

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

1	Introduction	3
	References	5
2	The Mind and the Body	7
	2.1 Mind-Body Problem	7
	2.2 Idealism	9
	2.3 Dualism	13
	2.4 Panpsychism.....	18
	References	23
3	Let's Get Physical	25
	3.1 Behaviorism	25
	3.2 Physicalism	27
	3.3 Eliminative materialism	29
	3.4 What is it like.....	31
	3.5 Colors	33
	3.6 Epiphenomenal Qualia	35
	3.7 Hard Problem of Consciousness	36
	3.8 Functionalism	39
	3.9 Representational Theory of Mind	41
	3.10 Extended Mind Hypothesis	42
	References	46
4	Cybernetics	49
	4.1 Origin and <i>Cyberspeak</i>	49
	4.2 Design for a Brain	54
	4.3 Homeostasis	58
	4.4 Cybernetics and Functionalism	60
	4.5 The Fall of Cybernetics	61

References	63
5 Artificial General Intelligence	65
5.1 Chinese Room and Intentionality	65
5.2 Strong AI and Weak AI	67
5.3 Objections and Replies to Searle	69
5.4 AI-completeness	73
5.5 Artificial General Intelligence?	75
References	78
 Part II How Machines Think	
6 The Imitation Game	83
6.1 Turing Machine and the Halting Problem	83
6.2 Cybernetics and Biology	86
6.3 Turing Test	86
6.4 Limitations of the imitations	88
6.4.1 Physical and Conceptual Constraints	88
6.4.2 The Theological Objection	89
6.4.3 The “Heads in the Sand” Objection	90
6.4.4 The Mathematical Objection	90
6.4.5 The Argument from Consciousness	91
6.4.6 Arguments from Disabilities	92
6.4.7 Ada Lovelace Objection	93
6.4.8 Argument from Continuity of the Nervous System	93
6.4.9 Argument from Informality of Behavior	94
6.4.10 Argument from Extra-Sensory Perception	94
6.5 Turing Test Today	95
References	98
7 Computation	101
7.1 Time and Space	101
7.2 P vs. NP	105
7.3 Computationalism	106
7.4 Logic	108
7.5 Design for a Brain	110
7.6 Artificial Neurons	111
7.7 Perceptron and Logic	112
References	116
8 Neural Networks and Machine Learning	119
8.1 Connectionism(s)	119
8.2 Machine Learning	120
8.2.1 Basic Concepts	120
8.2.2 Supervised Learning	124
8.2.3 Unsupervised Learning	129

8.2.4	Semi-supervised Learning	131
8.2.5	Reinforcement Learning	132
8.3	From Machine Learning to Deep Learning	133
	References	136
9	Deep Learning	139
9.1	Machine Learning and Deep Learning	139
9.2	How Deep Neural Networks Work	140
9.3	Backpropagation	146
9.4	Autoencoders	148
9.5	Types of Neural Networks	150
	References	154
10	Large Language Models	157
10.1	Transformers	157
10.1.1	How To Do Numbers With Words	158
10.1.2	Attention Is All You Need	161
10.2	Large Language Models	173
10.3	Learning from Human Preferences	174
10.4	Prompting and Prompt Injections	175
	References	179
Part III Why Consciousness Matters		
11	Consciousness	183
11.1	(Hard) Problem of Consciousness	183
11.2	Animal Consciousness	188
11.3	Qualia Objections	191
11.4	Swarm Minds	193
11.5	Emergent Computation	195
11.6	Towards Machine Consciousness	197
	References	202
12	Hallucinations	205
12.1	Brains in a Vat	205
12.2	Consciousness as a Hallucination	209
12.3	Anthropomorphizing Machine Errors	210
12.4	The Mechanics of False Generation	211
12.5	Truth-Seeking in Unreliable Systems	214
12.6	Ethical Issues	216
	References	218

13 Machine Consciousness	221
13.1 Testing Machine Consciousness	221
13.2 Can LLMs Be Conscious In Principle?	223
13.3 Can LLMs Lie?	225
13.4 Could LLMs Be Conscious Without Us Realizing It?	229
References	233
Part IV What Comes Next	
14 Explaining AI	237
14.1 What Explainable AI Is	237
14.2 What Explainable AI Is Not	239
14.3 Interpretability and Reasoning	243
14.4 Interpretability and Creativity	245
14.5 Interpretability and Medicine	246
14.6 Intentions	247
References	250
15 Quantum Minds	253
15.1 What Even Is Quantum	253
15.2 Quantum Computing	257
15.3 Quantum Mind?	258
15.4 From Quantum Brains to Quantum AI	260
References	261
16 Artificial Minds, Human Ethics	263
16.1 AI? Ethics?	263
16.2 Human Compatible Guardrails	264
16.3 AI Alignment	266
16.4 When Does My Smart Fridge Deserve Rights?	269
16.5 Transhumanism	272
16.6 Alignment Faking	273
References	276
17 Hard Problem of Consciousnesses	279
17.1 Embodied Cognition	279
17.2 Disembodied Cognition	282
17.2.1 Body	282
17.2.2 Proto-Qualia	283
17.2.3 Hard Problem of Consciousnesses	285
17.3 The Search for Other Minds	287
17.4 Final Tokens	288
References	291
About the Author	293

Index	295
--------------------	-----

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 ⁴
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

A

A-consciousness 185
A-machine 84
Aaronson, Scott 73, 105, 257
Absent qualia 192
Access consciousness 185
Actroid 88
Adamatzky, Andrew 281
AI alignment 266
AI ethics 263
AI hallucination 210, 211, 213, 225
AI safety 266
Algorithm 101
Alignment faking 273
Ambient intelligence 89
Animal consciousness 188
Antinativism 27
Artificial consciousness 221
Artificial general intelligence 68
Artificial intelligence 68
Artificial neural network 107
Artificial neuron 111
Artificial stupidity 76
Ashby, William Ross 52, 54, 110, 193, 267, 280
Asimov, Isaac 264
Astrobiology 89, 287
Attend (transformers) 164
Attention head 161
Attribute (machine learning) 123
Attribution graph 243
Autoencoder 148
Automata theory 112
Automatic machine 84
Ayumu 93

B

Baars, Bernard 199
Backpropagation 146
Bayes' theorem 242
Behaviorism 26, 92
Berkeley, George 9, 67
Berlin, Brent 42
BERT 173
Bias 122
Bias-variance tradeoff 122
Big O notation 102
Bit 257
Black box 238
Blackbody radiation 254
Block, Ned 40, 192
Bostrom, Nick 224, 272
Brain in a vat 3, 207
Brain simulator reply (Searle) 71
Brute-forcing 104

C

Causal theory of reference 208
Chain rule 146
Chalmers, David 8, 36, 37, 42, 71, 190, 191, 197, 224, 273, 281
Characteristica universalis 107
China brain 192
Chinese Room 66, 96
Churchland, Patricia 29
Churchland, Paul 29
Clark, Andy 42, 273, 281
Classification 124
Claude 174, 224, 247, 273
Clever Hans 125
Cluster 130

Cogito, ergo sum 4, 13
 Cognitive convenience 42
 Cognitive scaffolding 273
 Color 42
 Combination reply (Searle) 71
 Complexity class 102
 Computation 101
 Computation path 103
 Computational complexity 73
 Computational proto-qualia 284
 Computational theory of mind 106
 Computationalism 106
 Connectionism 119
 Consciousness 184
 Constant time 104
 Context window 172
 Controlled hallucination 209, 231, 289
 Convolution 143
 Convolutional neural network 150
 Creature consciousness 184
 Credit assignment path 140
 Crick, Francis 190
 Cross-layer transcoder 243
 Curse of dimensionality 241
 Cybernetics 50, 281, 285
 Cyberspeak 51

D

Dancing qualia 198
 De Broglie, Louis 255
 Decision problem 102
 Decision tree 126
 Decoder (autoencoder) 148
 Decoherence 259
 Deep learning 140
 Deepfake 151
 DeepSeek 224
 Dennett, Daniel 31, 38, 247
 Deontology 264
 Descartes, René 4, 13, 55, 89
 Deterministic Turing machine 74, 103
 Dick, Steven 287
 Dimension 159
 Discriminator (GAN) 151
 Disposition 26
 Distribution shift 267
 Dot product 166
 Double-slit experiment 255
 Dualism 8, 13, 283

E

Effective altruism 264

Effective procedure 106
 Einstein, Albert 255, 256
 Eliminative materialism 29
 Eliminativism 29
 ELIZA 95
 Embedding 158
 Embodied cognition 279
 Embodied mind 186
 Enaction 280
 Enactivism 186
 Encoder (autoencoder) 148
 Epiphenomenal qualia 33
 Epiphenomenalism 17
 EPR paradox 256
 Esse est percipi 10
 Essentialism 56, 124
 Ethics of artificial intelligence 263
 Evil demon 13
 Exclusive disjunction 113
 Explainability 238
 Explainable AI 237
 Explanatory gap 37
 Exploration-exploitation dilemma 132, 229
 Exponential growth 103
 Exponential time 104
 Exponential-time solution 74
 Extended mind hypothesis 42, 273, 281
 Extrinsic hallucination 212

F

Factorial time 105
 Fading qualia 197
 Feature 123
 Feature engineering 140
 Feature patching 244
 Feature space 123
 Feedforward neural network 147
 Fine-tuning 172, 266
 Fodor, Jerry 40, 41
 Folk psychology 29
 Forward pass 244
 Frame problem 267
 Functionalism 40, 60, 111, 286

G

Gödel's incompleteness theorem 258
 Gödel, Kurt 90
 Gated recurrent unit 151
 Generative adversarial network 151
 Generator (GAN) 151
 Gerovitch, Slava 51, 61
 GOF AI 120

GPT 174

Graphics processing unit 222

Grover's algorithm 258

Grover, Lov 258

Gulliver 95

H

Hallucination 3, 205

Halting problem 84

Hameroff, Stuart 258

Hard problem of consciousness 8, 36, 184, 191, 286

Harman, Gilbert 4, 207

Haugeland, John 120

Hidden variables 256

Hinton, Geoffrey 148

Homeostat 58, 60, 110

Hopfield, John 148

Human compatible 265

Hybrid computing 281

I

Idealism 9

Identity theory 28

Illusionism 38

Imitation game 86

Incompleteness 90

Incompleteness theorems 90

Inconsistency 90

Information theory 54

Intentional stance 247

Intentionality 66, 68

Interactionism 13

Interoception 209

Interpretability 238, 243, 274, 283

Intrinsic hallucination 212

Inverted qualia 191

J

Jackson, Frank 33

Jakobson, Roman 52

K

K-means 131

K-nearest neighbors 240

Kay, Paul 42

Key vector 166

Knowledge argument 35

Knowledge graph 215

Koch, Christof 190

L

Löwenheim–Skolem theorem 207

LaMDA 223

Language of thought 41

Latent representation 243

Latent space 149

Law of requisite variety 58

Layer (neural networks) 141

Layer normalization 170

Leibniz's mill 32

Leibniz, Gottfried 14, 19, 50, 107, 146

Lemoine, Blake 223

Levine, Joseph 37

Life 69, 89

Linear regression 126

Linear time 104

LLM+ system 224

Logarithmic time 104

Logic 90, 108, 111

Logical positivism 11

Logistic regression 126

Long short-term memory network 151

Lovelace, Ada 93

Lucas–Penrose constraint 91

M

Machine consciousness 221

Machine functionalism 40

Machine learning model 122

Malebranche, Nicolas 15

Mary, the scientist 35

Masked language modeling 173

Master Argument 10

Materialism 16

Matrix 145

Matsuzawa, Tetsuro 93

McCulloch, Warren 107, 111

McCulloch–Pitts neuron 111

Mechanicism 55

Meditations 13

Mentalese 41

Merleau-Ponty, Maurice 280

Microtubule 259

Mind-body problem 8, 283

Minimal cognition 285

Minsky, Marvin 37, 113

Mirror test 189

Mitchell, Tom 121

MNIST dataset 141

Model 121

Monadology 19

Monism 8

Moral patienthood 269
 Moral status 269
 Mori, Masahiro 88
 Morphogenesis 86
 Morphological freedom 272
 Multi-layer perceptron 244
 Multiple realizability 40, 54

N

Naïve Bayes 241
 Nagel, Thomas 19, 32
 Narrative consciousness 185
 Nativism 27
 Neocognitron 150
 Neural correlates of consciousness 190
 Neuromorphic computing 281
 Neutral monism 37
 Newspeak 51
 Next sentence prediction 174
 Non-linear separability 114
 Nondeterministic Turing machine 103
 NP complexity class 103
 NP problem 74, 103
 NP-complete problem 75
 NP-hard problem 74

O

O3 224
 Objective Reduction 259
 Occasionalism 15
 Orchestrated Objective Reduction 258
 Overfitting 122

P

P complexity class 103
 P problem 74, 103
 P vs. NP 75
 P-consciousness 185
 Panpsychism 19
 Papert, Seymour 113
 Parallelism 14
 Penrose, Roger 258
 Perceptron 112
 Phenomenal consciousness 185
 Philosophical zombie 36, 192
 Physicalism 28
 Pineal gland 13
 Pitts, Walter 107, 111
 Planck's constant 254
 Planck, Max 254
 Podolsky, Boris 256

Polynomial-time solution 74
 Postbiological universe 287
 Problem of other minds 4, 10, 91, 92, 288
 Prompt 175
 Prompt engineering 176
 Prompt injection 176
 Prompt tuning 215
 Property dualism 16
 Protocol sentence 12
 Prototype theory 129
 Prototypical representations 243
 Pseudohallucination 3
 Psychofunctionalism 193
 Putnam, Hilary 4, 39, 207

Q

Quadratic time 104
 Qualia 8, 31, 183, 191, 222, 283
 Quantization 254
 Quantum bit 257
 Quantum consciousness 258
 Quantum entanglement 256
 Qubit 257
 Query vector 165
 Quining qualia 38

R

Ratio Club 59
 Rectified Linear Unit 150
 Red-teaming 268
 Reductive materialism 28
 Regression 124
 Reinforcement learning from human feedback
 133, 174, 267
 Reliabilism 214
 Representation 41
 Representational theory of mind 41
 Representationalism 41
 Residual connections 170
 Restricted Boltzmann machine 149
 Retrieval augmented generation 215
 RoboMary 39
 Robot reply (Searle) 71
 Rosen, Nathan 256
 Rosenblatt, Frank 112
 Russell, Stuart 265
 Ryle, Gilbert 26

S

Safety evaluation 268
 Samuel, Arthur 121

Scalable oversight 267
 Schrödinger's cat 256
 Schrödinger, Erwin 256
 Searle, John 66
 Self-attention mechanism 161
 Self-consciousness 184
 Self-refinement 215
 Self-supervised learning 173
 Sentience 184
 Seth, Anil 209, 231, 289
 Shallow neural network 147
 Shannon, Claude 54
 Shared intentionality 43
 Sheldrake, Rupert 20
 Ship of Theseus 272
 Shor's algorithm 258
 Shor, Peter 258
 Skip connections 170
 Smart, J. J. C. 28
 Sobel filter 143
 Softmax 167
 Solipsism 12
 Source-reference divergence 212
 Space (complexity) 74, 102
 Spectrum inversion 191
 Spiking neural network 281
 Square-root time 104
 Stable equilibrium 58
 State consciousness 184, 185
 Strong AI 67, 68
 Subjective idealism 9
 Substance 13
 Superintelligence 68
 Superposition 255
 Supervenience 28
 Supervised learning 124
 Support vector machine 126
 Sutskever, Ilya 222
 Symbol grounding problem 66
 Synthetic consciousness 221
 System prompt 175
 Systems reply (Searle) 69

T

Temperature (LLM) 226
 Tensor 145
 Tensor processing unit 222
 Thought ordered mental expression 41
 Time (complexity) 74, 102
 Token 158
 Token-identity physicalism 39
 Tokenization 158, 162
 Tononi, Giulio 187

Top- k sampling 225
 Top- p sampling 226
 Training 122
 Training and test set 122
 Transformer 158
 Transhumanism 272
 Travelling salesman problem 195
 Turing machine 84, 102
 Turing test 86
 Turing, Alan 83, 110
 Type-identity physicalism 39

U

Ultraviolet catastrophe 254
 Uncanny valley 88
 Unconventional computing 281
 Underfitting 123
 Universal Turing machine 106
 Unsupervised learning 129
 Utilitarianism 264

V

Value alignment 264
 Value vector 168
 Vanishing gradient problem 151
 Variance 122
 Vector 123, 145
 Virtue ethics 264
 Vital variables 56
 Von Neumann, John 112

W

Wakefulness 184
 Wave-particle duality 255
 Weak AI 68
 Weight (neural networks) 141
 Weizenbaum, Joseph 95
 What is it like 32, 184, 284
 Wiener, Norbert 50, 267, 280
 Word embedding 159
 Word vector 159
 Word2vec 158

X

XAI 237
 XOR problem 113

Y

Young, Thomas 255