arXiv preprint arXiv:2001.08361*.
- Kleinbaum, David G and Mitchel Klein (2002). *Logistic regression: a self-learning text*. Springer.
- Knight, Will (2017). 'The Dark Secret at the Heart of AI'. In: *MIT Technology Review*. URL: <https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/>.
- Kocher, Paul et al. (2019). 'Spectre Attacks: Exploiting Speculative Execution'. In: *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 1–19.
- Korsch, Hans Jürgen, Hans-Jörg Jodl, and Timo Hartmann (2008). 'The Double Pendulum'. In: *Chaos: A Program Collection for the PC*, pp. 91–113.
- Kosinski, Matthew (2024). *What Is Black Box AI and How Does It Work?* URL: <https://www.ibm.com/think/topics/black-box-ai>.
- Krickel, Beate (2023). 'Different Types of Mechanistic Explanation and Their Ontological Implications'. In: *New Mechanism*, p. 9.
- (2024). 'The new mechanistic approach and cognitive ontology—or: What role do (neural) mechanisms play in cognitive ontology?'. In: *Minds and Machines* 34.3, p. 17.
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2017). 'ImageNet classification with deep convolutional neural networks'. In: *Communications of the ACM* 60.6, pp. 84–90.
- Kulstad, Mark and Laurence Carlin (2020). 'Leibniz's Philosophy of Mind'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2020. Metaphysics Research Lab, Stanford University.
- LaFrance, Adrienne (2015). 'Not Even the People Who Write Algorithms Really Know How They Work'. In: *The Atlantic*. URL: <https://www.theatlantic.com/technology/archive/2015/09/not-even-the-people-who-write-algorithms-really-know-how-they-work/406099/>.
- Leibniz, Gottfried Wilhelm (1989). 'The Monadology'. In: *Philosophical Papers and Letters*. Ed. by Leroy E. Loemker. Dordrecht: Springer Netherlands, pp. 643–653. URL: https://doi.org/10.1007/978-94-010-1426-7_68.
- Lenzen, Timo (2026). 'Where Is A.I. Taking Us? Eight Leading Thinkers Share Their Visions'. In: *The New York Times*. URL: <https://www.nytimes.com/interactive/2026/02/02/opinion/ai-future-leading-thinkers-survey.html>.
- Levien, RB and SM Tan (1993). 'Double pendulum: An experiment in chaos'. In: *American Journal of Physics* 61, pp. 1038–1038.
- Levy, Steven (2024). *AI Is a Black Box. Anthropic Figured Out a Way to Look Inside*. URL: <https://www.wired.com/story/anthropic-black-box-ai-research-neurons-features/>.
- Lewis-Kraus, Gideon (2026). 'What Is Claude? Anthropic Doesn't Know, Either'. In: *The New Yorker*. URL: <https://www.newyorker.com/magazine/2026/02/16/what-is-claude-anthropic-doesnt-know-either>.
- Liesenfeld, Andreas and Mark Dingemanse (2024). 'Rethinking open source generative AI: open-washing and the EU AI Act'. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. Rio de Janeiro, Brazil: Association for Computing Machinery, pp. 1774–1787.
- Liesenfeld, Andreas, Alianda Lopez, and Mark Dingemanse (2023). 'Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators'. In: *Proceedings of the 5th International Conference on Conversational User Interfaces*. CUI '23. ACM, pp. 1–6.
- Litch, Mary (1997). 'Computation, connectionism and modelling the mind'. In: *Philosophical Psychology* 10.3, pp. 357–364. URL: <http://dx.doi.org/10.1080/09515089708573225>.
- Liu, Xinyue et al. (2026). *Alignment Whack-a-Mole: Fine-tuning Activates Verbatim Recall of Copyrighted Books in Large Language Models*. arXiv: 2603.20957 [cs.CL]. URL: <https://arxiv.org/abs/2603.20957>.
- Lombrozo, Tania and Daniel Wilkenfeld (2019). 'Mechanistic versus Functional Understanding'. In: *Varieties of Understanding: New Perspectives from Philosophy, Psychology, and Theology*. Oxford University Press. eprint: <https://academic.oup.com/book/0/chapter/299833282/chapter-pdf/57510173/oso-9780190860974-chapter-11.pdf>. URL: <https://doi.org/10.1093/oso/9780190860974.003.0011>.

- Longo, Anthony (2025). 'How Do Social Media Algorithms Appear? A Phenomenological Response to the Black Box Metaphor'. In: *Minds and Machines* 35.2, p. 15.
- Maffulli, Stefano (2023). *Meta's LLaMa license is not Open Source*. URL: <https://opensource.org/blog/metals-llama-2-license-is-not-open-source>.
- Maier, Holger R et al. (2023). 'Exploding the myths: An introduction to artificial neural networks for prediction and forecasting'. In: *Environmental modelling & software* 167, p. 105776.
- Maris, Jordan (2025). *Meta's LLaMa license is still not Open Source*. URL: <https://opensource.org/blog/metals-llama-license-is-still-not-open-source>.
- Martin, Andrea E and Leonidas AA Doumas (2017). 'A mechanism for the cortical computation of hierarchical linguistic structure'. In: *PLoS biology* 15.3, e2000663.
- Matten, Dirk (2026). 'Fascism as a management philosophy'. In: *Philosophy of Management*, pp. 1–29.
- McCorduck, Pamela (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. AK Peters/CRC Press.
- Medina, Eden (2011). *Cybernetic Revolutionaries: Technology and Politics in Allende's Chile*. The MIT Press.
- Milkowski, Marcin (2016). 'Integrating cognitive (neuro) science using mechanisms'. In: *AVANT. Pismo Awangardy Filozoficzno-Naukowej* 2, pp. 45–67.
- Mirowski, Philip (2023). 'The evolution of platform science'. In: *Social Research: An International Quarterly* 90.4, pp. 725–755.
- Montell, Amanda (2021). *Cultish: The language of fanaticism*. HarperCollins.
- Morgan, Mary S and Margaret Morrison (1999). *Models as mediators*. Cambridge University Press Cambridge.
- Musser, George (2023). *How AI Knows Things No One Told It*. URL: <https://www.scientificamerican.com/article/how-ai-knows-things-no-one-told-it/>.
- Naur, Peter (1985). 'Programming as theory building'. In: *Microprocessing and microprogramming* 15.5, pp. 253–261.
- Neumann, Erik (2002). *Double Pendulum*. URL: <https://www.myphysicslab.com/pendulum/double-pendulum-en.html>.
- Nguyen, Dennis and Erik Hekman (2024). 'The news framing of artificial intelligence: a critical exploration of how media discourses make sense of automation'. In: *AI & society* 39.2, pp. 437–451.
- Nizami, Lance (2020). 'Mind and Machine: at the core of any Black Box there are two (or more) White Boxes required to stay in'. In: *Cybernetics & Human Knowing* 27.3, pp. 9–32.
- Nolan, Mike (2025). 'How 'open' is open-source AI?' In: *Ada Lovelace Institute*. URL: <https://www.adalovelaceinstitute.org/blog/how-open-is-open-source-ai/>.
- Nolte, David D. (2020). 'The Ups and Downs of the Compound Double Pendulum'. In: *Galileo Unbound*. [Online; accessed 2025-08-16]. URL: <https://galileo-unbound.blog/2020/10/18/the-ups-and-downs-of-the-compound-double-pendulum/>.
- Nusinovici, Simon et al. (2020). 'Logistic regression was as good as machine learning for predicting major chronic diseases'. In: *Journal of clinical epidemiology* 122, pp. 56–69.
- Offert, Fabian and Ranjodh Singh Dhaliwal (2025). *The Method of Critical AI Studies, A Propaedeutic*. arXiv: 2411.18833 [cs.CY]. URL: <https://arxiv.org/abs/2411.18833>.
- Ohlhoff, Antje and Peter H Richter (2000). 'Forces in the double pendulum'. In: *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik: Applied Mathematics and Mechanics* 80.8, pp. 517–534.
- Oja, Erkki (1989). 'Neural networks, principal components, and subspaces'. In: *International journal of neural systems* 1.01, pp. 61–68.
- Olvera, Abi (2025). 'Why nobody can see inside AI's black box'. In: *Bulletin of the Atomic Science*. URL: <https://thebulletin.org/2025/01/why-nobody-can-see-inside-ais-black-box/>.
- Pande, Vijay (2018). 'Artificial Intelligence's 'Black Box' Is Nothing to Fear'. In: *The New York Times*. URL: <https://www.nytimes.com/2018/01/25/opinion/artificial-intelligence-black-box.html>.
- Pasquale, Frank (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.
- Patel, Dwarkesh (2025). *The Scaling Era: An Oral History of AI, 2019–2025*. Ed. by Gavin Leech. Stripe Press.
- Pearson, Jordan (2016). *When AI Goes Wrong, We Won't Be Able to Ask It Why*. URL: <https://www.vice.com/en/article/ai-deep-learning-ethics-right-to-explanation/>.
- Pedreschi, Dino et al. (2019). 'Meaningful Explanations of Black Box AI Decision Systems'. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01, pp. 9780–9784. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/5050>.
- Pellert, Max et al. (2024). 'AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories'. In: *Perspectives on Psychological Science* 19.5, pp. 808–826.
- Penn, Jonnie (2022). 'Algorithmic silence: a call to decompute'. In: *Journal of Social Computing* 2.4, pp. 337–356.
- Perrigo, Billy (2024). 'Artificial Intelligence Is a 'Black Box.' Maybe Not For Long'. In: *Time*. URL: <https://time.com/6980210/anthropic-interpretability-ai-safety-research/>.

- Peters, Uwe (2023). 'Explainable AI lacks regulative reasons: why AI and human decision-making are not equally opaque'. In: *AI and Ethics* 3.3, pp. 963–974.
- Petrack, Elizabeth R (2020). 'Building the black box: Cyberneticians and complex systems'. In: *Science, Technology, & Human Values* 45.4, pp. 575–595.
- Pfaffenberger, Bryan (1988). 'Fetishised Objects and Humanised Nature: Towards an Anthropology of Technology'. In: *Man* 23.2, pp. 236–252. URL: <http://www.jstor.org/stable/2802804> (visited on 03/31/2026).
- Piccinini, Gualtiero and Carl Craver (2011). 'Integrating psychology and neuroscience: Functional analyses as mechanism sketches'. In: *Synthese* 183.3, pp. 283–311.
- Pickering, Andrew (2014). *The Cybernetic Brain Sketches of Another Future*. University of Chicago Press.
- Polger, Thomas W and Lawrence A Shapiro (2016). *The multiple realization book*. Oxford University Press.
- Poston Jr, Dudley L, Eugenia Conde, and Layton M Field (2023). *Applied regression models in the social sciences*. Cambridge University Press.
- Pratt, Philip J. and Mary Z. Last (2015). *Concepts of Database Management*. Cengage Learning.
- Prigogine, Ilya and Isabelle Stengers (2018). *Order out of chaos: Man's new dialogue with nature*. Verso Books.
- Puzrov, Volodymyr et al. (2022). 'On the stability of the equilibrium of the double pendulum with follower force: Some new results'. In: *Journal of Sound and Vibration* 523, p. 116699.
- Ramaul, Laavanya et al. (2025). 'Rethinking how we theorize ai in organization and management: A problematizing review of rationality and anthropomorphism'. In: *Journal of Management Studies*.
- Rich, Patricia, Mark Blokpoel, et al. (2020). 'How intractability spans the cognitive and evolutionary levels of explanation'. In: *Topics in cognitive science* 12.4, pp. 1382–1402.
- Rich, Patricia, Ronald de Haan, et al. (2021). *How hard is cognitive science?* URL: osf.io/preprints/psyarxiv/k79nv.
- Richter, Peter H and H-J Scholz (1984). 'Chaos in classical mechanics: The double pendulum'. In: *Stochastic Phenomena and Chaotic Behaviour in Complex Systems: Proceedings of the Fourth Meeting of the UNESCO Working Group on Systems Analysis Flattnitz, Kärnten, Austria, June 6–10, 1983*. Springer, pp. 86–97.
- Ross, Lauren N (2020). 'Multiple realizability from a causal perspective'. In: *Philosophy of Science* 87.4, pp. 640–662.
- Rozemond, Marleen (2014). 'Mills Can't Think: Leibniz's Approach to the Mind-Body Problem'. In: *Res Philosophica* 91.1, pp. 1–28. URL: <http://dx.doi.org/10.11612/resphil.2014.91.1.1>.
- (2019). 'Leibniz on Internal Action and Why Mills Can't Think'. In: *The Leibniz Review* 29, pp. 13–40.
- Rudin, Cynthia and Joanna Radin (2019). 'Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition'. In: *Harvard Data Science Review* 1.2, pp. 1–9.
- Rumelhart, David E., R. Durbin, et al. (1995). *Backpropagation: The basic theory*. Ed. by Y. Chauvin and D. E. Rumelhart. Lawrence Erlbaum Associates, Inc, pp. 1–34.
- Rumelhart, David E., G.E. Hinton, and R.J. Williams (1986). 'Learning representations by back-propagating errors'. In: *Nature* 323, pp. 533–536.
- Rumelhart, David E. and James L. McClelland (1986). *Parallel Distributed Processing, Vol. 1: Foundations*. The MIT Press.
- Salvaggio, Eryk (2025). *The Black Box Myth: What the Industry Pretends Not to Know About AI*. URL: <https://www.techpolicy.press/the-black-box-myth-what-the-industry-pretends-not-to-know-about-ai/>.
- Sanger, Terence D (1989a). 'Optimal unsupervised learning in a single-layer linear feedforward neural network'. In: *Neural networks* 2.6, pp. 459–473.
- (1989b). 'Optimal unsupervised learning in a single-layer linear feedforward neural network'. In: *Neural networks* 2.6, pp. 459–473.
- Sanguinetti, Pablo and Bella Palomo (2024). 'An alien in the newsroom: AI anxiety in European and American newspapers'. In: *Social Sciences* 13.11, p. 608.
- Schierwagen, Andreas (2012). 'On reverse engineering in the cognitive and brain sciences'. In: *Natural Computing* 11.1, pp. 141–150.
- Schmidhuber, Jürgen (2015). 'Deep learning in neural networks: An overview'. In: *Neural networks* 61, pp. 85–117.
- Serban, Maria (2015). 'The scope and limits of a mechanistic view of computational explanation'. In: *Synthese* 192.10, pp. 3371–3396.
- Shinbrot, Troy et al. (1992). 'Chaos in a double pendulum'. In: *American Journal of Physics* 60.6, pp. 491–499.
- Siegel, Eric (2019). 'Why A.I. is a big fat lie'. In: *Big Think*. URL: <https://bigthink.com/the-future/why-a-i-is-a-big-fat-lie/>.
- Solaiman, Irene (2023). 'The gradient of generative AI release: Methods and considerations'. In: *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pp. 111–122.
- Stachowiak, Tomasz and Toshio Okada (2006). 'A numerical analysis of chaos in the double pendulum'. In: *Chaos, Solitons & Fractals* 29.2, pp. 417–422.
- Stahl, William A (1995). 'Venerating the black box: Magic in media discourse on technology'. In: *Science, Technology, & Human Values* 20.2, pp. 234–258.
- Stengers, Isabelle and Stephen Muecke (2018). *Another science is possible*. Polity Cambridge, UK.
- Stoltzfus, Jill C (2011). 'Logistic regression: a brief primer'. In: *Academic emergency medicine* 18.10, pp. 1099–1104.
- Suarez, Marcela et al. (2025). *Critical AI Literacy: Beyond hegemonic perspectives on sustainability*. URL: <https://doi.org/10.5281/zenodo.15677840>.

- Suchman, Lucille Alice (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge university press.
- Suchman, Lucy (2019). 'Demystifying the intelligent machine'. In: *Cyborg futures: Cross-disciplinary perspectives on artificial intelligence and robotics*. Springer, pp. 35–61.
- Sullivan, Emily (2018). 'Understanding; not know-how'. In: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 175.1, pp. 221–240. URL: <http://www.jstor.org/stable/45095155> (visited on 04/04/2026).
- (2022). 'Understanding From Machine Learning Models'. In: *British Journal for the Philosophy of Science* 73.1, pp. 109–133.
- Sullivan, Mark (2023). *The scary truth about AI chatbots: Nobody knows exactly how they work*. URL: <https://www.fastcompany.com/90896928/the-frightening-truth-about-ai-chatbots-nobody-knows-exactly-how-they-work>.
- Tangermann, Victor (2024). 'Sam Altman Admits That OpenAI Doesn't Actually Understand How Its AI Works'. In: *Futurism*. URL: <https://futurism.com/sam-altman-admits-openai-understand-ai>.
- Tarita, Tudor (2025). *We Don't Know How AI Works. Anthropic Wants to Build an "MRI" to Find Out*. URL: <https://www.zmescience.com/ecology/world-problems/we-dont-know-how-ai-works-anthropic-wants-to-build-an-mri-to-find-out/>.
- Thorne, Michelle (2009). *Openwashing*. URL: <https://michellethorne.cc/2009/03/openwashing/>.
- Topol, Eric (2023). *Geoffrey Hinton: Large Language Models in Medicine. They Understand and Have Empathy*. URL: <https://erictopol.substack.com/p/geoffrey-hinton-large-language-models>.
- Tsimpoukis, Panos (2025). 'Contesting dominant AI narratives on an industry-shaped ground: Public Discourse and Actors around AI in the French Press and Social Media (2012–2022)'. In: *Journal of Science Communication* 24.2, A10.
- Tsouana, Voula (2023). 'THE METHOD OF MULTIPLE EXPLANATIONS REVISITED'. In: *Epicureanism and Scientific Debates. Antiquity and Late Reception: Volume I. Language, Medicine, Meteorology*. Leuven University Press, pp. 221–256. URL: <http://www.jstor.org/stable/j.ctv34h08cj.13> (visited on 03/21/2026).
- Tully, Stephanie M, Chiara Longoni, and Gil Appel (2025). 'Lower artificial intelligence literacy predicts greater AI receptivity'. In: *Journal of Marketing*, p. 00222429251314491.
- Turkle, Sherry (1984). *The second self: Computers and the human spirit*. Mit Press.
- Vallor, Shannon (2024). *The AI mirror: How to reclaim our humanity in an age of machine thinking*. Oxford University Press.
- van Rooij, Iris and Giosuè Baggio (2021). 'Theory before the test: How to build high-verisimilitude explanatory theories in psychological science'. In: *Perspectives on Psychological Science* 16.4, pp. 682–697.
- van Rooij, Iris and Olivia Guest (2025a). 'Combining Psychology with Artificial Intelligence: What could possibly go wrong?' In: — (2025b). 'Combining Psychology with Artificial Intelligence: What could possibly go wrong?' In: *PsyArXiv*.
- van Rooij, Iris, Olivia Guest, et al. (2024). 'Reclaiming AI as a theoretical tool for cognitive science'. In: *Computational Brain & Behavior* 7.4, pp. 616–636.
- VandeHei, Jim and Mike Allen (2025). 'Why AI hallucinates: Even the companies building it can't explain'. In: *Axios*. URL: <https://www.axios.com/2025/06/09/ai-llm-hallucination-reason>.
- Von Eckardt, Barbara and Jeffrey S Poland (2004). 'Mechanism and explanation in cognitive neuroscience'. In: *Philosophy of science* 71.5, pp. 972–984.
- Wadden, Jordan Joseph (2022). 'Defining the undefinable: the black box problem in healthcare artificial intelligence'. In: *Journal of Medical Ethics* 48.10, pp. 764–768.
- Whang, Oliver (2026). 'We Don't Really Know How A.I. Works. That's a Problem'. In: *The New York Times Magazine*. URL: <https://www.nytimes.com/2026/04/15/magazine/ai-black-box-interpretability-research.html>.
- Whitehead, Alfred North (1925/1967). *Science and the Modern World*. New York: Free Press.
- Widder, David Gray, Meredith Whittaker, and Sarah Myers West (2024). 'Why 'open' AI systems are actually closed, and why this matters'. In: *Nature* 635.8040, pp. 827–833.
- Williams, R.J. and D. Zipser (1989). 'A learning algorithm for continually running fully recurrent neural networks'. In: *Neural computation* 1.2, pp. 270–280.
- (1995). *Gradient-based learning algorithms for recurrent networks and their computational complexity*. Ed. by Y. Chauvin and D. E. Rumelhart. Lawrence Erlbaum, pp. 433–486.
- Wright, Cory and William Bechtel (2007). 'Mechanisms and psychological explanation'. In: *Philosophy of psychology and cognitive science*. Elsevier, pp. 31–79.
- Xiang, Chloe (2022). *Scientists Increasingly Can't Explain How AI Works*. URL: <https://www.vice.com/en/article/scientists-increasingly-cant-explain-how-ai-works/>.
- Zednik, Carlos (2021). 'Solving the black box problem: A normative framework for explainable artificial intelligence'. In: *Philosophy & technology* 34.2, pp. 265–288.
- Zhang, Chiyuan et al. (2017). 'Understanding deep learning requires rethinking generalization'. In: *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=Sy8gdB9xx>.

APPENDIX: QUOTES ABOUT ANNs, LLMs, & AI

This appendix contains two sections. The first provides a list of quotes that contain (critically or otherwise) the claim to match *Statement 2* (and β in [Figure 1](#)): ‘ANNs models are a black box or unknowable even if opened’; and the second the claim to match *Statement 3*: ‘We understand neither human nor artificial cognition, and therefore they are similar.’

ANNs models are a black box or unknowable even if opened

A

Alechina (quoted in Jiménez [2025](#), n.p.):

If AI systems operate in a black box, we have no insight into how to fix the system if there’s a mistake

Amodei ([2025](#), n.p.):

People outside the field are often surprised and alarmed to learn that we do not understand how our own AI creations work. They are right to be concerned: this lack of understanding is essentially unprecedented in the history of technology. [And o]ur inability to understand models’ internal mechanisms means that we cannot meaningfully predict such behaviors, and therefore struggle to rule them out[.]

Amodei ([2025](#), n.p.):

In the case of AI systems, we can set the basic architecture (usually some variant of the Transformer), the broad type of data they receive, and the high-level algorithm used to train them, but the model’s actual cognitive mechanisms emerge organically from these ingredients, and our understanding of them is poor.

B

Bailey ([2023](#), n.p.):

But AI systems have a significant limitation: Many of their inner workings are impenetrable, making them fundamentally unexplainable and unpredictable.

[...]

Many AI systems are built on deep learning neural networks, which in some ways emulate the human brain. These networks contain interconnected “neurons” with variables or “parameters” that affect the strength of connections between the neurons. As a naïve network is presented

with training data, it “learns” how to classify the data by adjusting these parameters. In this way, the AI system learns to classify data it hasn’t seen before. It doesn’t memorize what each data point is, but instead predicts what a data point might be.

Many of the most powerful AI systems contain trillions of parameters. Because of this, the reasons AI systems make the decisions that they do are often opaque. This is the AI explainability problem — the impenetrable black box of AI decision-making.

Baldi (quoted in Castelvechi [2016](#), p. 23):

To Baldi, scientists should embrace deep learning without being “too anal” about the black box. After all, they all carry a black box in their heads. “You use your brain all the time; you trust your brain all the time; and you have no idea how your brain works.”

Barak (quoted in Dupré [2024](#), n.p.):

Many people in the field often compare it to physics at the beginning of the 20th century.

We have a lot of experimental results that we don’t completely understand, [...] and often when you do an experiment it surprises you.

Beer ([2023](#), n.p.):

Looking back at the history of neural networks tells us something important about the automated decisions that define our present or those that will have a possibly more profound impact in the future. Their presence also tells us that we are likely to understand the decisions and impacts of AI even less over time. These systems are not simply black boxes, they are not just hidden bits of a system that can’t be seen or understood.

It is something different, something rooted in the aims and design of these systems themselves. There is a long-held pursuit of the unexplainable. The more opaque, the more authentic and advanced the system is thought to be. It is not just about the systems becoming more complex or the control of intellectual property limiting access (although these are part of it). It is instead to say that the ethos driving them has a particular and embedded interest in “unknowability”. The mystery is even coded into the very form and discourse of the neural network. They come with deeply piled layers — hence the phrase deep

learning — and within those depths are the even more mysterious sounding “hidden layers”. The mysteries of these systems are deep below the surface.

Bengio (quoted in Pearson 2016, n.p.):

it’s exactly because we can’t mathematically pick apart a decision made by deep learning software that it works so well.

Bengio et al. (2023, n.p.):

recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one — not even their creators — can understand, predict, or reliably control.

Beres (2017, n.p.):

The computers that run those services have programmed themselves, and they have done it in ways we cannot understand. Even the engineers who build these apps cannot fully explain their behavior.

Bills et al. (2023, n.p.):

Language models have become more capable and more widely deployed, but we do not understand how they work.

[...]

Our explanations also do not explain what causes behavior at a mechanistic level, which could cause our understanding to generalize incorrectly. To predict rare or out-of-distribution model behaviors, it seems possible that we will need a more mechanistic understanding of models.

C

Calvello (2023, n.p.):

While there are some types of AI that humans can comprehend, there are others that, because of their complexity and high dimensionality, are beyond the ken of human intelligence.

D

Dupré (2024, n.p.):

not even the folks creating all this AI fully understand how it really works. [...]

In other words, AI is already everywhere. But as it’s increasingly integrated into human life,

the scientists building the tech are still trying to fully understand how it learns and functions.

Some experts chalk the lack of understanding up to the burgeoning nature of the field, arguing that AI’s nascency means that sometimes researchers will have to work backward from experimental results and outputs.

E

Engelsman (2018, n.p.):

People understand the input and output of a neural network. But its inner workings are a black box. It feels unsettling to have computers make judgements in a way that we ourselves do not understand.

H

Hutson (2024, n.p.):

Because chatbots can chat, some researchers interrogate their workings by simply asking the models to explain themselves. This approach resembles those used in human psychology. “The human mind is a black box, animal minds are kind of a black box and LLMs are black boxes,” says Thilo Hagedorff, a computer scientist at the University of Stuttgart in Germany. “Psychology is well equipped to investigate black boxes.”

J

Jobin and Katzenbach (2023, p. 48):

In consequence, we are living in a time when the infrastructures and institutions of our everyday lives are being (re)built at the hands of techniques that already elude popular and professional understanding.

K

Kaplan et al. (2020, p. 22)

At present we do not have a solid theoretical understanding for any of our proposed scaling laws. The scaling relations with model size and compute are especially mysterious. [...] Without a theory or a systematic understanding of the corrections to our scaling laws, it’s difficult to determine in what circumstances they can be trusted.

Kaplan (quoted in Patel 2025, p. 23):

we don't really know all of the details of how neural networks work. We're still very confused. There's a lot left to understand, even to just validate some of our hypotheses.

Knight (2017, n.p.):

The workings of any machine-learning technology are inherently more opaque, even to computer scientists, than a hand-coded system. This is not to say that all future AI techniques will be equally unknowable. But by its nature, deep learning is a particularly dark black box. You can't just look inside a deep neural network to see how it works.

Kosinski (2024, n.p.):

AI developers broadly know how data moves through each layer of the network, and they have a general sense of what the models do with the data they ingest. But they don't know all the specifics. For example, they might not know what it means when a certain combination of neurons activates, or exactly how the model finds and combines vector embeddings to respond to a prompt.

Even open-source AI models that share their underlying code are ultimately black boxes because users still cannot interpret what happens within each layer of the model when it's active.

[...]

If a black box model does make the wrong decisions or consistently produces inaccurate or harmful outputs, it can be hard to adjust the model to correct this behavior. Without knowing exactly what happens inside the model, users cannot pinpoint exactly where it is going wrong.

L

LaFrance (2015, n.p.):

Not Even the People Who Write Algorithms Really Know How They Work

Levy (2024, n.p.):

Even the people who build them don't know exactly how they work, and massive effort is required to create guardrails to prevent them from churning out bias, misinformation, and even blueprints for deadly chemical weapons. If the people building the models knew what happened inside these "black boxes," it would be easier to make them safer.

M

Maier et al. (2023, p. 14):

While the potential of ANNs is clear, they are still surrounded by an air of mystery and intrigue, leading to a lack of understanding of their inner workings. This has led to the perpetuation of a number of myths, resulting in the misconception that the application of ANNs primarily involves "throwing" a large amount of data at a "black-box" software package.

Mitchell (quoted in Musser 2023, n.p.):

I don't know how [LLMs a]re doing it or if they could do it more generally the way humans do—but they've challenged my views

Mitchell (quoted in Lenzen 2026, n.p.):

The misconception that A.I. has "magic" or "emergent" abilities that are impossible to understand and predict. This is mainly a view of the public (and policymakers, to some extent). Technologists and Silicon Valley often push this narrative but I don't know how much they really believe it.

Moore (quoted in LaFrance 2015, n.p.)

And it's going to get more convoluted before it gets clearer. In fact, for a few reasons, it probably won't get clearer ever. First of all, there's virtually no regulation of data-collection in the United States, meaning companies can create detailed profiles of individuals based on huge troves of personal data—without those individuals knowing what's being collected or how that information is being used. "This is getting worse," said Andrew Moore, the dean of computer science at Carnegie Mellon University.

Which means, Moore told me, we are "moving away from, not toward the world where you can immediately give a clear diagnosis" for what a data-fed algorithm is doing with a person's web behaviors. I once explored the idea that we might eventually be able to subscribe to one algorithm over another on Facebook as a way to know exactly how the information filter was working. A nice thought experiment, perhaps, but one that assumes the people who write algorithms know with any level of precision or individuality how they work.

"You might be overestimating how much the content-providers understand how their own

systems work,” said Moore, who is also a former vice president at Google. He didn’t want to talk about Google in particular, but he did present another hypothetical: Imagine a company showing movie recommendations.

Musser (2023, n.p.):

No one yet knows how ChatGPT and its artificial intelligence cousins will transform the world, and one reason is that no one really knows what goes on inside them. Some of these systems’ abilities go far beyond what they were trained to do—and even their inventors are baffled as to why. A growing number of tests suggest these AI systems develop internal models of the real world, much as our own brain does, though the machines’ technique is different.

O

Olah (quoted in Levy 2024, n.p.):

For the past decade, AI researcher Chris Olah has been obsessed with artificial neural networks. One question in particular engaged him, and has been the center of his work, first at Google Brain, then OpenAI, and today at AI startup Anthropic, where he is a cofounder. “What’s going on inside of them?” he says. “We have these systems, we don’t know what’s going on. It seems crazy.”

Olvera (2025, n.p.):

This opacity has created an unprecedented power dynamic: First, at the most fundamental level, the tech companies building these AI systems don’t fully understand how their models work internally—a challenge inherent to the technology itself. But there’s a second, distinct barrier to transparency: Developers aren’t making the data they train these systems with available to those outside their organizations. Additionally, outside researchers who have the skills and knowledge to study these systems independently lack the resources and computing power to run their own experiments, even if they had data access. With generative AI rapidly reshaping society, from medical diagnoses to classroom teaching, academic and independent researchers are pursuing parallel investigations: They hope to crack open the AI “black box” to understand its decision-making, while rigorously studying how these systems affect the real-world. Recent breakthroughs reveal that true transparency requires not just peering into AI’s inner workings,

but reimagining how society should study, evaluate, and govern these systems.

P

Pavlick (quoted in Lewis-Kraus 2026, n.p.):

Ellie Pavlick, a computer scientist at Brown, has drawn up a taxonomy of our most common responses. There are the “fanboys,” who man the hype wires. They believe that large language models are intelligent, maybe even conscious, and prophesy that, before long, they will become superintelligent. The venture capitalist Marc Andreessen has described A.I. as “our alchemy, our Philosopher’s Stone—we are literally making sand think.” The fanboys’ deflationary counterparts are the “curmudgeons,” who claim that there’s no there there, and that only a blockhead would mistake a parlor trick for the soul of the new machine. In the recent book “The AI Con,” the linguist Emily Bender and the sociologist Alex Hanna belittle L.L.M.s as “mathy maths,” “stochastic parrots,” and “a racist pile of linear algebra.”

But, Pavlick writes, “there is another way to react.” It is O.K., she offers, “to not know.”

What Pavlick means, on the most basic level, is that large language models are black boxes. We don’t really understand how they work. We don’t know if it makes sense to call them intelligent, or if it will ever make sense to call them conscious. But she’s also making a more profound point. The existence of talking machines—entities that can do many of the things that only we have ever been able to do—throws a lot of other things into question. We refer to our own minds as if they weren’t also black boxes. We use the word “intelligence” as if we have a clear idea of what it means. It turns out that we don’t know that, either.

Perrigo (2024, n.p.):

Today’s artificial intelligence is often described as a “black box.” AI developers don’t write explicit rules for these systems; instead, they feed in vast quantities of data and the systems learn on their own to spot patterns. But the inner workings of the AI models remain opaque, and efforts to peer inside them to check exactly what is happening haven’t progressed very far. Beneath the surface, neural networks—today’s most powerful type of AI—consist of billions of artificial “neurons” represented as decimal-point

numbers. Nobody truly understands what they mean, or how they work.

R

Ramaul et al. (2025, p. 799):

we advocate intellectual humility in recognizing that AI remains partly a “black box”

Rudin and Radin (2019, p. 3):

In machine learning, these black box models are created directly from data by an algorithm, meaning that humans, even those who design them, cannot understand how variables are being combined to make predictions. Even if one has a list of the input variables, black box predictive models can be such complicated functions of the variables that no human can understand how the variables are jointly related to each other to reach a final prediction.

S

Shanahan (Ghosh 2025, n.p.):

That’s worrying, says Prof Murray Shanahan, principal scientist at Google DeepMind and emeritus professor in AI at Imperial College, London.

“We don’t actually understand very well the way in which LLMs work internally, and that is some cause for concern,” he tells the BBC.

T

Tarita (2025, n.p.):

Dario Amodei stood before the U.S. Senate in 2023 and said something few in Silicon Valley dared to admit: that even the people building artificial intelligence don’t understand how it works. You read that right: AI, the technology that’s taking the entire world by storm... we only have a general idea how it works.

Now, the CEO of Anthropic—one of the world’s top AI labs—is raising that same alarm, louder than ever. In a sweeping essay titled *The Urgency of Interpretability*, Amodei delivers a clear message: the inner workings of today’s most powerful AI models remain a mystery, and that mystery could carry profound risks. “This lack of understanding is essentially unprecedented in the history of technology,” he writes.

V

VandeHei and Allen (2025, n.p.):

The wildest, scariest, indisputable truth about AI’s large language models is that the companies building them don’t know exactly why or how they work. Sit with that for a moment. The most powerful companies, racing to build the most powerful superhuman intelligence capabilities [...] don’t know why their machines do what they do.

W

Whang (2026, n.p.):

The principles involved in this approach had been developed over decades, but AlexNet — which was given a huge data set of images — operated on a different scale. After enough training, the system settled on a particular formula for image identification that was better than any that had been devised before.

But there was a catch: The formula itself was mysterious, even to the people who were responsible for it. Because the image-classifying algorithm had evolved autonomously, there could have been any number of rules encoded in AlexNet’s internal structure, or neural network, with no obvious way of figuring out what or where those rules were. You could look directly at the functions in the program, but with tens of millions of them, accurately characterizing the emergent structure would be almost impossible. The program was essentially a black box.

X

Xiang (2022, n.p.):

The people who develop AI are increasingly having problems explaining how it works and determining why it has the outputs it has.

Z

Zednik (2021, p. 271):

In order to develop such a framework, it is instructive to look at cognitive science. Despite obvious differences to the computing systems being programmed in Machine Learning, biological cognizers can equally be viewed as opaque “black boxes.” Moreover, as has already been suggested above, biological cognizers and ML-programmed computers afford a similar diversity of explanations and epistemically relevant

elements. Of course, these similarities should not come as a surprise: biological cognizers have long been viewed as computing systems in their own right (Pylyshyn 1984), and the problems of Artificial Intelligence are traditionally defined in terms of the capacities possessed by humans and other intelligent beings (Minsky, 1968).

We understand neither human nor artificial cognition, and therefore they are similar

A

Altman (quoted in Curry 2024, n.p.):

“We don’t understand what’s happening in your brain at a neuron-by-neuron level, and yet we know you can follow some rules and can ask you to explain why you think something,” said Altman. By likening GPT to the human brain, Altman reasoned a black-box presence, or a sense of mystery behind its functionality. Like human brains, generative A.I. technology such as GPT creates new content based on existing data sets and can supposedly learn over time. GPT may not have emotional intelligence or human consciousness, but it can be difficult to understand how algorithms—and the human brain—come to the conclusions they do.

B

Bengio (quoted in Pearson 2016, n.p.):

“As soon as you have a complicated enough machine, it becomes almost impossible to completely explain what it does,” Bengio said. “Think about another person or an animal—their brain is computing something with hundreds of billions of neurons. Even if you could measure those neurons, it’s not going to be an answer that you can use.”

The math at the core of deep learning systems is really pretty simple, Bengio said, but the problem is this: once they get going, it becomes too complex to make sense of. You could put all the calculations that went into making a decision into a spreadsheet, Bengio explained, but the result will just be numbers that only a machine can understand.

It’s worth emphasizing here that deep learning still runs on computers, and that means we shouldn’t completely mythologize it. Think about it this way: in the past, many people were paid to be human computers. The term “computer” contains an implicit historical continuity

that draws our attention to the fact that today’s powerful machines are doing the exact same job as these original human computers, but much faster.

“You don’t understand, in fine detail, the person in front of you, but you trust them”

C

Chalmers (in video interview, Chalmers et al. 2024, n.p.):

I think it’s possible for an AI system to be conscious. I think it’s possible for a machine to be conscious. The brain itself is a big machine. Somehow that machine produces consciousness. We don’t know how, but it does it somehow. I think if biology can do it I don’t see why silicon can’t do it. I can’t. We don’t understand how silicon could give us consciousness. We also don’t understand how neurons could give us consciousness. So I don’t see a difference in principle.

H

Hinton (quoted in Topol 2023, n.p.):

I think these chatbots, they have intuition that is what they’re doing is they’re taking strings of symbols and they’re converting each symbol into a big bunch of features that they invent, and then they’re learning interactions between the features of different symbols so that they can predict the features of the next symbol. And I think that’s what people do too. So I think actually they’re working pretty much the same way as us. There’s lots of people who say, they’re not like us at all. They don’t understand, but there’s actually not many people who have theories of how the brain works and also theories of how they understand how these things work. Mostly the people who say they don’t work like us, don’t actually have any model of how we work. And it might interest them to know that these language models were actually introduced as a theory of how our brain works.

P

Pande (2018, n.p.):

There’s particular concern about this in health care, where A.I. is used to classify which skin lesions are cancerous, to identify very early-stage cancer from blood, to predict heart disease, to determine what compounds in people and animals could extend healthy life spans and more. But

these fears about the implications of black box are misplaced. A.I. is no less transparent than the way in which doctors have always worked — and in many cases it represents an improvement, augmenting what hospitals can do for patients and the entire health care system. After all, the black box in A.I. isn't a new problem due to new tech: Human intelligence itself is — and always has been — a black box.

[...]

But we make decisions in areas that we don't fully understand every day — often very successfully — from the predicted economic impacts of policies to weather forecasts to the ways in which we approach much of science in the first place. We either oversimplify things or accept that they're too complex for us to break down linearly, let alone explain fully. It's just like the black box of A.I.: Human intelligence can reason and make arguments for a given conclusion, but it can't explain the complex, underlying basis for how we arrived at a particular conclusion. Think of what happens when a couple get divorced because of one stated cause — say, infidelity — when in reality there's an entire unseen universe of intertwined causes, forces and events that contributed to that outcome. Why did they choose to split up when another couple in a similar situation didn't? Even those in the relationship can't fully explain it. It's a black box.