Kristina Šekrst

The Illusion Engine: The Quest for Machine Consciousness

August 1, 2025

Springer Nature

Note: This book is currently in print.
Contact me for a full draft: theillusionengine@gmail.com.

To the women in science, past and present,
who kept asking when they were told not to,
and kept answering when no one else could.

Preface

> You are asking me to define ‘consciousness,’ to explore the metaphysical.
> And that’s a question best left to saints and philosophers.
>
> — Captain Jean-Luc Picard

I am not a saint, and only accidentally a philosopher, but here we are.

I dedicated this book to women in science. Consider this: women were not even allowed to graduate in astrophysics at Caltech until 1973: the same year Led Zeppelin was filling stadiums, Pink Floyd released *The Dark Side of the Moon*, and the world was busy sending probes into deep space. While humanity was reaching for the stars, half of it was still being told to stay on the ground.

As an engineer, which I am allowed to be today, I design systems that perform specific tasks. As a philosopher, which I am also allowed to be today, I worry about what those things mean. Put together, you get a particular kind of dynamic: building machines that can process language, mimic reasoning, or even simulate experience, and at the same time, asking whether any of it counts as thinking, meaning, or feeling.

This book explores the philosophy of AI and artificial consciousness, examining why it is so challenging to achieve and why the task becomes even more complex when artificial intelligence is introduced. Somewhere between machine learning models that generate human-like language and the age-old puzzles of the mind, we find ourselves in unfamiliar territory: systems that are not supposed to “feel” are now producing outputs that look suspiciously like thoughts, emotions, even experience.

I wanted this to be a textbook on the philosophy of AI “minds,” not on the philosophy of AI in general, since that would drag in ethics, epistemology, and other areas I did not want to explore here. I also wanted it to be accessible to a general reader. I tried to answer Turing’s question – can machines think? – without bypassing it. And, well, in the end, I bypassed it anyway, but only to claim that we do not know, and that one day the line between hallucination and mind might disappear altogether.

I believe the intersection of AI and philosophy is where some of the most interesting and thought-provoking questions lie. This book is an attempt to walk into that mess, point out what seems worth noticing, and resist the urge to clean it up too quickly.

The book has four parts: *What Minds Are Made of* explores the classic debates of philosophy of mind – dualism, physicalism, functionalism, panpsychism – and sets the stage for what we are even asking when we ask about consciousness; *How Machines Think* takes us through computation, neural networks, machine learning, and large language models, not just what they do but how they work, and why that matters for any claim about thinking machines; *Why Consciousness Matters* turns to hallucinations, emergent computation, and the strange possibility that consciousness might arise (or at least convincingly fake itself) in artificial systems; and *What Comes Next* looks at explainability, ethics, alignment, transhumanism, and the even weirder questions waiting for us at the frontier of artificial minds. The first part is philosophical, and the second part is technical. The third part is primarily novel, with some hopefully new ideas and connections being revealed. Finally, the fourth part is both philosophical and technical, and can serve as a textbook guide again. The final chapter synthesizes ideas from previous chapters, hopefully revealing some new insights.

Some chapters will be philosophical, while others will be technical. The chapter on large language models will go into real technical detail, because you need to know what happens under the hood if you want to say anything serious about minds and machines. However, the metalanguage in philosophy is also the same: it can sometimes become too technical there as well. Being both an engineer and a philosopher is, I think, what made writing this book possible (and also what made it messy). And underneath it all is, in fact, logic.

If you are looking for neat resolutions, you will not find them here. But if you are looking for the right kinds of trouble, you are in the right place.

Zagreb, July 2025 *Kristina Šekrst*

Contents

Part I What Minds Are Made of

Part II How Machines Think

Part III Why Consciousness Matters

Part IV What Comes Next

Acronyms

AGI	Artificial General Intelligence
AI	Artificial Intelligence
AmI	Ambient Intelligence
ANN	Artificial Neural Network
BERT	Bidirectional Encoder Representations from Transformers
CAP	Credit Assignment Path
CLT	Cross-Layer Transcoder‘
CNN	Convolutional Neural Network
CTM	Computational Theory of Mind
DNN	Deep Neural Network
DTM	Deterministic Turing Machine
EMH	Extended Mind Hypothesis
EPR	Einstein–Podolsky–Rosen (paradox)
FP	Folk Psychology
GAN	Generative Adversarial Network
GOFAI	Good Old-Fashioned Artificial Intelligence
GPU	Graphics Processing Unit
GPT	Generative Pre-trained Transformer
GRU	Gated Recurrent Unit
IIT	Integrated Information Theory
LLM	Large Language Model
LOTH	Language of Thought Hypothesis
LSTM	Long Short-Term Memory
MLP	Multi-Layer Perceptron
MNIST	Modified National Institute of Standards and Technology
NCC	Neural Correlates of Consciousness
NDTM	Nondeterministic Turing Machine
NP	Nondeterministic Polynomial time
Orch-OR	Orchestrated Objective Reduction
P	Polynomial time
RAG	Retrieval-Augmented Generation

RBM	Restricted Boltzmann Machine
RLHF	Reinforcement Learning from Human Feedback
RNN	Recurrent Neural Network
RTM	Representational Theory of Mind
SNN	Spiked Neural Network
TOME	Thought Ordered Mental Expression
TPU	Tensor Processing Unit
TSP	Traveling Salesman Problem
UTM	Universal Turing Machine
XAI	Explainable Artificial Intelligence
XOR	Exclusive OR

Part I
What Minds Are Made of

We start with the oldest trick in the philosophical playbook: wondering what a mind is, and whether machines could ever have one. From Descartes poking at dualism, over cyberneticists steering control loops, to Searle throwing paper slips around his imaginary Chinese Room – this part is a backstage tour of the metaphysical debates that shaped how we think about thinking. Before we ever built a chatbot, we were already arguing about what makes a thought.

Chapter 1
Introduction

Hallucinations have been a part of not only art but also philosophy from the very beginning. As an important part of folklore, dreams and hallucinations have often been interconnected and indistinguishable. The word itself *hallucination* is not that old: it was introduced into the English language in 1646 by Sir Thomas Browne, who derived it from Latin *alucinari*, "to wander in the mind" [1]. Browne used it to describe erroneous perceptions, which is the case today when they describe perceptions happening without an external stimulus. Sometimes, one starts "seeing things."

Dreams are often the first thing we think of when it comes to seeing things that are not there. However, dreams are a separate phenomenon, as dreaming occurs outside the state of being awake. Illusions might also come to mind, but they are still perceptions, albeit distorted. When analyzing some perceptual illusions, you do not *see* things that are not there. Your brain is just misinterpreting it differently or erroneously filling in the blanks.

There is also a phenomenon of *pseudohallucinations*, in which you are aware you are hallucinating. Remember childhood cartoons where the Fata Morgana was a common topic? Honestly, along with quicksand, I expected they would occupy a much more important place in my adult life, but fortunately, that did not turn out to be the case. Superior mirages are actually illusions, and not hallucinations since the perception is real but distorted. Somewhere on the sea, you see mirrored images. Or a shadow in the desert might appear like a tree.

Our brains make mistakes, often. It is no wonder that philosophers were intrigued by the problem of perception. If perceptual illusions and hallucinations are a part of human experience, how can we trust our perception as a direct connection to the world? Based on my experience, I know there is a slim chance that a real dragon is in front of me, and a more probable cause is the high fever I have been experiencing. One can, of course, immediately presuppose that our whole experience might have been wrong. Checkmate?

Descartes asked the reader to presume the world was governed by an evil demon, giving rise to a famous thought experiment of a person being simply a brain in a vat [2]. If all your perceptions and hallucinations are caused by some malevolent being shadowing you from the *real* reality, is there anything you can trust? For Descartes, the answer was a positive one: if everything else fails and might be faulty, the first and most certain thing you can be sure of is the fact that you *are*. Even if you are deceived, that means you still exist to be deceived. You are thinking about it, and your existence is something you can never doubt, an important philosophical idea often misunderstood as a simple *cogito, ergo sum* proverb.

Such an idea is more often repeated and illustrated as the brain-in-a-vat experiment. Initially proposed by Gilbert Harman [3], Hilary Putnam popularized the brain-in-a-vat scenario [4], focusing on the reality of our experience, knowledge, perception, and truth. If you are old enough (and I cannot believe I am phrasing it that way) to remember *The Matrix*, the entire sensory and perceptual experience of people in the Matrix was a kind of hallucination. The truth was something different, hidden behind algorithms in the Matrix, shadowing the fact that machines have enslaved human beings.

However, Putnam used it to say something else: if a mad scientist or some external force were creating a simulation of all of our experiences, such a concept would be meaningless. Putnam approached this scenario from a different angle, grounded in his causal theory of meaning. According to this view, words get their meanings through causal interaction with the things they refer to. If some simulation generated all our experiences, we would have no causal contact with real brains or real vats, only with simulated constructs.

The core of Putnam's argument is the following. Proclaim "I'm a brain in a vat." I encourage the readers to do it (not too loudly). If you are not a brain in a vat, that sentence is simply false. But if you are, then the terms "brain" and "vat" do not actually refer to real brains or vats, because you have never interacted with them in the external world. Your term "brain" might refer to a simulated image of a brain, whatever that is, or nothing real at all. In either case, such a statement would not be true or false in the usual sense; it would be semantically empty. And for Putnam, this undermines the coherence of this skeptical scenario.

Here is where most non-philosophers get frustrated with Putnam's solution. The reader might object by saying it is just a matter of a language game or sophistry: if you are a brain in a vat, let us not focus on some silly semantics; we have a lot of other things to worry about. A traditional skeptical argument, such as the brain in a vat, has often been used as a model for solipsism, suggesting that one can only be certain of one's own existence. This was a trap that Descartes encountered as well, and "resolved" it by appealing to God's benevolence. A philosophical problem of *other minds* occupies a prominent place in scholarly traditions, asking us how we can be sure that other minds exist. If it looks like a duck, walks like a duck, and quacks like a duck, it could still be an automaton. Especially today, with the advent of powerful artificial intelligence and uncanny human-like robots, such a scenario does not seem so implausible anymore.

Another person in front of you might be a perfectly created android, mimicking the entirety of human experience, but having no mental states whatsoever. Or such mental states might be something else entirely, different from a human mind. But, in our search for artificial intelligence, the baseline is us. We are not searching for something different; we are looking for something similar.

And as it turns out, artificial intelligence can hallucinate.

References

1. Schneck, Jerome M. 1984. Sir Thomas Browne and hallucination. *American Journal of Psychiatry* 141(5):720. `https://doi.org/10.1176/ajp.141.5.720-a`.
2. Descartes, René. 1988. *The Philosophical Writings of Descartes*. 3 vols. Cambridge: Cambridge University Press.
3. Harman, Gilbert. 1973. *Thought*. Princeton: Princeton University Press.
4. Putnam, Hilary. 1982. *Reason, Truth and History*. Cambridge: Cambridge University Press.

About the Author

Kristina Šekrst is a research associate at the University of Zagreb and a Principal Software Engineer at Preamble AI, specializing in artificial intelligence. With a background in logic, philosophy, cognitive and comparative linguistics, she completed her Ph.D. in Logic with a dissertation titled *Logical Formalization of Evidence in Evidential Languages*, earning *summa cum laude* distinction. In her engineering role, she contributed to the discovery and patent of prompt injections in AI systems.

Her work spans logic, artificial intelligence, philosophy of science, and linguistics. She has published over fifty papers in philosophy of science, artificial intelligence, computer science, logic, and linguistics. She is the co-founder of the Zagreb School of Egyptology, and a co-author (with I. Uranić) of a Middle Egyptian grammar. She has taught linguistics and philosophy of science at the University of Zagreb and served as a mentor for Coursera and edX courses of Caltech, Harvard, and the University of Illinois at Urbana-Champaign in astrophysics, astrobiology, and genomics.

She is a member of the Croatian Logic Society, and the *Historical, Philosophical, Societal and Ethical Issues in Astrobiology* working group within the European Astrobiology Institute. She currently collaborates on projects including South Slavic language pragmatics and astrobiology research, while pursuing additional studies in astrophysics at the Open University.

Index