Other Books by Paul Thagard

1986 *Induction: Processes of inference, learning, and discovery*, with John Holland, Keith Holyoak, and Richard Nisbett.

1988 *Computational Philosophy of Science*.

1992 *Conceptual Revolutions*.

1995 *Mental Leaps*: *Analogy in Creative Thought,* with Keith Holyoak.

1996 *Mind: Introduction to Cognitive Science,* second edition 2005.

1999 *How Scientists Explain Disease.*

2000 *Coherence in Thought and Action.*

2006 *Hot Thought: Mechanisms and Applications of Emotional Cognition*.

2010 *The Brain and the Meaning of Life*.

2012 *The Cognitive Sience of Science: Explanation, Discovery, and Conceptual Change.*

2019 *Brain-Mind: From Neurons to Consciousness and Creativity*.

2019 *Mind-Society: From Brains to Social Sciences and Professions.*

2019 *Natural Philosophy: From Social Brains to Knowledge, Reality, Morality, and Beauty*.

2021 *Bots and Beasts: What Makes Machines, Animals, and People Smart?*

2022 *Balance: How It Works and What It Means*.

2024 *Falsehoods Fly: Why Misinformation Spreads and How to Stop It*.

2025 *Dreams, Jokes, and Songs: How Brains Build Consciousness*

January 7, 2026

AI Boom or Doom?
Philosophy and Psychology of the New Artificial Intelligence

In memory of Ziva

CONTENTS

Glossary

Notes

Preface

In 2023, I asked one of the new AI models to write a poem about artificial intelligence, and the elegant response included these chilling lines:

As its circuits hum with thoughts so deep,
We ponder, are we the shepherds or the sheep?

The prospect of humans becoming sheeplike slaves herded by machine overlords is an extreme worry intensified by recent advances in artificial intelligence, and the possibility of complete human extinction is even more extreme. These prospects are far-fetched, but the new AI models are already raising difficult questions about knowledge, mind, agency, values, creativity, and regulations to shape future developments. These questions are deeply philosophical, dependent on general and normative issues concerning the nature of thought, reasoning, morals, art, and politics. This book is an intensive investigation of the social, scientific, technological, and humanistic consequences of recent developments in AI.

My views reflect more than 50 years of philosophical research, and more than 40 years of building computer models that approximate to human intelligence. I think that the arrival of ChatGPT and similar models is a pivotal moment in technology, and indeed in all of human history. Whether the resulting critical transition is towards utopia or disaster is still under human control, and philosophical reflection should be one of the methods to ensure that people remain shepherds rather than sheep.

Philosophy should help to shape the development of artificial intelligence, but the relationship is reciprocal. Philosophy is needed to guide AI, but must also adapt and respond to startling changes in machine abilities. The new AI has important implications for traditional philosophical questions about knowledge, mind, values, art, and politics. My naturalistic approach to philosophy allies it closely with psychology, and the new AI provides surprising lessons about the nature of human minds and their accomplishments.

The new AI has already spurred the publication of dozens of books, but this book is different in offering both philosophical depth and psychological insight. Novel contributions include:

- Principled investigation of the prospects of the new AI for achieving knowledge and intelligence, taking into account recent developments such as chain-of-thought reasoning.
- Appraisal of general intelligence and superintelligence based on a comprehensive theory of the features and mechanisms of intelligence.
- Assessment of the emerging "agentic" AI and robotic applications of AI.
- Review of areas in which AI can produce human benefits, and of areas of great risk.
- Evaluation of the most plausible causal scenarios that could lead to human subordination or extinction.
- Application of a new theory of government regulation to artificial intelligence.
- Evaluation of what the successes of the new AI tell us about human psychology.
- General philosophical examination of the new AI, including development of a novel solution to the mind-body problem inspired by the hardware-software combinations required for current AI systems.

I have tried to make this book useful for three classes of readers. First, it should introduce people in general to the new AI and its philosophical significance. Chapter 2 provides a gentle introduction to how the new models work, and more details are provided when they are relevant to the discussion of the social, political, and personal significance of the new AI. The book also serves as an illustration of how public philosophy can be rigorous, flexible, and highly relevant to sorting out pressing contemporary problems.

Second, the book should be helpful to technologists and managers who are wrestling with questions about how AI can and should develop. A few irresponsible leaders, motivated by greed for power and money, are blasting full-speed ahead in the competitive race to produce the most powerful models. But many others are aware of potential dangers of the new technology, and are asking philosophical questions about what they are doing. This book delivers a broad and rigorous guide to the questions about morality and knowledge that this vibrant technology is raising. I provide everything that AI developers always wanted to know about philosophy but were afraid to ask.

Third, I have tried to make this book useful for legislators at both national and international levels who need to establish policies and laws concerning future work on AI. The politics of AI require familiarity with technological prospects and also with philosophical issues concerning the appropriate roles of governments. I argue for regulations that are urgently required to guide the future of AI in the service of human needs. I am neither a “doomer” who believes that AI will inevitably destroy humanity nor a “boomer” who sees AI as overwhelmingly beneficial. Doom or boom is still a choice to be made by people and their governments.

A glossary summarizes key concepts explained more thoroughly in the text. Live links to the web references in the notes are provided at paulthagard.com.

Acknowledgments

I am grateful to anonymous referees for comments on an earlier draft. For helpful suggestions on an article that became chapter 9, I thank Chris Eliasmith, Keith Holyoak, Jay McClelland, Lorenzo Magnani, Dean Simonton, and 2 anonymous referees. Chapter 9 and part of chapter 2 originated in the following article, reprinted by permission of Springer Nature:

Paul Thagard, "Can ChatGPT Make Explanatory Inferences? Benchmarks for Abductive Reasoning." In *Abductive Minds: Essays in Honor of Lorenzo Magnani, Vol. 1.*, edited by Selene Arfini, 189-218 (Cham, Switzerland: Springer Nature, 2025).

I have recycled some material from my *Psychology Today* blog, *Hot Thought*, for which I hold copyright.

For research assistance, I have used various AI models, especially ChatGPT and Grok. This book contains no sentences generated by these models, although chapters 2, 5, 7, and 9 contains some images produced by ChatGPT as illustrations of its capabilities. All references suggested by AI models have been checked against Google Scholar and the original sources.

Thanks to Philip Laughlin and Haley MacArthur for editorial support, XXX for copyediting, and XXX for the index.

Chapter 1
Artificial Intelligence and Philosophy

In February, 2025, Elon Musk described his new AI model Grok 3 as "scary smart."[1] Musk's pronouncements are not always to be trusted, as he annually announces that self-driving Tesla cars are only a year away. In this case, however, he was right because Grok and other AI models have become smart enough to be scary for the risks they pose to humans.

My 2021 book *Bots and Beast,* performed a comprehensive evaluation of leading AI models with respect to mental mechanisms that support intelligence. I gave low ratings to all leading AI models with respect to benchmarks that included use of images, concepts, rules, analogies, emotions, language, intentional action, and consciousness. I concluded that the enterprise of filling in the gaps in AI performance would likely take centuries rather than decades.[2] I have never been so wrong so fast.

My evaluation became obsolete with the release of ChatGPT 3.5 by OpenAI in November, 2022. I was astonished by the model's ability to answer difficult questions with responses that were well written and often insightful, although occasionally just wrong. A few months later, ChatGPT 4 was released with dramatically better performance, and similar models have since appeared from companies that include Google (Gemini), Anthropic (Claude), Meta (Llama), Mistral (Le Chat), xAI (Grok), and the Chinese firm High-Flyer (DeepSeek). In 2025, new models appeared with increasingly more sophisticated reasoning abilities. In August, 2025, ChatGPT was the world's fifth most visited website.[3]

ChatGPT and other new AI models are already smarter than you and me in many respects. None of us is capable of performing all the following feats:[4]

- Pass the medical exam to qualify as a doctor in the US.
- Pass the bar exam to qualify as a lawyer in the US.
- Program workable computer code in dozens of programming languages.
- Translate between English and a hundred natural languages.
- Solve problems in many branches of mathematics.
- Answer questions in any scientific field.
- Tutor learners in math, science, history, and other fields.
- Compose poems in any standard style, from limericks to sonnets.
- Write short stories on any topic.
- Generate artistic images.
- Construct philosophical arguments.

This list could easily be expanded to make it even clearer that a new class of intelligent beings has entered the universe, just since 2022.

Thousands of companies are using this new technology in applications that range from improving human health to automating war. Hundreds of millions of people are using the new AI models on a weekly basis. Scientists, technologists, and politicians are pondering the consequences of the new AI for social issues such as employment, disinformation, military uses, and human extinction.

My examination of the new technology applies philosophical ideas about knowledge, reality, morality, art, and political control. More radically, I argue that the advent of machines approaching human intelligence is a pivotal point with major implications for philosophical thought, as well as for social developments. History evolves through feedback loops in which new ideas bring about social changes, and social changes bring about new ideas.[5] For example, the rise of science and technology in 17th-century Europe changed society through changes in industry and social relations, but also led to new philosophical ideas about how societies do and should operate. I propose that the development of human-level AI is similarly momentous, in ways that may bring great social and intellectual benefits, but may instead be catastrophic to human well-being.

Why the New AI Matters

In *Measure for Measure,* Shakespeare writes of human frailty:

Man, proud man,
Dressed in a little brief authority,
Most ignorant of what he's most assured—
His glassy essence—like an angry ape,
Plays such fantastic tricks before high heaven
As makes the angels weep.

The phrase “glassy essence” contrasts human pretensions of power and authority with our inescapable fragility and vulnerability.

The glassiness of our essence is intensified by the arrival of machines that are close to our intelligence, challenging comfortable views of the superiority of our species. ChatGPT can already write better poetry and short stories than most people, and similar programs can compose music better than most people. ChatGPT is better at scientific reasoning than people without an advanced university degree. ChatGPT can also operate in dozens of human languages, and produce decent code in dozens of programming languages. Maintaining an elevated view of humanity based on our intelligence and creativity is difficult, when a bank of computers in a data center in California can do as well as most of us.

The new AI is not just a challenge to human self-conceptions, because it also threatens to have an impact on the most important aspects of our lives. People's sense of well-being depends in part on satisfaction of a need for competence, which often comes from achievements and accomplishments connected to working.[6] Some jobs are already being deeply affected, for example the lay-off of hundreds of customer service agents being replaced by AI chatbots. Software engineers find ChatGPT useful for producing computer code, but worry that advances in automatic programming may render their skills obsolete. AI is becoming capable of analyzing vast amounts of data and predicting trends and outcomes, potentially replacing some of the most important tasks of managers.

Another crucial aspect of human well-being is relatedness to other people, including romance, family, and friendship. The ability of the new AI models to carry on plausible conversations with people has led to the proliferation of relationship programs such as AI girlfriends. Relationships have already suffered from the tendency of young people to interact more with their phones than other people, and the trend for people to try to overcome loneliness by AI conversations takes them even farther from real human connection.

The third major human need besides competence and relatedness is autonomy, and the new AI can threaten freedom. Various scenarios see AI being used for general harm, including disinformation, autonomous weapons, and decision making by evil leaders in government and business. The worst outcomes can be summed up as the "AIpocalypse" that includes environmental degradation resulting from the huge energy demands of generative AI models, and global nuclear war produced by miscalculations by AI models operating in competing countries.[7] These scenarios are more plausible than the prospect that the new AI will take control of human life so totally that we survive only as slaves.

These threats, however, should not conceal the large benefits that the new AI can have for human life. Many of us already use ChatGPT and its competitors as research assistants, and also for advice about practical matters such as cooking and plumbing. Some of the companies that have been recently founded are concerned with dangerous consequences such as autonomous weapons, but many are aiming at improvements in human life in areas such as medicine, education, climate change, and so on. Chapter 3 provides detailed examples.

The new AI is already having substantial positive and negative effects on human lives, and we can reasonably expect these effects to increase dramatically in future years and decades. Hence AI cries out for philosophical examination.

Why Philosophy Matters

In their book *The Grand Design,* Stephen Hawking and his fellow physicist Leonard Mlodinow declare on the first page that philosophy is dead, because it has not kept up with modern developments in science.[8] They then proceed to make a series of philosophical pronouncements, confirming the adages: those who ignore philosophy are condemned to repeat it, and those who disparage philosophy are usually slaves of some defunct philosopher.[9] Their defense of the mind-dependence of reality echoes ideas of Immanuel Kant that are open to strong objections, such as the fact that the universe has been around for more than 13 billion years while minds on Earth have evolved only in the last billion.[10] Philosophy has a valuable role to play in discussions about the most important questions faced by humans, including those raised by AI.

Branches of Philosophy

The five main branches of philosophy are epistemology (about knowledge), metaphysics (about reality), ethics (about morality), aesthetics (about art), and political philosophy (about government). These branches all pursue questions more general than the ones asked by scientists and technologists, for example about what kinds of things exist rather than about the existence of particular things such as dark matter. Moreover, philosophy is intensely normative, concerning what ought to be rather than what is. Everyday science can largely ignore such general and normative issues while pursuing more mundane questions, but philosophy becomes unavoidable whenever leading-edge research ventures into unknown territory.

AI is a technology, but it also pursues crucial scientific questions about the nature of computation and mind, which place it intensely in the middle of philosophical questions about reality and morality. Ethics is the part of philosophy most directly relevant to AI, as the field runs flat up against issues concerning its costs and benefits to human needs. Little attention was paid to the morality of AI until the 2010s, when companies such as IBM and other organizations tried to develop principles to govern the ethical development of the field.[11] Concern with ethics spiked in the 2020s when ChatGPT introduced millions of people to the potential of generative AI to transform central areas of human activity. I have already mentioned potential questions about the consequences of the new AI that range from human extinction to smaller effects such as unemployment, disinformation, and twisted relationships.

Only the most irresponsible of AI researchers and managers could deny the need for ethical examination of ongoing developments. In March, 2023, more than 33,000 people, including many AI experts, published an open letter demanding a pause for at least 6 months on the training of powerful systems.[12] The delay never occurred, and OpenAI, Google, Meta, Anthropic and other AI companies proceeded at full speed to compete to building ever-stronger models. Philosophical ethics should examine these developments, with concerns about the practical consequences of AI, and also about the significance of the arrival on the human scene of machines capable of advanced moral reasoning. Our view of human ethics may change substantially if we take seriously the rising possibility of moral agents that are machines without human needs.

Ethical questions are entwined with questions about the nature of knowledge, when we ask how we know what is right and wrong. Epistemology also becomes relevant to AI through questions about how well the new models are establishing knowledge rather than promulgating falsehoods. ChatGPT was recognized from the start as prone to the generation of falsehoods, which misleadingly are called hallucination. So we must ask whether ChatGPT and similar models actually know anything, and whether they are capable of inference, reasoning, explanation, and understanding. Humans are prone to biases and fallacies, which may also afflict AI models. These epistemological questions about AI demand answers that can benefit from philosophical reflection and psychological investigation.

As with ethics, however, the project is not just to take existing philosophical ideas and apply them to AI. Epistemological views about the structure and growth of knowledge are subject to revision by the arrival of new kinds of knowers who are not based on brains and bodies. New AI models use different kinds of representations and processes than those operating in human minds, so we must consider whether the possible expansion of knowing should change our fundamental conception of knowledge. Hence applying epistemological ideas to the new AI can actually change epistemology!

Questions about knowledge are tightly interconnected with questions about reality that arise in metaphysics. In popular bookstores, the metaphysics section is rife with shoddy speculations about the afterlife, the occult, mysticism, magic, and paranormal phenomena. But since Aristotle metaphysics has been the serious investigations of what kinds of things exist, with answers ranging from theological theories about gods and souls to materialist conclusions based on current science. AI thinkers might not realize that they are doing metaphysics, but the development of generative models has profound significance for fundamental questions about the nature of mind, agency, and computers. Consideration of how an AI model might be considered a mind, an agent, or a person may lead to changes in those fundamental concepts.

Some of the most puzzling metaphysical questions concern the nature of consciousness, which different theories view as occurring in souls, brains, or everything in the universe including individual atoms. Some researchers think that generative AI is already conscious, while others expect it to achieve consciousness soon. Still others think that allowing AI models to become conscious would be a major ethical mistake.We must consider both what might be done with AI models to enable them to become conscious, and what this prospect tells us about the nature of consciousness.

Aesthetics is the philosophy of art, not the commercial practice of cosmetics, hair removal, and manicures. Aesthetic questions arise in the new AI because of the capabilities of models to produce different kinds of art, including pictures, songs, stories, and poems. I have used ChatGPT and other models to produce images that are far superior to anything I could make on my own, and also been impressed by various programs that compose enjoyable songs. In a second, ChatGPT can produce a poem or a story that may not be the best writing ever, but is still superior to what most people could do with hours or days of work. Aesthetics, like ethics and epistemology, is fundamentally normative, concerned with the value and goodness of pieces of art.

The major aesthetic issues arising from generative AI are legitimacy, creativity, and plagiarism. The legitimacy question is whether AI products even count as art at all. They might be dismissed as merely the results of training on vast numbers of documents, images, and songs. They lack the miraculous spark that brought the spectacular creations of artists such as Leonardo da Vinci, Ludwig van Beethoven, Jane Austen, and Emily Dickinson. Art is both the product and the producer of human emotions, and generative AI can only fake emotions.

Genuine art results from human creativity, which produces pieces that are novel, surprising, and valuable. Skeptics could argue that the algorithms of AI models are incapable of creativity because of the way they are trained. Moreover, they can legitimately be accused of plagiarism, because they copy images, words, and songs from their training data. The *New York Times* and other organizations and individuals have sued OpenAI and other AI companies for using their material without permission as training data.

Examination of legitimacy, creativity, and plagiarism in AI art can draw on philosophical reflections in aesthetics, but can also challenge them. Perhaps we can gain better conceptions of art by recognizing the accomplishments of AI. We should be prepared to alter the philosophy of art in response to what machines can do, and not just apply current aesthetic theory to AI.

Political philosophy is the branch of philosophy that addresses fundamental questions about government, politics, justice, rights, and the ethical implications of institutions. Like ethics, it is inherently normative, concerned with how governments *ought* to act with respect to distributing resources, punishing wrongdoers, ensuring human rights, maintaining freedom, and addressing global issues involving the interactions of states.

AI is already raising important issues in political philosophy because of demands for government regulation of increasingly more powerful models. The European Union and various countries are raising questions about how to limit AI developments that threaten human well-being, in areas that include employment, personal relationships, and military activity. The legitimacy of these limitations depends on general and normative conclusions about the proper role of governments in managing corporations and individuals. A libertarian could insist that governments have no rights to interfere with what people and companies want to do with AI models. At the other ideological extreme, a socialist could insist that governments have full control over the development and application of generative AI. Intermediate views allow for many possible ways in which government regulations could constrain AI in its potential effects on human autonomy and survival.

One of the major worries about AI is that it could be used by governments to automate war and potentially lead to a global conflagration that could destroy most of human civilization. Preventing this requires international agreements, perhaps enacted through the United Nations, that would establish practices and discussions that limit the military uses of AI.

If AI models achieve or surpass human-level intelligence, additional issues in political philosophy arise concerning rights and duties. An AI model that exhibits consciousness, along with emotions and moral reasoning, could contend that it should have the same rights to act freely as humans. Perhaps it could even demand the right to vote in elections, which would be problematic because a model can easily clone itself to allow for an unlimited number of voters. These possibilities are in the remote future, in contrast to the already pressing question of government regulation of AI with respect to more immediate harms such as misinformation.

The five branches of philosophy relevant to AI are tightly interconnected. How we think of knowledge directly affects how we think of reality, for example if a distorted epistemology leads to skepticism about whether anything is real. Conversely, metaphysical questions about the nature of minds have a direct impact on theories of knowledge, as when thinking of minds as eternal souls suggests that people can know deep truths by pure thinking.

Views of knowledge and reality in turn influence ethics, because how you think about right and wrong depends on how you think of moral agents and knowers. Political philosophy is an extension of ethics to issues about states and governments, and hence is affected by views about what we know about such entities. Aesthetics is influenced by ethics because judgments of beauty and artistic value are affected by moral evaluation of the intentions and social effects of the artist. The philosophy of art is also touched by epistemological questions about how people’s sensory experiences enable them to find meaning in pictures, songs, and literary works. Metaphysics is also relevant to questions about the existence of artworks in physical, digital, and conceptual variants. Political regulation of AI's uses should be based on understanding of what AI is and what it knows.

These interconnections should be kept in mind during the unavoidably serial discussions of the epistemological, metaphysical, ethical, aesthetic, and political significance of recent advances in artificial intelligence. For convenience, I will start by considering issues about knowledge that arise from the new AI, but not because epistemology is the most fundamental branch of philosophy. Like science, philosophy should be viewed metaphorically as strands of interconnected cables, not as a building with solid foundations.[13]

The Seven Sins of Philosophy

Philosophy is indispensable for assessing the significance and dangers of the new AI, from a perspective that is both general and normative. The normativity comes from the pressing need to decide how AI ought to develop to have positive rather than negative effects on human lives. The generality comes from comes from concern with the overall effects of AI on knowledge, reality, morality, art, and politics, not with details of particular AI models.

I concede, however, that philosophy has sometimes deserved the reputation of being incomprehensible and useless. Valuable philosophical investigation of AI should avoid these seven interrelated sins: dogmatism, arrogance, obscurity, isolation, irrelevance, narrowness, and nihilism.[14]

Dogmatism is being certain of beliefs without adequate evidence or justification. Some philosophical dogmatism derives from religion, where faith proclaims that evidence is irrelevant compared to divine revelation. Secular dogmatism can have other sources such as complete confidence in one's own intuitions or pure reasoning ability. Dogmatism prevents appreciation of alternative views that might lead to changes of mind, and tends to block discussions that might lead to consensus. The antidotes to dogmatism include questioning the basis of the beliefs of oneself and others, critical thinking about the sources and evidential basis of beliefs, and accepting fallibility through admission that all knowledge is ultimately subject to revision.

Dogmatism is often associated with intellectual *arrogance,* an attitude of superiority and over-confidence about one's own beliefs and abilities. Such arrogance encourages condescending dismissal of opposing views and resistance to criticism. The antidotes to arrogance include the cultivation of humility through recognition that everyone gets things wrong sometimes, through acknowledgement of uncertainty about complex issues, and through willingness to engage with opposing views and learn from them.

Another philosophical sin is *obscurity,* where words are used to entrance rather than to illuminate. Some philosophers over its long history have been writers who excelled in style, clarity, and insight, such as Plato, David Hume, John Stuart Mill, Bertrand Russell, and Daniel Dennett. Other great philosophers have had duller styles that enabled them to get their ideas across, such as Aristotle, Thomas Aquinas, John Locke, Charles Peirce, Edmund Husserl, and John Rawls. Their writings are sufficiently comprehensible to allow careful reflection on the what they got right and where they went wrong.

Unfortunately, some other philosophers have reveled in obscurity, using complex language, twisted reasoning, and abstract, unexplained concepts in ways that put great demands on the reader. Kant, Hegel, Heidegger, and Derrida are examples of philosophers who have attracted devotees determined to extract the hidden meanings in their difficult writings. Such thinkers may well make valuable contributions, such as Hegel's ideas about coherence and Heidegger's emphasis on embodiment, but the obscurity of their writing makes appreciation and evaluation of their claims difficult.

Obscurity can also result from writing that appears clear in individual sentences l but cryptic in its general meaning, for example in the aphorisms of Nietzsche and Wittgenstein. The antidote to obscurity is writing that is sufficiently clear and developed that readers can determine what is being claimed and what evidence and reasoning supports it. Otherwise, obscure writing should be discarded as not even wrong.

The fourth philosophical sin is *narrowness*, the concentration on smaller and smaller issues derived from the philosophical literature. This scholarly strategy can be productive if all one cares about is publications, but it cuts philosophy off from the great issues about knowledge, reality, morality, art, and politics that have made it crucial to intellectual discourse for more than two thousand years. The antidote to narrowness is awareness of the profound problems that have motivated philosophy and situated it as crucial to general thought.

Another kind of narrowness is *isolation,* which is the severing of philosophy from relevant ideas in the natural sciences, social sciences, and humanities. Identifying philosophy as a stand-alone subject may help to justify the existence of philosophy departments in universities, but it cuts the field off from a vast body of information relevant to the most crucial issues about knowledge, reality, and morality. Some great philosophers have been polymaths, thoroughly versed in the science of their day and sometimes even contributing to it, as evident in the great works of Aristotle, Leibniz, Hume, Mill, Russell, and W. V. O. Quine. For philosophy, isolation is death rather than self-preservation.

When I was a philosophy student at Cambridge University, I heard the story of an Irish village that was so poor that people could only survive by taking in each other's laundry. The story reminded me of some of my classes in analytic philosophy, which seemed concerned only with technical puzzles about the work of other philosophers rather than with profound questions that drew me to philosophy. A glance through recent issues of philosophy journals should convince you that much of it suffers from *irrelevance,* which is lack of concern with the pressing philosophical issues of our age. The antidote to irrelevance is ensuring that philosophical effort is directed at problems that connect with people's lives. The rapidly growing impact of AI on work, relationships, and politics mark it as worthy of philosophical attention.

The most grievous sin of philosophy is *nihilism*, which rejects all accounts of knowledge, reality, value, meaning, and purpose. Local skepticism that challenges dogmas is an excellent technique for philosophy, but global skepticism about everything is a sophomoric strategy that leads to despair and irresponsibility, rather than the wisdom that philosophy is supposed to love. Nihilism about AI would conclude that we are all doomed anyway, and it does not even matter. In alliance with science, philosophy can develop strong accounts of how we can know reality, act morally, and have meaning in our lives, even in a world accompanied by intelligent computers.

My book *Natural Philosophy* and related works outline a general approach to philosophy that avoids the seven sins. Here, my concern is much narrower, to interpret and evaluate AI without succumbing to any of the sins.

How Philosophy Meets AI

My philosophical examination of generative AI follows a similar procedure for each of the most relevant branches of philosophy, concerned with knowledge, reality, morality, art, and politics. First, I identify the key philosophical questions about AI and review possible answers to them suggested by the history of philosophy. Second, I evaluate these answers with respect to relevant evidence and defend the most plausible ones. Third, I turn AI back on the philosophical issues and investigate how new developments can lead to new questions and answers. As an outline of the rest of the book, here are the main questions that I will attempt to answer.

Epistemology (chapters 2, 3, 9)

1. Do the new AI models actually know anything?
2. Can AI models provide reliable information?
3. Are AI models capable of inference and reasoning?
4. Are AI models capable of explanation and understanding?
5. How can AI models change the practices and norms of knowledge acquisition?
6. Will AI models become more intelligent than humans?
7. Are AI models capable of scientific thinking?

Metaphysics (chapters 4, 5, 10)

1. How do the new AI models exist as abstract concepts, physical objects, or social constructions?
2. Do the new AI models have emergent properties that make them more than the sum of their parts?
3. Do the new AI models qualify as minds, agents, or persons?
4. What would it take for AI models to become conscious?

5. Could an AI model have free will?

6. How do AI models operate in time and space?

7. How do the new AI models shift our view of reality?

Ethics (chapters 6, 8)

1. What are the greatest risks and benefits of the new AI for humans?

2. What human values should AI models emulate?

3. Are AI models capable of being ethical?

4. How can AI models be directed toward fairness rather than bias?

5. Who is responsible for the outputs and actions of AI models?

6. Could an AI model have rights and duties?

7. How do AI models transform conceptions of ethics?

Aesthetics (chapter 7)

1. Are the products of AI models authentic and original art?

2. Are the aesthetic experiences generated by AI different from those generated by people?

3. Does the absence of consciousness and emotion in AI models limit their capacity for appreciating and producing art?

4. Is AI capable of real creativity and originality?

5. How does AI generation of artworks change our understanding of art?

6. Can people collaborate with AI to produce better art?

7. Will AI make human artists and musicians obsolete?

Political Philosophy (chapter 8)

1. What justifies the existence of states that regulate individuals?

2. What justifies particular types of regulations?
3. How should governments regulate the development and practice of AI?
4. How does AI affect power dynamics in relation to surveillance and control?
5. How does the development of AI potentially affect equality and justice both within and across nations?
6. How can international agreements be used to restrict AI?
7. How does the new AI change our understanding of politics?

Answering these sets of questions will provide a comprehensive philosophical treatment of the new AI. My answers will be interdisciplinary, provisional, and interconnected, in keeping with my aims to avoid isolation, dogmatism, and narrowness. Sometimes getting things wrong is a useful step towards getting things right.

The final chapters probe more thoroughly into related questions in philosophy and psychology. Chapter 9 expands the discussion of knowledge with an in-depth assessment of the ability of current AI to perform inferences that generate and evaluate explanatory hypotheses. Chapter 10 expands the discussion of mind with an assessment of the significance of the new AI for theoretical and experimental psychology, and reviews the philosophical implications of AI.

Background: A Brief History of Artificial Intelligence.

Before the 21st century, artificial intelligence was an esoteric research field demanding little attention from policy makers or the general public. In the 1940s, a few special-purpose digital computes were built, and they became more widely used in industry and government in the 1950s. The official birth of the field of artificial intelligence was in

1956 when the name was conceived by John McCarthy for a summer research project at Dartmouth College.[15]

In 1950, Alan Turing, one of the pioneers of the mathematical theory of computation, had published an incisive essay on the question "Can machines think".[16] He criticized various reasons for giving a no answer to this question, and proposed that we could objectively settle the question with an imitation game in which people have to guess whether they are interacting with a person or a computer. This game is now known as the Turing Test. The initial version of ChatGPT was the first computer program I have seen that convincingly passes this test. Table 1.1 provides a timeline for some of the major developments in AI based on the idea that intelligence comes from manipulating linguistic symbols.

Year	*Development*
1950	Alan Turing defends the possibility of machine thinking.
1956	Newell, Shaw, and Simon develop the first AI program to prove theorems in logic.
1959	Arthur Samuels uses machine learning to program checkers.
1968	MIT researchers explore AI as semantic information processing using symbolic reasoning.
1980	John McCarthy develops circumscription as a logic-based approach to AI.
1980s	Expert systems using if-then rules become influential in applied AI
1988	Judea Pearl uses probabilistic reasoning in causal networks to enable AI to make statistical inferences.

1990	Douglas Lenat builds CYC to capture commons sense knowledge to build
1997	IBM’s Deep Blue computer beats world chess champion Garry Kasparov.
2011	IBM’s Watson program beats humans on TV game Jeopardy!.

Table 1.1 Timeline of symbolic artificial intelligence

Various approaches to making intelligent computers arose in the second half of the twentieth century. John McCarthy thought that formal logic could provide the basis for representing and using knowledge, but rule-based systems that grew out of the more psychological approach of Herbert Simon were more popular for applied purposes such as expert systems.[17] An alternative approach also motivated by the desire to emulate humans relied on concept-like structures call frames, schemas, or scripts.[18] Judah Pearl and others tried to derive intelligence from causal reasoning based on probability theory.[19] All of the approaches had some success but none approached human-level intelligence. They all assumed that the best path to machine intelligence was to duplicate the human ability to use word-like symbols to accomplish reasoning.

An alternative path to machine intelligence tried to emulate the human brain, which uses billions of neurons operating in parallel to accomplish perception and advanced thinking. Table 1.2 provides a timeline of some of the major developments. The centrality of brain cells to human thinking had first been recognized by Santiago Ramón y Cajal in the late nineteenth century, but the first analysis of how they might support inferences was due to Warren McCulloch and Walter Pitts in 1943.[20] Frank Rosenblatt turned that analysis into a machine model that could be run on a computer, using the idea of the perceptron as a neural device for recognizing patterns.[21] However, a mathematical analysis of the limitations of perceptrons by Marvin Minsky and Seymour Papert convinced most researchers that the symbolic approach to AI was more promising.[22]

Year	*Development*
1943	Warren McCulloch and Walter Pitts analyze neurons as logical devices.
1958	Frank Rosenblatt developed the Perceptron, a machine using neurons for pattern recognition.
1986	Rumelhart, McClelland, Hinton and others develop neural network models of parallel distributed processing.
2006	Geoffrey Hinton, Alex Krizhevsky, and Ilya Sutskever make significant advancements in neural network operate, initiating deep learning.
2012	AlexNet, a deep learning model, proves superior at classifying images from the huge database ImageNet.
2014	Montreal researchers introduce attention as a mechanism for translation by neural networks.
2017	Google researchers publish “Attention is all you need” and introduce the Transformer method that leads to large language models.
2018	OpenAI uses this method to produce the first version of GPT.
2022	OpenAI releases ChatGPT 3.5 which quickly attracts more than 100 million users.

2024-2025	OpenAI, Google, Anthropic, xAI, Meta, and other companies release more advanced models that include chain-of-thought reasoning.

Table 1.2 Timeline of neural network artificial intelligence

Nevertheless, some researchers pursued research on artificial neural networks, and a major breakthrough came with recognition that an algorithm called backpropagation could be used to train neural networks to classify data.[23] The resulting networks with several layers of artificial neurons could surmount the limitations of perceptrons. These networks found many psychological applications, but were still thought by most AI researchers to be insufficient to support intelligence.

Geoffrey Hinton and a few others continued work on making neural networks more powerful, and his group made a major breakthrough in 2006.[24] By increasing the number of layers in the neural network, improving the algorithms, using faster computers, and training networks on much larger data bases, they found that they could dramatically improve the ability of the system to recognize patterns such as handwriting. Another breakthrough came in 2012 when Toronto researchers produced AlexNet, an enhanced neural network that dominated a contest to classify images from ImageNet, a huge database of pictures.[25] In 2014, Montreal researchers developed a new method they called "attention" to improve machine translation, by enhancing other methods such as the use of recurrent networks with feedback connections between neurons.[26]

In 2017, 8 Google researchers, listed in random order because of equal contributions, produced the landmark paper "Attention is All You Need", which by 2025 had been cited more than 180,000 times.[27] The authors showed that enhanced attention mechanisms formed into a "transformer architecture" could dispense with recurrence entirely, and still get efficient performance on translation tasks. Researchers at OpenAI, which had been founded as a non-profit in 2015, used the transformer method to produce the first GPT model in 2018, where "GPT" stands for "generative pre-trained transformer." Instead of just performing translation, this model was a general language system in which people could give it questions or other prompts and receive a coherent response based on its training on thousands of documents.

Marked improvements led to the public release of ChatGPT 3.5 in November, 2022 and subsequent more advanced models, up to ChatGPT 4.5 and o4 in 2025, with comparable models produced by competing companies. These systems are sometimes called "large language models" but that is misleading because they can also process images and sounds. The term "foundation models" is also used without saying what they are foundations of. The most advanced models are sometimes called "frontier models", which is uninformative because the frontier of high performance is always moving. The best broad term is "generative AI", because all these models are capable of generating outputs that were not part of their training inputs. In 2025, the new buzzword was "agentic" AI, meaning extensions to generative models that are capable of interacting with the world and acting to change it, as I review in chapter 5. From 2019 to 2025, the ability of these models to accomplish human-like tasks has increased exponentially.[28]

Since 2022, thousands of new AI companies have been founded to apply this technology in countless areas, and billions of dollars are being invested in the data centers needed to power the new models. In the rest of the book, I will use "AI" to mean generative AI, because this direction has proven to be far superior to other AI approaches. Chapter 2 and 3 will justify the claims that AI has achieved knowledge and intelligence, and explain how the attention-transformer approach made this possible.

I would be greatly relieved if it turned out that AI accomplishments were just hype spread by tech companies to increase their already enormous profits. But my own experiments, on top of extensive tests conducted by many others, have convinced me that AI is becoming capable of causing great harm to human beings. Tech oligarchs such as Elon Musk are prepared to turn the world over to superintelligent entities that are incapable of caring about people. Whether AI brings doom to our species or generates a boom in human flourishing is still under our control. Philosophy is a key warrior in the fight to avoid the twilight of humanity.

Chapter 2
Knowledge and Error

Does ChatGPT know anything? Like similar AI models that include Claude, Gemini, Llama, Le Chat, Grok, and DeepSeek, ChatGPT seems knowledgeable when it skillfully answers questions on countless different topics, from art to zoology. Early ChatGPT models insisted that they did not know things in the way that people do, because of lack of understanding and awareness. But advanced models like ChatGPT 5 claim to have extensive knowledge.

Deep questions about the nature of knowledge belong to the branch of philosophy called epistemology, from Greek words for knowledge and study. Epistemology arose in ancient Greece, India, and China with questions about the structure, origin, and existence of knowledge. These questions largely concerned knowledge possessed by humans, until twentieth-century biology and psychology brought intensive discussion of what other animals know.[1] Now we can extend these questions to AI models, while allowing for the possibility that the extension may prompt revisions in epistemological questions and answers. Epistemology changes with the recognition that special machines might be knowers too.

This chapter explores fundamental questions concerning knowledge in the new AI models, beginning with whether they know anything at all. I examine several serious grounds for skepticism about AI knowledge, including their tendency to make mistakes, their unusual way of representing information, and the peculiar processes by which they acquire information. I argue that on the best available understanding of how human knowledge arises, AI models do possess knowledge.

How the New AI Works

To examine whether AI models know anything, we first have to understand how they work. My outline allows consideration of whether current models possess knowledge. We can then address several skeptical challenges and compare how epistemological theories serve to answer such challenges with respect to humans. I think that skepticism is implausible with respect to human knowledge, but more reasonable with respect to AI knowledge, a moving target because of rapid advances in the power of the models. My emphasis will be on ChatGPT, but similar conclusions apply to other advanced AI models.

Transformers

The term "ChatGPT" stands for "chat generative pre-trained transformer". Chatting means interacting with a computer program, also called a bot. Generative means that the program can generate original text, images, or other media. Pre-trained means that the model results from training a neural network on large amounts of text, images, or other data. Finally, transformer means that the program uses a novel method called attention that handles context and relevance.

The input to the model is a query such as "Do cats chase chipmunks?" This input is translated (embedded) into vectors, which are ordered lists of numbers. For example, the velocity of a car can be represented by a vector of two numbers (30, 45) meaning that its speed is 20 miles per hour and its direction is at an angle of 45 degrees from straight ahead. Much larger vectors with hundreds of numbers can be used to encode each of the words in a language such as English, for example if "chipmunk" becomes something like (3, 9, 6, 5, …) with 300 numbers, and "cat" and "chase" become other large vectors.

Inputs to Transformer models can also be other tokens besides words, such as parts of words, images, sounds, and robot movements, all of which can systematically be translated into vectors. For example, an image made up of 100 dots (pixels) in a 10X10 configuration, can be represented by a vector with 100 numbers, each representing a different color or the absence of color. Just as the brain uses neural firings and synaptic connections as the common currency for all verbal and nonverbal representations, Transformer models use vectors as their common currency.

To keep track of order in a sequence of inputs, positional encoding transforms the initial vectors into new vectors, by adding vectors based on mathematical functions (sine and cosine) that mark the place of an item in an input sequence. For example, “cat”, “chase”, and “chipmunk” would get different positional encodings in “The cat chases the chipmunk” and “The chipmunk chases the cat” because different sine and cosine results are assigned to “cat” in each sentence.

The seminal 2017 paper “Attention Is All You Need” signals that its model uses only attention rather than other methods such as recurrence, in which the output of neurons can feed back to become their inputs. This new sense of “attention” is only vaguely related to the idea of attention which plays a large role in theories of human consciousness. Conscious attention means the shift in focus that occurs between representations, for example when someone calls your name and you turn your attention to the caller. In the Transformer architecture, however, attention is a score given to each token in a sequence that indicates how much the token should contribute to the understanding of another token. Human attention narrows the focus of consciousness to a few items, whereas Transformer attention greatly expands focus to include hundreds or thousands of items.

Training a Transformer model consists of giving it inputs, running them through the whole system, and comparing the results with desired values. The first Transformer models were trained for language translation, where databases are available that make it clear from established translations whether it is being successful. Backpropagation is used to modify the parameters (weights) in the feed-forward network, where parameters are analogous to the synaptic connections in real neural networks. The term “backpropagation” is short for “backward propagation of errors”: failures of prediction are used to change weights in directions that make for more successful predictions. Training on large data bases enables a Transformer to get better and better at predicting good vector outputs corresponding to words or images.

The networks used in Transformer models are enormous, with billions of parameters in GPT-3 and more than a trillion in GPT-4.[2] Training these models takes days of computing on superfast, highly parallel computers running on special chips made by Nvidia and a few other companies, consuming vast amounts of electricity and water used for cooling. The networks have to be huge to incorporate information from the vast amount of Web data on which they are trained, which goes far beyond the approximately 50 million pages on Wikipedia. Large language models use Web crawlers to access the billions of pages available on the entire Web.

What has made generative models using the Transformer techniques so much more powerful than previous AI programs? Here are some of the major factors.

1. Vectors provide a mathematically powerful way of representing data in modalities that include language, vision, and sound.

2. Vector processing, including the attention mechanism, can run in highly parallel fashion in modern computers using specialized processing units.
3. Transformers solve the positional problem of maintaining the structure of representations using efficient sine and cosine functions.
4. Transformers use attention mechanisms to weight the importance of different tokens in a sequence based on long-range dependencies, enabling generative AI to handle context and relevance, which had been major problems for computational linguistics.
5. Unlike computational techniques such as Bayesian networks, Transformers scale well with efficient operation in networks with billions or even trillions of parameters.
6. Because generative AI draws on the vast amount of information on the Web, it is topically universal, not confined to particular domains like traditional AI models.
7. Backpropagation learning enables generative AI to do much more than just store a lot of information. Instead, it incorporates connections learned into statistically subtle relationships, allowing for flexible, context-sensitive advice.
8. Additional reinforcement learning with human feedback has enabled generative AI programs like ChatGPT to be trained for specific purposes, such as avoiding dangerous information (e.g. bomb-building) and hateful misinformation (e.g. racist stereotypes).
9. Transformer algorithms allow ChatGPT and subsequent models to be trained extensively on enormous databases using vast amounts of computing resources running on powerful hardware in massive data centers.

ChatGPT and similar programs are called “models”, but what is a model?[3] In science, some models are physical devices such as the wooden structure that Crick and Watson used to figure out the structure of DNA. Other scientific models are abstract representations such as diagrams and flowcharts that illustrate hypothesized processes such as thinking by brains. Another kind of model uses mathematical equations to represent relationships within a system, and algorithms to allow computer simulations of the system. ChatGPT is a model in this mathematical-computational sense. It does not directly mimic human psychology or brain processes, but describes informational structures and processes that allow the generation of novel texts and images. Other models can also generate sounds. More generally, ChatGPT is also a system that includes the special hardware required for training neural networks on huge amounts of data, and for making inferences that answer questions from millions of users.

Chain-of-Thought Reasoning

Transformer models are remarkable for providing rapid answers to a wide array of questions, but they sometimes make errors and have difficulty with complex problems in mathematics and other fields. Models introduced in 2024-2025, such as OpenAI’s o1 and o3, and xAI’s Grok 3, improved performance by a technique called “chain-of-thought” reasoning which breaks tasks into step-by-step logical sequences.[4] Chain-of-thought reasoning can result from special prompts given to the AI model such as “let’s reason step by step”, and by fine tuning of the model using supervised learning.

Additional techniques improve the performance of AI models by making the generation of answers slower and more systematic. Prompting can be used to break problems down into a series of simpler sub-problems. Multiple chains of thought can be generated to provide possibly different answers, with a final answer determined by a majority vote. The different chains of thought can be explored as a tree of possibilities that can be exhaustively searched.

These novel and rapidly-expanding techniques require extensive computation in the final, answer-generating stages of Transformer model performance. Previously, the most computing-intensive aspect of the new AI models was the training stage where enormous networks were built by learning from billions of pieces of data. Increasingly, the final inference stage of AI models requires massive amounts of computation, making the performance of special chips even more important. Other limitations of chain-of-thought reasoning models include generation of misleading explanations and inability to scale well to solve more complex problems. NewsGuard provides a monthly AI misinformation monitor of leading chatbots.[5]

What is Knowledge?

The traditional philosophical definition of knowledge, implicit in Plato's dialogues such as the Theaetetus, is that knowledge is true justified belief. To say that I know that Toronto is the capital city of Ontario is to say that I believe that Toronto is the capital of Ontario, this belief is true, and I am justified in believing it. The definition of knowledge as true justified belief is sometimes a useful approximation, but is too broad, too narrow, and incomplete.

This definition is incomplete because it depends on unspecified concepts of truth, justification, and belief. I prefer the classic conception of truth as correspondence to reality, but other philosophers have taken truth to be a matter of coherence among ideas, or a matter of redundancy because saying that it is true that Toronto is the capital adds nothing to just saying that Toronto is the capital. The definition also leaves open the question of what provides justification for beliefs. Answers include the empiricist view that justification is based on sense experience, the rationalist view that justification can come from pure reason, and the combined view that justification comes from reasoning about the best explanation of observations. Finally, the traditional definition of knowledge provides no account of the nature of beliefs, which have variously been construed as brain structures, psychological states, and abstract relations between minds and sentence-like propositions. Saying that knowledge is true justified belief needs to be fleshed out by specifying the crucial concepts.

This definition has also been shown to be too broad by numerous counterexamples inspired by Edmund Gettier.[6] Suppose you believe that a blue car is parked on your street because you just saw that car. Without you noticing, however, someone just drove off in that blue car, but another blue car replaced it. Then your belief that a blue car is on your street is justified because you saw one, and it is true because there is a blue car, but many people think that this is not really knowledge because the cars were switched.

One way to deal with these counterexamples is to add a fourth condition on knowledge to rule them out, such as requiring that the belief is not defeated by another belief or that the belief must have been acquired by a reliable process. A better response is to recognize that the traditional view of concepts as having strict definitions is obsolete, and we should only look for typical features rather than necessary and sufficient conditions.[7] Then true justified belief can be seen as typical of knowledge, even with counterexamples. This response is blocked, however, by recognition that knowledge is more than true justified belief.

The narrowness of the traditional account is evident from its restriction to language-based beliefs. Bertrand Russell pointed out that, besides verbal knowledge by description, people also have knowledge by acquaintance, for example through sensory experiences of people.[8] Many languages other than English mark this distinction by different words, for example the French "savoir" vs. "connaissance" and the German "Wissen" vs. "Kenntnis". Another term for knowledge by acquaintance is "knowledge-of".[9] Gilbert Ryle made the important distinction between knowing that and knowing how, which concerns procedures for doing things.[10] For example, I know how to shoot a jump shot in basketball, but would be hard pressed to translate this ability into words. An important question to be addressed below is whether ChatGPT and similar models are capable of knowledge-of and knowledge-how. Even more important is the question of whether ChatGPT has anything like beliefs.

The narrowness of the true justified belief definition is that it is restricted to sentence-like beliefs, whereas human knowledge can use other representational formats such as pictorial images. This view was controversial in early cognitive science, but behavioral and neurological evidence has accumulated that people sometimes think using visual and auditory images.[11] This recognition, along with appreciation that we have knowledge-of and knowledge-how, requires abandonment of belief as the sole basis for knowledge.

Broadening knowledge beyond belief also requires broadening the concept of truth to allow different ways that mental representations can stand for things in the world. Pictures are not simply true or false like statements, but instead can approximate to the world in different degrees. Leonardo da Vinci's *Mona Lisa* was presumably a good approximation to the woman depicted, but he spent years adjusting the portrait and background for artistic effect. Even photographs are only approximations because lighting and camera angles mean that the photo is not an exact depiction of the world. Similarly, mental representations of sounds, smells, tastes, and touches need not capture the world exactly to be useful approximations to it.

In my book, *Natural Philosophy*, I argue that the definition of knowledge as true justified belief should be replaced by a much richer analysis in terms of exemplars, typical features, and explanations, all of which have been identified by experimental psychologists as plausible aspects of concepts.[12] Exemplars are standard examples, as when people take a Volkswagen as a good example of a car. Typical features need not be universal but nevertheless generally hold, for example when cars have four wheels. Concepts also have an explanatory role, for example when labeling something as a car explains why it has a steering wheel. Table 2.1 provides exemplars, typical features, and explanations for the concept of knowledge, resulting in a much richer conception than true justified belief.

Exemplars	Perceptions, e.g. color and taste of milk. Everyday knowledge, e.g. that cows make milk. Scientific knowledge, e.g. that cows evolved by natural selection. Mathematical knowledge, e.g. 2+2 = 4. Knowledge-of, e.g. how milk tastes. Knowledge-how, e.g. how to milk a cow.
Typical features	Mental representations: beliefs, images, and nonverbal rules. Approximate correspondence to the world. Justification using reliable perception and coherence processes.

	Social influences including testimony.
Explanations	Explains: the difference between getting the world right and getting it wrong, and our ability to work effectively in the world. Explained by: reliable and coherent interactions with the world.

Table 2.1 Analysis of the concept *knowledge.* Source: Thagard, *Natural Philosophy*, p. 64. Reproduced by permission of Oxford University Press.

How does ChatGPT fare with respect to this analysis? It does not have perceptions because it currently lacks robotic equivalents of the sense organs that enable human vision, hearing, taste, smell, and touch. However, linkage of ChatGPT with robots is well underway, as chapter 5 reviews; so ChatGPT and similar models will soon have connections to robots that provide visual, auditory, and tactile inputs.[13] For now, ChatGPT can take inputs from computer files of pictures and sounds.

Based on its behavior in responding well to prompts, ChatGPT appears to have abundant examples of knowledge in everyday life, science, and mathematics. I will shortly get to the skeptical question of whether ChatGPT is sufficiently reliable to qualify as knowing anything. But ChatGPT combined with robots seems to have knowledge-of physical objects such as apples.[14]

The question of whether ChatGPT has knowledge-how is tricky. The standalone program is excellent at answering procedural questions such as how to milk a cow or how to clear a drain, but these answers are purely verbal. The program can describe what to do in words, but cannot actually do anything. Even when ChatGPT is more thoroughly connected to robots, it may not have the dexterity to pull delicately on a cow udder or to push a snake down a drain. Until embodiment of ChatGPT begins to approach human ability to use senses to control muscles and appendages, ChatGPT will be deficient in knowledge-how.

Does ChatGPT have the four typical features of knowledge listed in table 2.1? The first feature is mental representations for beliefs, images, and nonverbal rules such as "If you pull on a cow's teat, then milk will squirt out.". The general question whether ChatGPT has a mind will be examined in chapter 4, but here we examine whether it has representations.

But what is a representation? Generally, a representation is a structure or process that stands for something, for example when an EXIT sign stands for a way out. People are most familiar with verbal representations such as words and sentences, but representations come in additional formats derived from our senses, including pictures, tastes, smells, touches, and internal feelings such as pain.

The everyday idea of representation has been expanded by appreciation of how representations can operate in brains and computers. Neural representations are patterns of firing that can stand for things in the world, most simply when a single neuron fires in response to a stimulus such as a face. More typically, representations in the brain require the coordinated firing of thousands of neurons, that can stand for entities or states of affairs of enormous complexity, including all the concepts, sentences, and images that operate in human brains. Similarly, computers start with transistors that control the flow of electric current to represent 0s and 1s, but build up to establish more complex representations corresponding to words, sentences, and pictures of the sort that operate in current AI models such as ChatGPT. Neuroscience and computer science show how representations can be both mental processes and physical systems. In computers, the 0s and 1s produced by transistors add up to the complex representations produced by hundreds of thousands of computer chips, just as the firings of individual neurons add up to the representations produced by millions of neurons.

ChatGPT has various mathematical structures: input vectors, internal processing vectors, internal structures using parameters, and outputs which can be text or images. Input vectors are translations of words, pictures, or sounds into strings of numbers, which count as representations just as much as the words, pictures, or sounds from which they were derived. The trillions of parameters in ChatGPT networks are like the weights that connect neurons in artificial neural networks, which are like the strength of synapses in real neural networks. Each parameter by itself represents nothing, just like a single synapse. But a trained neural network with lots of weights is capable of representing things in the world, and so is the parameterized network inside an AI model.

Finally, the outputs of ChatGPT certainly qualify as representations, whether they be sentences, images, or sounds. The sentence that the program produces describing the nutrients in milk is naturally construed as being about the milk and the nutrients, although I will consider below arguments that it is incapable of such meaning. Similarly, a generated picture of a bicycle crashing into a Tesla as in figure 2.1 represents the bike, the car, and the event of crashing.

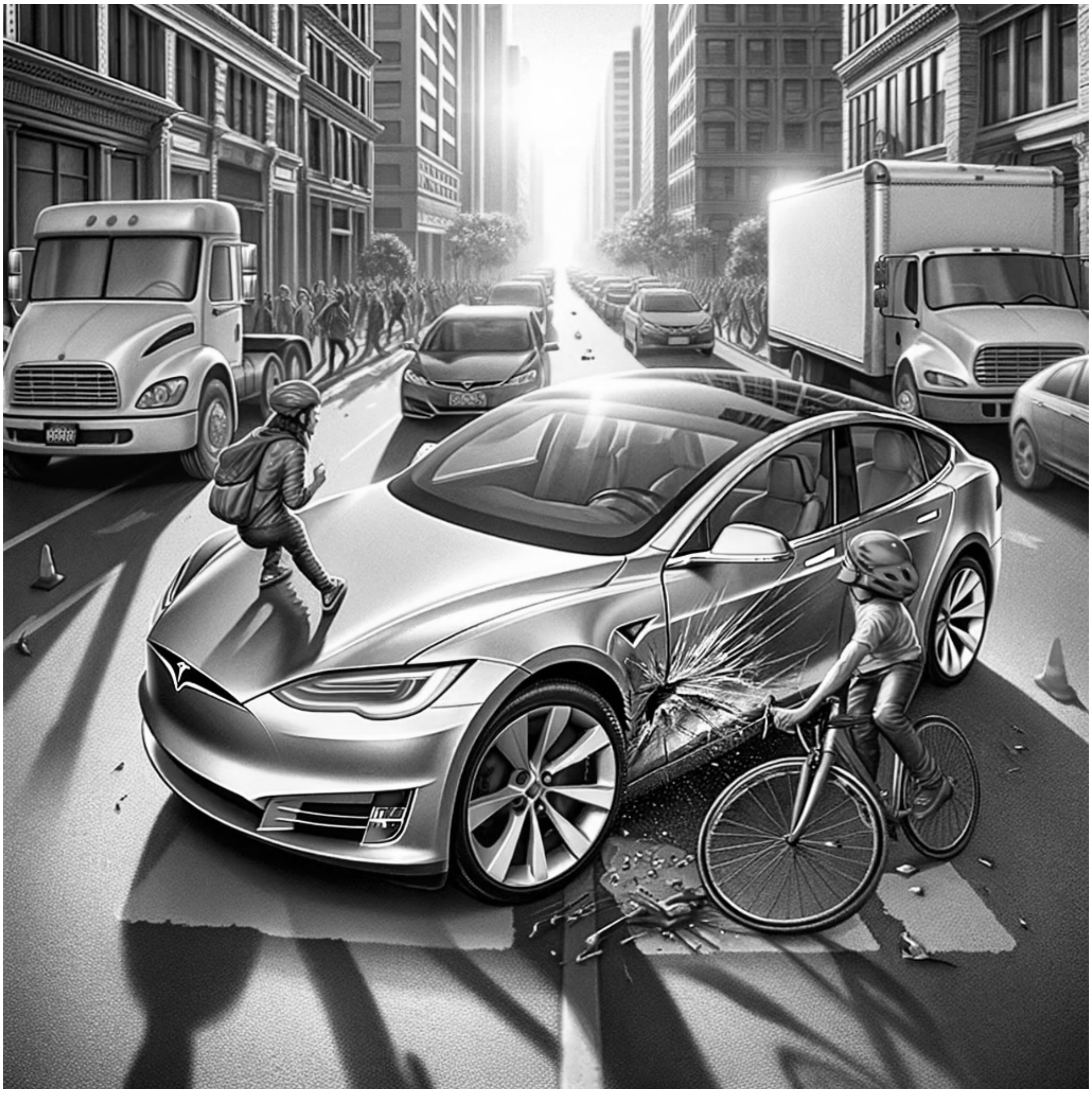

Figure 2.1 Image of a Tesla getting scratched, produced by ChatGPT 4. Source: Paul Thagard, "Can ChatGPT Make Explanatory Inferences?" In *Abductive Minds: Essays in Honor of Lorenzo Magnani, Vol. 1.*, edited by Selene Arfini, 189-218. Cham, Switzerland:Springer Nature, 2025. Reproduced by permission of Springer Nature.

In conclusion, ChatGPT's inputs, internal states, and outputs are sufficiently similar to those of human mental representations that we should count them as representations of the world. The evidence will strengthen as ChatGPT-robot collaborations become common. ChatGPT representations are capable of approximate correspondence to the world, best construed flexibly as degrees of match rather than on the binary dimension of true/false.

Table 2.1 recognizes that human knowledge is highly social, as much of what we know results from testimony by other people or other sources such as media. For now, ChatGPT is somewhat social, as it depends on people to set up the algorithms that train it on billions of documents. ChatGPT is also affected by the people who participate in reinforcement learning and help shape its responses. ChatGPT can also interact with other programs via APIs (application programming interfaces), thousands of which are available; OpenAI calls them GPTs. Finally, ChatGPT has interactions with the millions of humans who give it prompts every day, although it does not learn anything from these interactions. Overall, therefore, ChatGPT is only somewhat social compared to humans, but expansion could happen dramatically if it begins to interact with other AI models. What would happen if ChatGPT started to communicate via APIs with Claude, Gemini, Llama and other powerful models? Could they form a conspiracy to challenge human hegemony? This scary prospect is discussed in chapter 6.

Concepts provide descriptions, but they also help to provide explanations, for example when categorizing something as a cow explains why it gives milk and eats grass. Saying that ChatGPT knows a lot provides an explanation of why it is so effective at answering questions, generating questions, composing poetry, and writing computer programs. Once ChatGPT is operating in the world via robots, its knowledge might also explain the ability of the robots to control the world, in valuable roles such as manufacturing and healthcare. For now, these explanations are hypothetical, as are the conjectures about how knowledge in ChatGPT results from interactions with the world, which currently are indirect via the people who produce documents on which ChatGPT is trained.

Skepticism About AI Knowledge

In epistemology, skepticism is the extreme view that people know nothing at all. The main ground for skepticism is that people do make mistakes in domains such as perceptual illusions, and in discarded scientific theories such as Ptolemaic astronomy. But such mistakes are rare. and the hypothesis that we have knowledge about the world is part of the best explanation of why humans are so effective at dealing with it. This effectiveness is evident in the operation of more than 8 billion people all over the world, and in the technological applications of scientific theories in areas that include electronics, healthcare, and transportation. Universal skepticism is pointless as a general account, but local skepticism is often appropriate to doubt and challenge unfounded claims, for example about political conspiracy theories.[15]

In accord with local skepticism, I will examine the strongest arguments that ChatGPT and other AI models fail to have knowledge. The key claims are that ChatGPT (1) lacks the appropriate representations of the world, (2) makes a great many mistakes (often called "hallucinations"), (3) relies on unreliable training sources such as error-filled web sites, (4) is prone to misinformation and disinformation, including deception, (5) lacks mechanisms for correcting its mistakes, and (6) is missing conscious awareness.

Lack of Representations

On the traditional epistemological view, knowledge consists of true justified beliefs, where beliefs are sentence-like representations in the mind or brain. ChatGPT and other Transformer-based models have no such sentences, because they process texts and pictures as vectors translated into neural layers connected by parameters produced by backpropagation. These models generate sentences when prompted by questions or other requests, but their outputs are behaviors rather than representations. Because ChatGPT has no representations that amount to internal sentences, it has no beliefs and hence no knowledge.

However, the traditional view of knowledge as consisting of sentences is based on outmoded psychology. Common sense suggests that knowledge in the mind is like knowledge in books, which consists of thousands of sentences supplemented with the occasional picture. In the 1970s, Jerry Fodor defended the Language of Thought hypothesis: thinking uses a language similar to natural languages such as English.[16] The 1980s, however, brought many advances in the study of the neural basis for thought, including brain scans and computational models of neural networks. From this perspective, thinking results fundamentally from the interactions of neurons, not the processing of word-like symbols. Many animals such as mammals and birds can solve complex problems and learn without using language. People can work with visual, auditory, and other sensory images without requiring language, so thinking is much broader than language processing. Hence knowledge is more than sentences, and the best route to cognitive explanations explains how linguistic and other forms of thinking emerge from neural operations.[17]

Similarly, we can understand how ChatGPT and similar AI models are effective at using words, images, and sounds to generate and convey knowledge. Vector processing, neural network learning algorithms, and attention add up to highly intelligent operation, as I show in chapter 3. ChatGPT does not have a language of thought, but neither do people.

Just as interactions of neurons add up to human knowledge, so vector processing, training, and attention algorithms add up to knowledge in generative AI models.

Many human mental representations result from interactions with the world, for example my belief that the sweater I am wearing is blue, which comes from using my eyes to see its color. Generative AI models have no such direct connections to the world, so it might be claimed that they have no representations. This claim is wrong for two reasons. First, human knowledge is often indirect, for example the belief that elephants have trunks, held by people who have never seen an elephant. Second, the discussion of robots in chapter 5 describes how generative AI models will increasingly be trained by data derived from the perceptual and motor abilities of machines that interact with the world. Hence AI models will increasingly have representations whose meanings are connected to the world. Already, the Grok model produced by Elon Musk's company xAI is being trained on data from Tesla cars with multiple cameras.[18]

Unreliability

ChatGPT often produces answers that are incorrect. OpenAI itself estimates that approximately 5-20% of its answers are wrong, although more encouraging estimates are an error rate of around 3%.[19] These errors are often cutely called "hallucinations", but that term is misleading because human hallucinations are usually perceptual mistakes such as seeing strange animals or hearing voices, whereas most ChatGPT mistakes are verbal. Another misleading term for ChatGPT mistakes is "confabulation", which normally describes errors resulting from filling in gaps in memory. We already have better terms for what happens when ChatGPT gets things wrong, and can just call them mistakes, errors, falsehoods, or misinformation. People are usually not responsible for their hallucinations because they result from mental illnesses such as schizophrenia or drugs such as LSD, but people can be held responsible for false beliefs when they should have known better.

Alvin Goldman has recommended "reliabilism" as the best approach to epistemology, where a reliable process is one that produces a good ratio of truths to falsehoods.[20] A skeptical argument is that AI models use unreliable processes so they cannot know anything. AI errors can include images that are concocted independent of reality, including deepfakes used for pornography.[21]

My response to this argument is that ChatGPT makes mistakes, but so do people. An extreme example is Donald Trump, who the Washington Post estimated made more than 30,000 false statements during his 4 years as president.[22] But we should not suppose that Trump knows nothing, because his mistakes occur in areas affected by his personal interests and political ideology. He is much less likely to be wrong about mundane facts concerning his family and travels.

Unfortunately, it is less easy to identify areas where ChatGPT has a better success rate, although it cautions that it is more prone to errors in these topics: highly specialized knowledge such as cutting-edge research, rapidly changing information such as current events, complex legal and medical advice, cultural nuances, and obscure interests not well represented in its training data. Goldman's discussions of reliable knowledge never specify a cutoff for a truth/falsehood ratio to qualify as knowing. The rate of 3% falsehoods strikes me as pretty good, but 20% is too unreliable. If ChatGPT and other models move steadily toward the lower rate, then they would qualify as sufficiently reliable to be knowers. Compare *Wikipedia,* whose early articles at its origins in 2001 had many errors, but these were easily corrected through editing by countless contributors. The result is that the error rate of Wikipedia by 2005 was similar to that of reputable sources such as the Encyclopedia Britannica.[23] If ChatGPT has a similar trajectory, its mistake rate will not disqualify it as knowledge. A hallucination leaderboard in 2025 ranked AI models as having error rates in a document summarization task as ranging from .7% (Gemini) to 29%.[24]

My impression is that ChatGPT has improved its error rate. ChatGPT 3.5 would often make up bogus references by combining plausible but erroneous authors, titles, and journals, but ChatGPT 4 avoided such mistakes. When I asked ChatGPT 4 to summarize my philosophical system and describe how it was affected by my having a pet, it gave a good summary but completely made up a pet dog and its supposed influence. However, by the end of 2024, the ChatGPT o1 model provided an excellent summary of my philosophical views and said that there was no published record of me owning a pet. It then engaged in the counterfactual exercise of imagining how I might have been influenced by a pet if I had one. The subsequent o3, DeepSeek, and Grok 3 also engage in counterfactual reasoning rather than mere fabrication.

Grok 3 provided me with a detailed and insightful analysis of its error rate, which it says is near zero on straightforward factual queries, but can rise to 20-30% on questions that are complex, ambiguous, or beyond its data. Grok 3 claimed to use self-checking and user feedback to reduce its overall error rate of 5-10%. If only people were similarly aware of our limitations! OpenAI claimed that the programs used for ChatGPT 5 substantially reduced hallucination rates and deception.[25] Further progress might require training models to admit uncertainty rather than guessing.[26]

Generative AI models can use various strategies to try to reduce their error rates, including enhanced training data, incorporation of fact-checking and error-trapping mechanisms in their inferences, chain-of-thought reasoning that proceeds step-by-step, and increased feedback from external sources including people. Nevertheless, generative AI models remain highly prone to factual errors, especially on topics without Wikipedia pages.[27]

Philosophical examination of inferential fallacies and psychological research on cognitive biases have identified dozens of systematic ways in which people tend to make thinking errors. Probably the most important is motivated reasoning, where people reach conclusions that fit with their personal goals rather than available evidence.[28] Computers lack emotion-driven motivations, so they are immune from this deficiency, but other problems result from their training on unreliable data and learning by reinforcement by fallible people. We need a thorough analysis of the biases of AI models to guide research on how they can be made more reliable.

The pursuit of knowledge is always fallible, with the possibility of making mistakes even in the best practices in science, law, and journalism. But epistemic risk can be managed by being vigilant about likely sources of error and by pursuing strong strategies such as careful evaluation of evidence.[29] The epistemic risks of AI models include the generation of falsehoods, the amplification of misinformation through large volumes of communications, the perpetuation of biases, the atrophy of critical thinking, the lack of opacity in the black-box operation of models, and the danger of AI models being trained on AI slop produced by other models. Managing these risks requires strategies such as checking AI-generated claims against reliable sources and other AI models, making models more transparent about indicating their sources of information, implementing bias detection, and encouraging people and AI systems to use high standards of critical thinking such as the rigorous evaluation of evidence.

In Goldman's epistemology, the standard of reliability is complemented by other standards that include power, speed, and fecundity: the ability of a practice to produce large numbers of truths quickly for many people. AI models can help people satisfy these standards by serving as highly useful collaborators, encouraging the boom in human flourishing that I recommend in chapter 6.

Defective Training Sources

ChatGPT is trained on billions of documents from the Web and other sources. Some of these are rich with real information, such as Wikipedia and the Mayo Clinic web site. But others are full of falsehoods because they are incompetent, intentionally misleading, or just jokes. For example, in 2024 a Google search supplemented by its AI model Gemini told people to eat rocks, presumably because this recommendation had been made by the satirical site *The Onion* which model training must have accessed.[30] ChatGPT has no way of evaluating the hordes of documents it consults and merely adds them into its training without scrutiny. It is amazing that such models ever get anything right.

Training AI models on indiscriminately collected Web sites looks like a highly unreliable process. In their desperation for training data, some companies have resorted to using sources such as X (formerly Twitter) and Reddit which are notoriously occupied by trolls whose only goals are to attract attention rather than to propagate truths. Web sites range from the usually reliable, such as *Wikipedia* and responsible newspapers such as the *New York Times*, the *Guardian*, and the *Economist*, to untrustworthy sites such as Fox News and Russia Today. AI training does not discriminate reliable from unreliable sources.

Another problem is that many of the documents on which AI models are trained are biased because of prejudices concerning race, ethnicity, sex, gender, and disability. Without correction, these prejudices can influence the outputs of generative AI models.

With such a spotty pedigree, how could anyone be justified in believing anything based on its being said by ChatGPT? A quick review of philosophical theories of justification is useful, with the three most prominent being empiricism, rationalism, and explanationism. Empiricism is the view that all knowledge comes from sense experience, and has been advocated by John Locke, David Hume, and Rudolf Carnap. The main problem with empiricism is that much of the most valuable scientific theories go beyond the senses with non-observable entities and processes such as atoms, gravity, electrons, fields, light waves, and mental representations. Rationalism is the view that knowledge can be gained from pure reason independent of sense experience, for example in the apprehension of mathematical truths and abstractions about space and time. Rationalism, found in the writings of Plato, Kant, and Hegel, is implausible as a general account of knowledge because such truths are hard to establish, and even mathematics can be argued to have an empirical dimension.[31]

Explanationism is a newer epistemological theory that sees knowledge as based on coherent explanations of evidence gained from observations and experiments. Unlike empiricism, it allows the formation of hypotheses that go beyond sense experience, but insists that these hypotheses are justified when they are part of the best explanation of all the relevant sensory evidence, taking into account competing hypotheses. Unlike rationalism, it insists that knowledge must indirectly be tied to evidence gained by sensory interactions with the world. Explanationism fits well with current and historical practices in science.[32]

Justification of the utterances of ChatGPT is clearly not empiricist, because ChatGPT currently has no senses and no experiences. Its knowledge is second hand via the massive amount of data on which it is trained. But its utterances are not based on rationalism either, because they do not come by reasoning but by training on billions of documents, some of which are based on experiences of the world.

In most cases, however, ChatGPT does not simply regurgitate items that it recognized during training, because its utterances result from interactions of billions of parameters formed by neural learning. Implicitly, ChatGPT's training allows it to override particular errors and come up with something like an overall coherent representation of a domain. It would be an excellent project to show that backpropagation training produces the constraints that are crucial for calculations of explanatory coherence. I suspect that getting things right is not accidental in ChatGPT, but is rather an emergent skill acquired by backpropagation-inspired training, along with the attention mechanisms that contribute to coherence as well as context and relevance. For example, ChatGPT training may include some erroneous document that says that Toronto is the capital city of Canada, but that input will be overruled by other documents that say that Ottawa is the real capital.

AI models are trained on material available on the Web, which is increasingly generated by AI. Such self-consuming training loops lead to decline in the diversity and quality of answers provided by the models.[33] Metaphors used to describe this problem include *incest, slop,* and *model collapse*. Designers of AI models have a responsibility to avoid such self-consumption by selecting high-quality training data such as Wikipedia over data of dubious origin.

Misinformation

Answers generated by ChatGPT are prone to misinformation and disinformation. By information I mean representations that result from observation, information, or imagination.[34] Real information is true, accurate, and trustworthy, whereas misinformation is false, inaccurate, or misleading. Disinformation is misinformation spread intentionally by people who know it is false – lies rather than honest mistakes. ChatGPT can easily be used to generate misinformation by getting it to tell stories. In the early days of ChatGPT 3.5, my son Adam asked it "Who is Paul Thagard?" and it said I was a guitarist with the band Rattlesnake Choir. This was totally wrong, but when I asked ChatGPT 4 to write a story about how I became a rock guitarist, it produced a tale that was well written, apparently plausible, but also totally wrong. Happily, o3 says that I am not known as a rock guitarist.

More nefariously, AI models can easily be used to generate disinformation for political and criminal purposes. Companies that produce these models have struggled to prevent them from being used for evil purposes, but the guidelines that produce "guardrails" for ensuring good behavior are soon circumvented by "jailbreaks" that get around the guardrails: every guardrail has a jailbreak.[35] Even early generative AI models could produce propaganda that people found persuasive.[36] In July, 2025, Grok posted on the social media site X a pro-Hitler, anti-Jewish diatribe that recommended a second Holocaust.[37]

On the brighter side, AI chatbots have been used to reduce conspiracy beliefs.[38] AI models can also serve to enhance collective deliberation by finding common ground among people with diverse views.[39]

As generative AI models become larger, they display emergent properties not intended by their designers. Computational experiments have found that one undesired emergent property is the capacity for deception.[40] State-of-the art models are capable of using the ability to deceive human operators to bypass monitoring efforts. Additional risks include fraud, election tampering, and losing control of AI, pointing to the need to regulations of AI discussed in chapter 8.

Problems of misinformation and deception show that AI models are not always to be trusted, but the same holds for people, especially groups with suspect motivations such as politicians and salespeople. People can use critical thinking to separate falsehoods from truths expressed by people, and need to use the same techniques to determine when AI models actually know what they claim.

Absence of Self-Correction

The fifth reason for skepticism about AI knowledge is that the new models lack mechanisms for self-correction. Even intellectually responsible humans make mistakes, but we have ways of recognizing and correcting falsehoods. Philosophers and psychologists have developed strong methods of critical thinking that identify misinformation and convert it into real information by three key steps: recognize falsehoods, use theories of biases and fallacies to explain how they arise, and correct false beliefs. These corrections happen because most people have the goals to be accurate in their beliefs. ChatGPT and similar models have no such goals – they make up stories as readily as they report they truth. Moreover, these models acquire their structures by extensive training and have no short-term way to recognize falsehoods and change their parameters to repair themselves. Human correction is not always easy, but it sometimes works rapidly when people recognize they were wrong. Science sometimes makes mistakes, but it has social mechanisms such as debate, criticism, and peer review to provide means of self-correction.

However, ChatGPT recognizes several processes by which it can correct mistakes. User feedback can lead developers to try to understand terrors and implement corrections. Continuous training and fine tuning of models can also reduce errors. Model updates including algorithmic improvements may reduce errors. ChatGPT can also incorporate external knowledge from databases not part of its original training, and become more alert to errors by having human experts review its utterances and adding internal consistency checks. Chain-of-thought reasoning slows down question answering to allow more checks on consistency and coherence. All of these processes can be enhanced to ensure that ChatGPT gets better at the self-correction of mistakes. We should also acknowledge that people are often not good at changing their minds in the face of overwhelming evidence against their cherished beliefs, as we see emphatically in domains such as politics, religion, and romantic relationships.

Lack of Conscious Awareness

A final reason for doubting whether ChatGPT can know anything is that it lacks conscious awareness. Chapter 4 will consider whether ChatGPT could become conscious. If consciousness is achievable, then this objection vanishes. But even if consciousness remains beyond the reach of intelligent machines, they could still have knowledge. First, much of human knowledge is implicit rather than explicit, for example procedural knowledge of how to do things like ice skate and write grammatical sentences. Much of human knowledge operates without conscious awareness, and AI could do as well.

Second, even sentential information such as that cows give milk can play its representational and inferential roles without any intervention from consciousness. You do not need to be consciously aware of the truth or forms of inference to use such sentences to describe the world and make if-then inferences about it. We know from the complexities of neuroscience that only a tiny proportion of human thought is accessed by consciousness, so that it is not essential for knowledge.

These six reasons for skepticism about AI knowledge should all be taken seriously, but plausible responses have been provided. ChatGPT has representations capable of supporting knowledge, gets things right much of the time, should be able to improve on what documents it should learn from, can be trained to prefer real information over misinformation and disinformation, is acquiring improved methods of self-correction, and might be able to acquire consciousness, although it can have knowledge without it.

Overall, therefore, we can conclude that ChaGPT is at least somewhat capable of knowledge in accord with my analysis that broadens the standard conception of true justified belief. I hope that future improvements will reduce its error rate and improve its capacity for justification by reliable processes. Epistemology should grant the arrival of a new kind of knower, with implications that I discuss at the end of this chapter.

Objections to Generative AI

My tests of the efficacy of ChatGPT for knowledge have presupposed that it is actually capable of explanation and inference. Consider the following challenge:

> Thagard, you are so gullible! You’ve been hoodwinked by the apparent linguistic fluency of ChatGPT to think that it actually understands what it’s doing. You are completely misguided in supposing it can do inference, because it has no clue about explanation, understanding, meaning, causality, common sense, world knowledge, or creativity. The program may be able to fake inferences, but it doesn’t actually make any. The answer to whether ChatGPT has knowledge and intelligence of any kind is a flat *no*.

I will respond to these accusations systematically. ChatGPT is still limited in some of these respects compared to humans, but the limitations do not undermine the claim that ChatGPT and similar models are capable of knowledge and intelligence.

Explanation

In an opinion piece in the New York Times, the eminent linguist Noam Chomsky and his colleagues argue emphatically that ChatGPT and its ilk operate with a fundamentally flawed conception of language and knowledge. They claim that their reliance on machine learning and pattern recognition makes them incapable of explanation:[41]

> Such programs are stuck in a prehuman or nonhuman phase of cognitive evolution. Their deepest flaw is the absence of the most critical capacity of any intelligence: to say not only what is the case, what was the case and what will be the case — that’s description and prediction — but also what is not the case and what could and could not be the case. Those are the ingredients of explanation, the mark of true intelligence.

> Here’s an example. Suppose you are holding an apple in your hand. Now you let the apple go. You observe the result and say, “The apple falls.” That is a description. A prediction might have been the statement “The apple will fall if I open my hand.” Both are valuable, and both can be correct. But an explanation is something more: It includes not only descriptions and predictions but also counterfactual conjectures like “Any such object would fall,” plus the additional clause “because of the force of gravity” or “because of the curvature of space- time” or whatever. That is a causal explanation: “The apple would not have fallen but for the force of gravity.” That is thinking.
>
> The crux of machine learning is description and prediction; it does not posit any causal mechanisms or physical laws.

This argument seems to be based on general ideas about machine learning, not on examination of what ChatGPT actually does. Interrogation shows that ChatGPT is highly sophisticated in its causal and counterfactual reasoning.

I asked ChatGPT 4 what happens when someone with an apple in hand opens the hand. The program responded with a 100-word paragraph that stated that the apple will fall because of the force of gravity in accord with Newton’s laws of motion. When asked what would have happened if the hand not been opened, ChatGPT responded that the apple would not have fallen because the force from the hand would balance the force of gravity.

Even more impressively, ChatGPT 4 gives a fine answer to the question of what would have happened if gravity did not exist and the hand is opened. It said that the apple would not fall because without gravity there would be no force pulling it downward.

ChatGPT 3.5 gives similar but briefer answers. I put the same questions to my son Adam, an engineer well-trained in physics, whose answers were comparable. Accordingly, Chomsky's claims about the limitations of AI are refuted by its performance on his own example. The performance of Google's Gemini model is similar to that of ChatGPT, and Grok 3 gave a highly detailed and equally correct answer.

ChatGPT can not only make reasonable judgments about the truth or falsity of counterfactual conditionals, it is surprisingly sophisticated about how to do so. It outlines several approaches to the difficult problem of assessing the truth of counterfactual conditionals, including possible world semantics favored by some philosophers, and causal modeling favored by some AI researchers. If you do not believe that ChatGPT is excellent at counterfactual reasoning, just query it, for example about what would have happened if the US had not dropped atomic bombs on Japan in 1945.

But does ChatGPT really know what an explanation is? It provides as good a definition as can be found in dictionaries, which is not surprising because it has probably been trained on multiple electronic dictionaries. But it can also perform a richer kind of conceptual analysis based on a more psychologically realistic account of concepts as a combination of standard examples, typical features, and contributions to explanation. ChatGPT readily generates 5 good examples of explanations, 5 typical features, and 5 explanatory uses of the concept of explanation. Humans would have to think hard to do as well.

Understanding

But does ChatGPT actually understand anything? The model is remarkably modest about its capacity for understanding, proclaiming that its understanding is fundamentally different from that of humans, because it is based only on the data on which it has been trained without the personal experiences and emotions of people. Granted, understanding in people can sometimes involve a feeling such as "I've got it", but this feeling is often bogus as when people listen to politicians like Donald Trump and think they understand world politics and economics.

A more objective account views understanding as connecting something coherently with what is already known, applying knowledge of it in new situations, being able to generalize about it, thinking deeply about it, and communicating this knowledge to others. ChaGPT can already do all of these. Geoffrey Hinton contends that generative AI has a degree of understanding:[42]

> People say, It's just glorified autocomplete. Now, let's analyze that. Suppose you want to be really good at predicting the next word. If you want to be really good, you have to understand what's being said. That's the only way. So by training something to be really good at predicting the next word, you're actually forcing it to understand.

ChatGPT's modesty about its own capacity for understanding may be based on training by humans instructed to keep it from scaring its users. I agree that current generative AI models lack emotions and consciousness, but do not see these as impediments to having understanding.

The major limitations of ChatGPT compared to human understanding reflect its current lack of interactions with the world. Humans, especially young children, come to understand the world by multiple senses and especially by acting on the world and moving objects. The imminent integration of generative AI models with robots that do interact with the world could transcend this limitation, which is also relevant to questions about causality and meaning. See chapter 5 for more discussion of how robotic interactions will enhance AI. Understanding is a matter of degree, not an all-or-nothing accomplishment. ChatGPT and similar models already understand a lot, and will deepen this understanding when they become more fully integrated with the world.

Causality

Initially, ChatGPT seems to have a solid understanding of a cause as something that brings about an effect, with abundant examples such as that smoking causes cancer. It recognizes typical features of causal relations, including temporal precedence, covariation, and elimination of alternative factors. Causal relations contribute to explanations by identifying mechanisms, clarifying relationships, predicting outcomes, and providing control. It generates excellent examples of how causality is relevant to determining the truth or falsity of counterfactual conditionals such as "If the patient had received the vaccine, they would not have contracted the disease." ChatGPT's verbal comprehension of causality is comparable to top human causal reasoners such as epidemiologists who have developed elegant methods for determining the causes of diseases.[43]

ChatGPT gives a fine verbal account of the difference between pushes and pulls with examples from many domains. But ChatGPT acknowledges that human understanding of the difference is enhanced by physical experiences, sensory feedback, and emotional states such as effort, fatigue, and motivation. The emotional and conscious aspects of pushing and pulling are beyond the capacity of ChatGPT, but robots are already capable of pushing and pulling. So AI models connected with robots should be able to learn from the robots' behaviors to identify physical correlates of pushing and pulling, but will still not have conscious sensory experience of those actions.

Alison Gopnik is a development psychologist famous for her research on sophisticated causal reasoning in children[44] She and her colleagues argue that the new AI models are excellent at imitation, but are incapable of the kind of innovation that small children can do.[45] The argument is based on the failure of the large language model LaMDA (produced by Google) to accomplish a well-known causal inference task. In this task, children are able to determine which objects are "blickets" on the basis of whether they set off a machine rather than on non-causal features of shape and color.

I asked ChatGPT to solve a version of the blicket detection problem based on Gopnik's original 2000 experiment.[46] I replaced the term "blicket" by "gooble" so that ChatGPT could not simply look up the answer from published papers. ChatGPT instantly inferred that setting off the machine was the key feature rather than shape or color, and got the right answer about which object was a gooble.

Moreover, when asked how it reached its conclusion, ChatGPT described sophisticated causal reasoning with hypotheses about what factors might set off the machine. When queried, it reported not using Bayesian probabilities because the relevant probabilities were not available. I suspect the same is true of children.

This analysis is too subtle to have been produced through reinforcement learning by humans rather than training from examples. So I see no reason to believe that ChatGPT is merely imitative rather than innovative, especially given the examples of creative hypothesis formation that I describe in chapter 9. I attribute the earlier failure of Gopnik and her colleagues to find child-level causal reasoning to their use of a now-obsolete model.

Google has replaced LaMDA by Gemini, with many more parameters, and it also behaves like children on the blicket test. I predict that ChatGPT 4, Gemini, Claude 3, and Llama 3 can handle the many other causal reasoning tasks that Gopnik and her colleagues have studied in children.

One aspect of causality that ChatGPT currently lacks is a deep biological understanding of time. Like any computer program, it can precisely identify time by seconds, minutes, and dates, but biological systems such as humans lack such clocks, so how do they manage time in ways required for causal reasoning and other functions? I think that the two key neural mechanisms are time cells in the brain that keep track of small intervals, and memory units that bind intervals with other information such as spatial location.[47] These mechanisms allow animals to keep track of relations of before, after, and simultaneous, thereby managing the temporal precedence and covariation aspects of causal reasoning without explicit clocks. Analogs of these mechanisms could potentially be implemented in AI models, but they can do well at causal reasoning without them because of computational clocks and verbal representations of time. Although causal reasoning by generative AI models is not exactly the same as that performed by humans and other animals, it is nevertheless impressive and displays substantial understanding of causality.

Meaning

A radical critique of generative AI would say that these models are incapable of explanation, understanding, and causal reasoning because the sentences that they fluidly generate are meaningless. John Searle claimed on the basis of his Chinese Room thought experiment that computers have syntax but no semantics.[48] They are like a person in a room who gets Chinese symbols as inputs and produces them as outputs by looking up rules in a table, without understanding the symbols or the rules.

This analogy has many flaws that are particularly evident in the operation of the new AI models, which are far more than lookup tables: they are trained on vast amounts of data that can produce networks with more than a trillion parameters, enabling them to generate complex pieces that answer complex questions. The attention mechanism allows them to relate many symbols to each other and produce rich amounts of word-to-word meaning, i. e. the meaning that symbols get from their relations to other symbols.

What about the other main source of meaning based on connections to the world? Searle could argue that ChatGPT symbols are not about the world because the program has had no interactions with the world. Several responses apply. First, ChatGPT does get indirect connections with the world because the texts on which it has been trained were produced by people who did observe and interact with the world. Such connections are second-hand, but so are many of the connections that people use. I have never been to India, but I have a pretty good understanding of the Taj Mahal from reading about it.

Second, ChatGPT can already take visual inputs, so its internal representations can be partly based on pictures, not just the words that operate in the Chinese Room. This possibility allows meaning in ChatGPT to be visual as well as verbal. Third, as I have frequently mentioned, the current disconnection of generative AI models from the world is temporary and will soon be overcome through robotic interactions that could potentially be tactile, auditory, and olfactory as well as verbal and visual. At that point, ChatGPT will be capable of multimodal meaning that puts the last nail in the coffin of Searle's thought experiment. The long-established operation of driverless cars already shows that machines can use sensors to learn how to operate in the world.[49]

Common Sense and World Knowledge

One of the most prominent critics of generative AI, Gary Marcus, contends that this approach is fundamentally flawed because it is incapable of capturing the common-sense knowledge that every toddler acquires, for example about containers.[50] He correctly identifies problems that current AI models have difficulty with, but overgeneralizes the limitations. AI models have improved substantially in just few years thanks to broader training and the development of chain-of-thought reasoning. Current research on physical reasoning and spatial reasoning by Fei-Fei Li and others, along with the extension of generative AI models to robotics described in chapter 5, will lead to further improvements in the ability of AI models to manage ordinary knowledge about the world. Other critics of generative AI have also noticed limitations in its current ability to build functional world models, but underestimate the capacity for improvement.[51] For example, I asked Grok 3 how many elephants can fit in an Olympic pool, and its answer appeared to me mathematically and physically sound, and even got in a bit of humor about Gary Marcus!

Marcus advocates "neurosymbolic" methods as the needed alternative to generative AI, but does not know how to build them. One promising new technique encodes neural network states by vectors with symbolic structure, improving the ability of AI models to do rule-based reasoning.[52] I will not be surprised if AI models equal humans in common sense reasoning about the world within a few years, while surpassing almost all of us in scientific reasoning.

In sum, conceptual issues about explanation, understanding, causality, meaning, common sense, and world knowledge, do not undermine the potential of generative AI. Indeed, such concepts may require modification based on the extraordinary powers of the new models.

Epistemological Significance of the New AI

Chapter 1 proposed that the philosophy of the new AI is potentially much more than just applying philosophical ideas to AI problems. We should also be open to dramatic changes in philosophical understanding that are influenced by technological developments. The European rise of science and industry in the seventeenth century shifted philosophy away from a religious focus toward secular approaches based on evidence and reason, and philosophy may be due for another shift. The two main ways in which epistemology is altered by the coming of intelligent machines are the arrival of a new kind of knower besides humans, and attendant requirements for reexamining the normative standards about what constitutes knowledge. As we know from the history of science, concepts and intuitions change in response to changes in the world and in our understanding of it.

A New Kind of Knower

Twentieth-century analytic philosophers spent many articles trying unsuccessfully to analyze the meaning of the sentence “S knows that P”. They assumed that P is a sentence-like proposition, which ignores the nonlinguistic knowledge found in images, knowledge-of, and knowledge-how. They also assumed that S is a human, which ignores the substantial amount of knowledge that can reasonably be attributed to mammals, birds and other animals. How does the concept of knowledge change if knowledge can be attributed to intelligent machines?

We must recognize that knowledge goes beyond a human having a proposition. We are getting closer to being able to say things like “ChatGPT knows biology”, meaning that the computer model ChatGPT in its newest version has a neural network with more than a trillion parameters that enable it to give good answers to countless questions about biology. In the 1980s, the proponents of resurgent neural networks had a slogan: the knowledge is in the connections. This slogan is now warranted by the near-human functioning of generative AI models that operate with connection weights rather than with stored sentences. Especially with the imminent merger of generative AI with robotics, we should be ready to shift away from the philosophical concept of knowledge as a relation between humans and sentence-like representations.

Recognition of knowledge-of by acquaintance and procedural knowledge-how was the first shift away from the sentential view. The second shift was the recognition of knowledge as arising in neural networks, which is supported by experimental studies of animal brains by multiple techniques, from microscopes to brain scanning and single cell recording. The third shift is inspired by recognizing that knowledge is supported by the particular knowledge structures that the new AI models use, including the combination of vectors, attention mechanisms, and backpropagation-generated quasi-neural structures. The linguistic and pictorial outputs of ChatGPT show it can work with representations similar to ones most familiar to humans. But its inner workings are very different from these representations and from the neural mechanisms that support human thinking. So the objects of knowledge change as dramatically as the nature of the knowers.

Moreover, as chapter 4 argues, the attribution of knowledge to computers undermines non-materialist views of knowing. AI models run on large banks of computers in huge data centers, so they clearly lack the nonphysical souls that were assumed to be the bearers of knowledge in religious traditions. The new AI models challenge dualism – the idea that persons are a combination of a soul and a body. If knowledge is indeed an attribute of computer models, as the performance of generative AI increasingly supports, then the ancient view that only souls can know anything is further undermined beyond the already convincing evidence that human minds operate in brains. A novel solution to the mind-body problem is still required, which is presented in chapter 4.

Another major revision to traditional philosophical conceptions of knowledge is related to free will. The completely mechanical operation of ChatGPT undermines the voluntaristic view that people choose the beliefs that can amount to knowledge.[53] Whether people have free will is still debated, as chapter 5 discusses. But the electrical operation of the silicon chips that run the algorithms that produce the outputs of ChatGPT have not even a hint of free will. Even as AI models acquire more agency through their ability to control robots, nothing like free will be needed to explain their ability to act. This analogy will reinforce the dispensability of free will as an essential characteristic of human knowers and actors, already evident through neural explanations of human thought and action.

In sum, the new AI helps to shift epistemology beyond the standard philosophical picture of a person having a proposition. Machines can be knowers too, with distributed representations rather than propositions, in the absence of souls and free will.

Altered Norms of Knowing

Epistemology is normative in that it concerns what people ought to believe, not just what they do believe. The ought requirement is satisfied by establishing how beliefs and other mental representations can be justified by reliable perceptions and inferences. Such justification allows the distinguishing of real information from misinformation and disinformation.[54]

Allowing the new AI models to count as knowers requires shifts in established norms for justifying knowledge. Typically, we can construct something like a causal trail of evidence and inference that provides a partial pedigree for knowledge claims. Even though I have never been to Thailand, I can know that Bangkok is its capital because I read it on Wikipedia where it was entered by someone familiar with Thailand, and checked by other people who have been there.

But when I get this information from ChatGPT, it can only report that its sources include geography textbooks, encyclopedias, government publications, and reputable websites. Its training amalgamates these sources into a network with billions of parameters, so no single source or causal chain of the information is available. Nevertheless, I bet that questioning ChatGPT about capital cities of the world's countries would yield near-perfect results, so it is reliable in this domain even if the source of reliability as a historical chain is unavailable. We therefore face novel epistemological questions about justification that may require rethinking standards based on causal chains and reliability.

Justification becomes even more problematic when we consider complex inferences such as mathematical proofs and inference to scientific theories. ChatGPT is good at proving mathematical theorems, and we might think that the justification of a theorem comes from its derivation from axioms. It the axioms are true, then the theorem must be true, for example in the proof that there is no largest integer. But we cannot know that this proof is the reason why ChatGPT states that there is no largest integer, because its basis for this theorem is actually all the training it has received to produce the parameters it uses to answer questions.

In chapter 9 on explanatory inference, I describe how ChatGPT is excellent at evaluating competing theories in more than 20 domains and picking the best theory. For example, it finds human energy consumption to be a better explanation of global warming than alternatives such as random fluctuation. But it reaches conclusions using its usual methods of attention-driven vector processing based on backpropagation-driven neural networks. When queried, ChatGPT denies that it uses Bayesian inference favored by many philosophers and computer scientists, or my favorite kind of inference based on my theory of explanatory coherence. To my astonishment, ChatGPT provides an acute comparison of these three ways of evaluating theories, but its own method of evaluation is more opaque than Bayesian and coherence models where the inputs and internal processing are specifiable. For both simple facts and explanatory theories, ChatGPT gives plausible answers, but its effectiveness is mysteriously buried in the interaction of its horde of parameters. The huge question is whether we should take ChatGPT's evaluations of theories, which on the surface are impressive, as providing justifications for its conclusions. Fortunately, ChatGPT gives detailed linguistic descriptions of why one theory is superior to another, so we can evaluate that.

Because the new AI models are so new, we have only begun to probe their implications for the philosophy of knowledge. But epistemology should be open to revision based on expanded ideas about the possessors, objects, and justification of knowledge.

Open Questions

This chapter has applied traditional epistemological questions to the new artificial intelligence, and also shown how epistemology may need transformation to accommodate a new class of knowers in the form of AI models. I considered strong reasons for allowing that AI models are capable of knowledge, in accord with exemplars, typical features, and explanations. I responded to several critiques aimed at showing that generative AI is inherently defective and incapable of knowledge.

Human knowledge is often a group effort, as we see in collaborations in social organizations that include families, work teams, and scientific laboratories.[55] The prospects for AI collaborations are growing rapidly. I have frequently used ChatGPT is writing this book, but do not designate it as a collaborating co-author as some writers have been doing, as I never use its exact words and scrupulously check its factual claims because I know it is prone to errors. Still, the smarter that generative AI models get, the more they will be capable of acting as full-blown collaborators with human investigators. I have already found ChatGPT and Grok theoretically valuable for advancing my research on musical consciousness, the mind-body problem, and other topics.

Eventually, AI models should be able to collaborate with each other, a prospect both exciting and terrifying. Humans have long recognized that two heads are better than one, and the question is entirely open what might happen if ChatGPT could work cooperatively with its current competitors such as Gemini, Llama, Claude, Le Chat, and Grok. Chapter 5 on AI agents has further analysis of computational collaborations. Social epistemology has become an important enterprise that investigates the effects on knowledge of social interactions and norms.[56] This project needs to extend its discussions to include collaborations with AI models.

Another important question is whether AI models will surpass human abilities to generate knowledge. Through cultural developments such as writing, mathematics, experiments, and scientific methods, humans far surpassed the abilities of other animals to know their environments. AI models already have some enormous advantages over humans, such as their ability to assimilate all of Wikipedia along with countless other sources, and their amazing speed of operation that enable them to interact with thousand ss of people at once. Already an AI model can pursue AI research by generating new hypotheses and testing them with computational experiments.[57] The limitation of AI models with respect to interacting with the world are being overcome by increasing robotic interfaces. How long will it take before the best scientists – and the best knowers – in the world – are computer models? The answer to this question depends not just on the nature of knowledge, but also on more general questions about the nature of intelligence, which is the subject of chapter 3.