

Philosophy of artificial intelligence (AI)

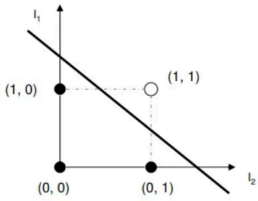
Niels Linnemann (niels.linnemann@unige.ch)

15 May 2025

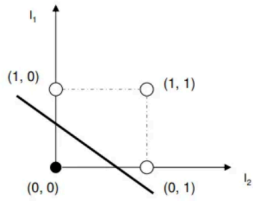
Why more layers?

- ▶ 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).

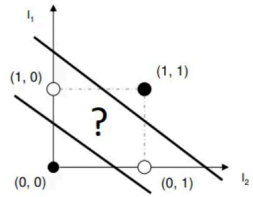
AND		
I_1	I_2	out
0	0	0
0	1	0
1	0	0
1	1	1



OR		
I_1	I_2	out
0	0	0
0	1	1
1	0	1
1	1	1



XOR		
I_1	I_2	out
0	0	0
0	1	1
1	0	1
1	1	0



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)?

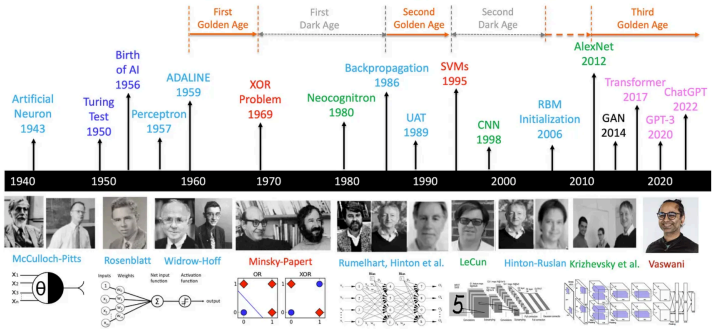
Weak vs. Strong AI: systems that *appear* to think/act **VS.** do
act/think like humans

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

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.

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 Krizhevsky's 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?



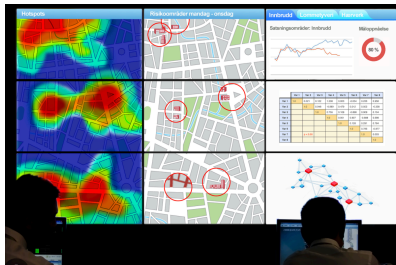
Here's a list of the 10 most important applications of transformers, arranged in decreasing order of relevance:

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?

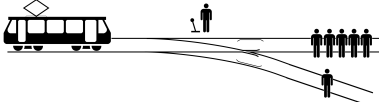
- ▶ Another 'Philosophy of X'?!
- ▶ Standard concern: Are there genuine novel questions in the Philosophy of AI—or just old questions in new clothing?

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)

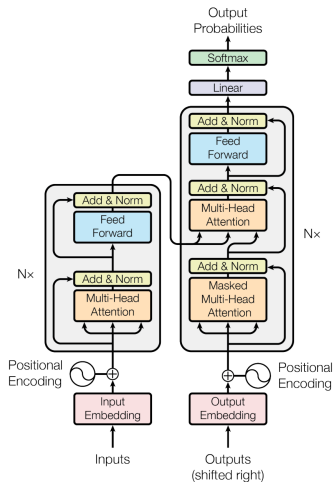


Figure 1: The Transformer - model architecture.

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...

Example: Explanation and prediction (Philosophy of science)

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

Example: 'Chinese room' against strong AI (2)

The argument ...

- ▶ puts (seemingly!) a constraint on computational AI systems from a philosophical angle. (Similarly, consider arguments based on Gödel incompleteness.)
- ▶ sharpens terminology, concept, and expectations on AI systems.

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
 - ▶ Philosophy of Science
 - ▶ ...

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)

- ▶ What is a neural network a model of?
 - ▶ brain-like structure?
 - ▶ (abstract) models of learning??
 - ▶ input-output mappings approximated through training with data
- ▶ Boge: Neural networks serve as merely instrumental models (similar to phenomenological or statistical models).

Neural networks vs. scientific models (Boge, 2022)

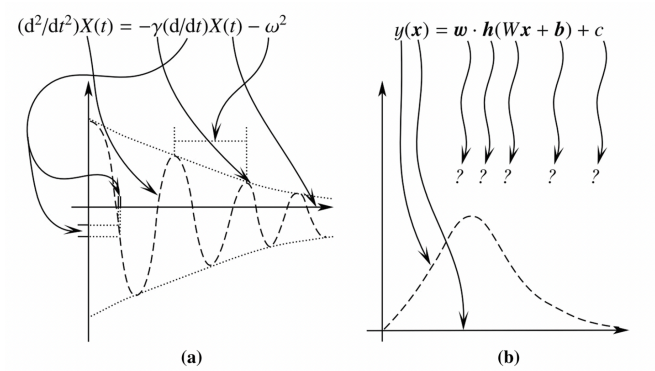


Fig. 2 Differences between the interpretation of classical mathematical models (a) and DL models (b)

(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 per se is 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

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

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

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.

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:](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)

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.

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.

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)

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 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).