

Philosophy of artificial intelligence (AI)

.. meets Philosophy of Science

Niels Linnemann (niels.linnemann@unige.ch)

23 April 2026

Matthew Schwartz on automating physics

The rules were strict: - Only give text prompts to Claude Code. No editing files directly. - Dont cut and paste my own calculations into the chat. - But pasting Gemini or GPT calculations was OK, as long as they were only text-prompted.

My question was: is there a set of prompts, like instructions to a talented G2, that can guide an AI to produce a high-quality physics paper (one that is genuinely interesting and pushes the field forward)? <https://www.anthropic.com/research/vibe-physics>

Table of Contents

- 1 A primer on 'Artificial intelligence'
 - *In medias res*: the neuronal network
 - Central terms
 - Some history
- 2 What is 'Philosophy of AI'?
- 3 Philosophy of Science meets AI

Why more layers?

- ▶ Without a hidden layer, many functions cannot be learned (such as famously the XOR function).
- ▶ In principle, from one hidden layer one (and a non-linear activation function), any continuous function can be arbitrarily approximated.
- ▶ Practically you need (many) more layers for good generalisation (among other things).

Learning the weights of the net

- ▶ The weights of the neural networked are **learned** (→ machine learning). They get updated so that for a given batch of training inputs (with known answers!) the difference between true answer and predicted answered becomes minimal.
- ▶ (Technically, one has been using backpropagation here since the 1980s.)

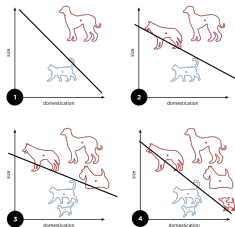


Figure: The classifier gets updated with new training data added.

Applications of neural networks

... including variants such as convolutional networks, networks with skip-connection, etc.:

- ▶ Image classification (handwriting, ...), image segmentation (tumor segmentation, pedestrian segmentation, ...)
- ▶ Building block for larger architectures ...

Table of Contents

- 1 A primer on 'Artificial intelligence'
 - *In medias res*: the neuronal network
 - Central terms
 - Some history
- 2 What is 'Philosophy of AI'?
- 3 Philosophy of Science meets AI

Some AI definitions

	Human-Based	Ideal Rationality
Reasoning-Based:	Systems that think like humans.	Systems that think rationally.
Behavior-Based:	Systems that act like humans.	Systems that act rationally.

Four Possible Goals for AI According to AIMA

Figure: Goal-oriented definitions of AI à la Russell & Norvig

Other ways of defining AI

Technical architectures/paradigms: Logic-based? Probabilistic?
Neurocomputational? Learning (supervised vs.
non-supervised)?

Weak vs. Strong AI: systems that *appear* to think/act **VS.** do
act/think like humans

General vs. Specialised AI: systems that *perform* a broad **VS.**
narrow range of tasks at the level of humans or higher

Current paradigms

1. **Supervised Learning (Discriminative AI)** Learns a mapping from inputs to labelled outputs.
Example: - Image classification (AlexNet, ResNet)
2. **Unsupervised Learning** Learns structure in data without explicit labels.
Example: - Vector representation of ordinary language (Word2Vec)
3. **Self-Supervised Learning** Generates labels from the data itself. Major advance in pretraining large models.
Example: - Next-token prediction (GPT): For the input: "Cat is on" the 'label' "mat" is assigned.
4. **Reinforcement Learning** Agents learn to act in environments to maximize cumulative reward.
Examples: - AlphaGo / AlphaZero

Current paradigms (2)

5. **Generative AI** Learns to produce novel outputs from training distributions.

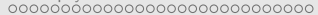
Subtypes and Models: – Diffusion models (Stable Diffusion, Midjourney, DALLE 2) - Autoregressive transformers (GPT-2/3/4, Claude, Gemini)

6. **Foundation Models** Large pretrained models that can be adapted to many downstream tasks, often trained using self-supervised or generative objectives.

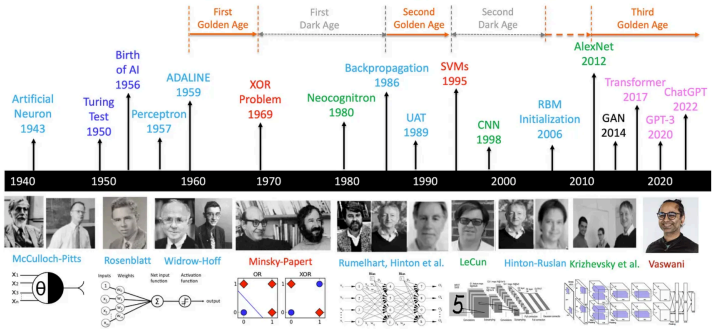
Examples: - GPT-3/4 - AlphaFold (scientific foundation model for protein structure prediction)

7. **Hybrid Systems / AI Agents** Combine reasoning, perception, memory, and action over time, often built on top of foundation models.

Examples: - BabyAGI



A rapid history of AI



Good old fashioned AI (GOF AI)

- ▶ Logic Theorist (1956) — one of the first AI programs (designed to prove theorems)
- ▶ ELIZA chatbot (1966) — one of the first chatbots
- ▶ MYCIN (1972) — one of the first expert systems

The history of AI neural networks

- ▶ 1943: McCulloch and Pitts invent the artificial neuron.
- ▶ 1958: Rosenblatt implements the special case of a perceptron.
- ▶ 1969: Minsky and Papert show limitations of one-layer perceptrons (thus contribute to AI winter).
- ▶ 1986: Hinton, Rumelhart, and Williams *popularize* backpropagation, allowing for systematic weight training.
- ▶ Early 1990s: LeCun applies CNNs to zip code digit recognition first practical success.
- ▶ 2012: AlexNet shows deep (!) nets + GPUs outperform earlier models.

AI spring (since 2010s)

- ▶ 2012 AlexNet wins ImageNet competition → Deep learning wave begins
 - ▶ 2014 GANs are introduced → Start of synthetic media generation
 - ▶ 2015 Deep Q-Networks (on Atari games) → Reinforcement learning
 - ▶ 2016 AlphaGo beats Lee Sedol
 - ▶ 2017 Transformers introduced in "Attention is all you need"
 - ▶ BERT, GPT-2, GPT-3, DALLE-E, AlphaFold, ...
-
- ▶ 2022 (30 November) AI BOOM begins with the release of Chat-GPT-3

The story of AlexNet

At the University of Toronto, Alex Krizhevsky had taken a class on writing code for GPUs, and decided to try it on neural networks. He applied himself to a popular image-recognition benchmark called CIFAR-10, which contained thumbnail-sized images that each belonged to one of ten categories: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, or truck. Krizhevsky built a network and began using a GPU to train it to categorize CIFAR-10 images. Shockingly, he was able to train his network from a random starting configuration all the way to state-of-the-art accuracy. In eighty seconds. It is at this point Krizhevskys labmate, Ilya Sutskever, takes notice and offers him what will become a kind of siren song. I bet, Sutskever says, you can make it work on ImageNet. (Christian, 2021, 20)

Example applications of transformers

Can you give the 10 most important applications for transformers in decreasing relevance?



Here's a list of the 10 most important applications of transformers, arranged in decreasing order of relevance:

1. **Text Generation:** Generating coherent, human-like text (e.g., GPT models), widely used in conversational AI, content creation, and creative writing.
2. **Machine Translation:** Translating text between languages (e.g., Google Translate), a key application in breaking down language barriers.
3. **Text Summarization:** Condensing large bodies of text into shorter, meaningful summaries, crucial for handling large information loads.
4. **Question Answering:** Providing answers to questions based on a given context, widely used in search engines and virtual assistants.
5. **Text Classification:** Classifying text into categories (e.g., spam detection, sentiment analysis), fundamental for sorting and organizing data.

Philosophy of AI?

- ▶ Another 'Philosophy of X'?!
 - ▶ Standard concern: Are there genuine novel questions in the Philosophy of AI — or just old questions in new clothing?
 - ▶ Some problems are indeed not completely new but they arise with new urgency...

Example 1: Fakes vs. deep fakes (Epistemology)

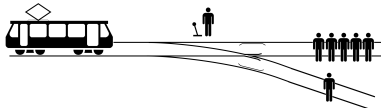


Example 2: Individual/societal biases vs. data biases (Epistemology/Ethics)



Word2Vec (2015) output for:
doctor - man + woman = nurse

Example 3: Trolley problem (Ethics)



Example 4: Black boxing (Epistemology/PhilSci)

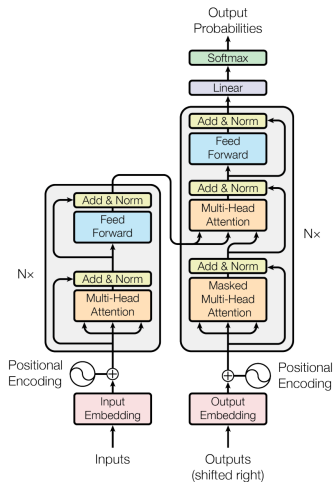


Figure 1: The Transformer - model architecture.

Philosophy of AI?

- ▶ Another 'Philosophy of X'?!
- ▶ Standard concern: Are there genuine novel questions in the Philosophy of AI — or just old questions in new clothing?
 - ▶ Addressing old problems in a new context brings new urgency — and sometimes even novel insight (including novel problems).
- ▶ But there is more to Philosophy of AI in any case...

Philosophy of AI

The discipline of AI informs philosophy (in particular: epistemology, .. of mind/of science), and philosophy (in particular: ethics, through philosophical 'toolkit') informs the discipline of AI.

Cf. to Wüthrich's definition of the Philosophy of Physics.

Philosophy of AI

The discipline of AI informs philosophy (in particular: epistemology, .. of mind/of science), and philosophy (in particular: ethics, through philosophical 'toolkit') informs the discipline of AI.

Example: Computationalism and functionalism (Philosophy of mind)

- ▶ Computational theory of mind: explain cognition and even qualia from computational processes

Example: Creativity (Epistemology)

- ▶ Understand creativity of humans from how 'creativity' gets engineered in computational systems

Example: Explanation and prediction (Philosophy of science)

- ▶ Do we need models to explain, or is prediction sufficient?

Philosophy of AI

*The discipline of AI informs philosophy (in particular: epistemology, .. of mind/of science), and **philosophy (in particular: ethics, through philosophical 'toolkit')** informs the discipline of AI.*

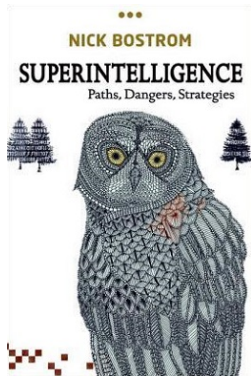
Example: 'Chinese room' against strong ('understanding') AI



1. *If a system does only syntactical manipulation, it will not acquire meaning [understanding].*
2. *A computer does only syntactical manipulations.*
3. *A computer will not acquire meaning [understanding].*
(Müller, 2024)

(read Cole (2023) for more)

Example: Bostrom's AI control problem;



- ▶ Conceptual framework for controlling 'superintelligences.'

Instead of a sharp characterisation...

- ▶ Another way to put it: 'AI' raises all sorts of questions: societal, social, technological, ethical, political, psychological, scientific, ...
- ▶ Philosophers take interests from all sorts of different angles:
 - ▶ Logic
 - ▶ Philosophy of technology
 - ▶ Ethics
 - ▶ Epistemology
 - ▶ Philopty of Science
 - ▶ ...

AI and science

- ▶ AI systems are ubiquitous in science by now.
- ▶ Some examples:
 - ▶ Physics: signal-from-noise separation, particle tracking, and anomaly detection for particle accelerator data
 - ▶ Biology: protein-folding problem
 - ▶ Medicine: clinical prediction from patient data
 - ▶ Mathematics: Physics-informed neural networks (PINNs) to solve PDEs
 - ▶ ...

Many standard applications of deep learning [...] focus on prediction—learning to output specific target values given an input. Scientific applications, on the other hand, are often focused on understanding—identifying underlying mechanisms giving rise to observed patterns in the data. (Raghu and Schmidt, 2020)

- ▶ What challenges (and chances) arise from using AI systems in scientific contexts?
- ▶ How to think of such systems as models?
 - ▶ How do they represent?
 - ▶ What about their black box character? (Opacity)
 - ▶ Does scientific realism still apply?

Table of Contents

- 1 A primer on 'Artificial intelligence'
- 2 What is 'Philosophy of AI'?
- 3 Philosophy of Science meets AI**
 - Representation
 - Opacity
 - Scientific realism

Neural networks vs. scientific models (Boge, 2022)

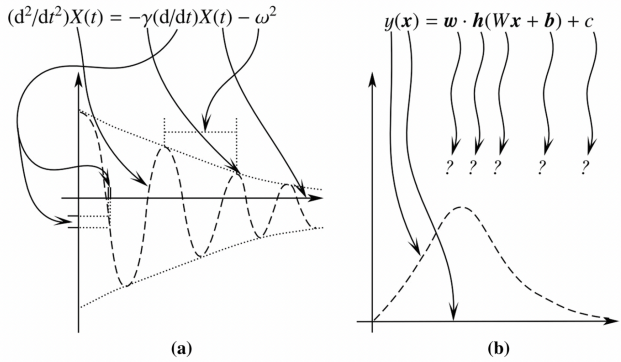


Fig. 2 Differences between the interpretation of classical mathematical models (a) and DL models (b)

(Boge, 2022)

Table of Contents

- 1 A primer on 'Artificial intelligence'
- 2 What is 'Philosophy of AI'?
- 3 Philosophy of Science meets AI
 - Representation
 - Opacity
 - Scientific realism

Opacity (Boge, 2022)

Deep neural networks are opaque in three senses:

- ▶ How-opacity: opacity with respect to **how** the machine learns.
- ▶ What-opacity: opacity with respect to **what** the machine learns.

How-opacity

- ▶ H-opacity is a special case of opacity due to complexity: the 'logic' behind the learning is per se not opaque (and thus in principle understandable).
- ▶ H-opacity does not significantly impair understanding.

What-opacity

- ▶ W-opacity occurs automatically as the model generates its own high-level features:

The surprising result of Baldi et al. (2014) was that the DNN always outperformed the other algorithms when given access only to the low-level features, and had a modest additional increase when given access also to the high-level features. The other algorithms instead exhibited major differences in performance between these situations. From this, Baldi et al. (2014, p. 7; emph. added) concluded "that [DNNs] are automatically discovering the insight contained in the high-level features". (p. XXX)

Opacity (wider literature)

Boge (2022)	Zerilli (2022)	Creel (2020)
i. How-opacity ii. What-opacity	1. Tractability 2. Fathomability 3. Intelligibility	a. Algorithmically b. Structurally c. Runtime

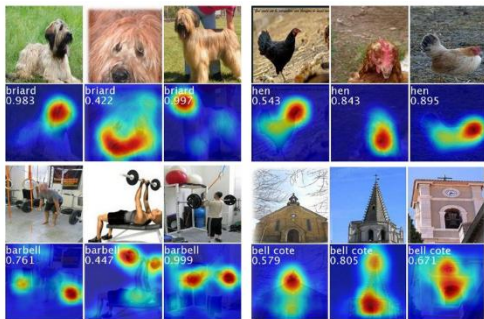
Remedy through post-hoc explanations?

The issue of opacity is of course well-familiar in AI practice. A reaction is to provide post-hoc explanations of what a net does:

Type	Method
Local Explanations	Feature Importance Explainer assigns to each feature an importance value which represents how much that particular feature was important for the prediction under analysis
	Rule Based Explicitly state the decision support system's decision boundary between the given and the contrasting advice, which can be viewed as "if... then..." statements
	Saliency Maps Generally used with image or video processing applications and is supposed to show what parts are most important to a network's decisions
	Prototypes Based A prototype is an object representing a set of similar records that the user can easily view, understand and appreciate the similarity to other validation samples
	Counterfactuals Explanation that provides a link between what could have happened had the input to a model been changed in a particular way
Global Explanations	Collection of Local Explanations Pick subset of k local explanations to constitute the global explanation after generating a local explanation for every data instance using one of the above approaches
	Representation Based Derive model understanding by analyzing intermediate representations (DNN) and determine model's reliance on 'concepts' that are semantically meaningful to humans
	Model Distillation Leverage model distillation to learn feature shapes that describe the relationship between input features and model predictions
	Summaries of Counterfactuals Construct global counterfactual explanations which provide an interpretable and accurate summary of recourses for the entire population

Figure 2: List of popular approaches for Post-hoc Explanation.

Remedy through post-hoc explanations: saliency maps



(taken from

<https://debuggercafe.com/saliency-maps-in-convolutional-neural-networks/>)

Intepretable AI

Others push for a shift to fully interpretable AI (explainable AI, or XAI in short):

Let us define what constitutes interpretability by considering how people explain to each other the reasoning processes behind complicated visual classification tasks. As it turns out, for classification of natural images, domain experts often direct our attention to different parts of the image and explain why these parts of the image were important in their reasoning process. The question is whether we can construct network architectures for deep learning that can also do this. The network must then make decisions by reasoning about parts of the image so that the explanations are real, and not posthoc. (Rudin (2019), my emphasis)

Is opacity always relevant?

Opacity of DNNs does **not** necessarily exclude

- ▶ (a) ... explanation/understanding via DNNs (Sullivan, 2022)
- ▶ (b) ... an instrumental use of DNNs as hypothesis generators (Duede, 2023)

ad (a): Link uncertainty as the central obstacle for explanation/understanding for DNNs

- ▶ Claim 1 (DNN): We can explain/understand a phenomenon *via* neural nets without being able to explain/understand all of the model.
 - ▶ Modelers have access to “wealth of knowledge and research about what methods to follow to **build** a generalizable model for the task at hand" (p. 122)
 - ▶ Saliency testing allows for high-level understanding of what the model is doing.
- ▶ Claim 2 (DNNs): Whether the explanation is actual or just possible, depends on link uncertainty.

ad (a): Link Uncertainty: Examples

Model	Input	Label	Theory	Link uncertainty
Melanoma	Skin image	Melanoma y/n	Dermatologist inspection	Little
Sexual orientation	5 facial photos	Straight y/n	Hormone theory vs. grooming cues	Yes

See: https://callingbullshit.org/case_studies/case_study_ml_sexual_orientation.html

[//callingbullshit.org/case_studies/case_study_ml_sexual_orientation.html](https://callingbullshit.org/case_studies/case_study_ml_sexual_orientation.html)

ad (b): Opaque models as hypothesis generators

- ▶ In a scientific context, opaque deep neural networks may be taken to suggest hypotheses. These hypotheses are to be tested independently though.
- ▶ In other words: neural networks are to be seen as relevant for the context of discovery, not the context of justification.

ad (b): Opaque models as hypothesis generators: an example from earthquake predictions

- ▶ A conventional model for predicting aftershocks from mainshocks had low performance.
- ▶ A trained DNN model then showed that there is room for improvement.
- ▶ A new conventional model could be proposed based on the post-hoc analysis of the trained DNN model — and successfully tested.

ad (b): Opaque models as hypothesis generators: an example from earthquake predictions

By iteratively sweeping through known geophysical properties and correlating them with DLM distributions, they find that three parameters (maximum change in shear stress, the von Mises yield criterion, and aspects of the stress-change tensor), that had not been considered by geophysicists as relevant, in fact explain nearly all of the variance in predictions generated by the neural network, thereby providing novel physical insight into the geophysics of earthquakes ... (Duede, 2023, p. 11)

ad (b): Opaque models as hypothesis generators: overall scheme

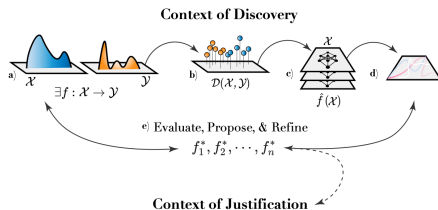


Figure 1: The confinement of epistemically opaque, neural network outputs to the context of discovery. **a)** posit the existence of some theory $\exists f$ that connects two phenomena; **b)** generate a dataset D that represents the assumed connection; **c)** train a deep learning model to learn a function that approximates the posited theory; **d)** examine the behavior of \hat{f} ; **e)** iteratively evaluate **(b)-(d)**, formulate, and refine hypotheses f_i^* connecting phenomena. Justify f^* by means distinct from those used to produce it.

Figure: Taken from Duede (2023)

Table of Contents

- 1 A primer on 'Artificial intelligence'
- 2 What is 'Philosophy of AI'?
- 3 Philosophy of Science meets AI
 - Representation
 - Opacity
 - Scientific realism

The example of AlphaFold2 (2)

- ▶ **Accuracy:** "AF2 is the first system to achieve near-atomic accuracy when compared to experimental methods."
- ▶ **Opacity:** "Despite its impressive performance, much is unknown about how AF2 works. In particular, little is known about the specific structural hypotheses AF2 learns. This is due in large part to AF2s opacity. Although AF2s high-level architecture is well-understood, it is far from clear how to translate its learned parameter values into humanly-graspable principles for identifying structure from sequence. Research in this area is ongoing."
- ▶ **The unexplained miracle:** "what needs to be explained is why AF2, given its current (trained) state, is predictively successful. The reason is that it accurately represents its target domain so the modified NMA suggests."

From using AI systems as modelling devices to the full automation of science ...

? distinguishes *overall* field of automated scientific discovery into:

- 1 AI-aided data collection and curation for scientific research
- 2 Learning meaningful representations of scientific data
- 3 AI-based generation of scientific hypotheses
- 4 AI-driven experimentation and simulation.

What lessons will there be for the philosophy of science?

References I

- Boge, F. J. (2022). Two dimensions of opacity and the deep learning predicament. Minds and Machines 32(1), 43–75.
- Christian, B. (2021). The alignment problem: How can machines learn human values? Atlantic Books.
- Cole, D. (2023). The Chinese Room Argument. In E. N. Zalta and U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy (Summer 2023 ed.). Metaphysics Research Lab, Stanford University.
- Creel, K. A. (2020). Transparency in complex computational systems. Philosophy of Science 87(4), 568–589.
- Duede, E. (2023). Deep learning opacity in scientific discovery. Philosophy of Science 90(5), 1089–1099.

References II

- Müller, V. C. (2024). Philosophy of AI: A structured overview. In N. A. Smuha (Ed.), Cambridge handbook on the law, ethics and policy of Artificial Intelligence, pp. 1–25. Cambridge University Press.
- Raghu, M. and E. Schmidt (2020). A survey of deep learning for scientific discovery. arXiv preprint arXiv:2003.11755.
- Rowbottom, D. P., W. Peden, and A. Curtis-Trudel (2024). Does the no miracles argument apply to AI? Synthese 203(5), 1–20.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence 1(5), 206–215.
- Sullivan, E. (2022). Understanding from machine learning models. The British Journal for the Philosophy of Science.
- Sullivan, E. (2023). Do ML models represent their targets?

References III

Zerilli, J. (2022). Explaining machine learning decisions.
Philosophy of Science 89(1), 1–19.

Recommended complementary resources on (philosophy of) AI

History of AI

- ▶ Wikipedia article on the History of AI — a concise overview.
- ▶ *The Quest for Artificial Intelligence* by Nils Nilsson.
- ▶ Short documentary on AlexNet:
<https://www.youtube.com/watch?v=gwzkwv2h05k>

Neural Networks and Transformers

- ▶ *Speech and Language Processing* by Jurafsky and Martin — a clear and comprehensive introduction to neural networks (chapter 7) and transformers (chapter 9)

Further recommendations

- ▶ *The Alignment Problem* by Brian Christian — explores ethical and safety issues.
- ▶ *Quand la machine apprend.* by Yann Le Cun — a compact overview (in French).