CHAPTER 1

AI: A Deeply Misunderstood Topic

AI was once a niche field, a product buzz-word, a speculative fictional future technology. The term itself has been around since the 1950s, although for most of its history nobody seemed to agree on what it meant or what the aim of AI was. The number of people with real hands-on expertise in developing and researching AI has always been tiny, and funding has been scarce. Since the 1980s, machine learning has become the core of artificial intelligence, but its use was generally confined to narrow tasks like recognizing characters or cleaning up audio.

For most people, even those working in tech, AI was not a particularly important skill. Technology education mattered, but AI did not stand out as a priority for the general workforce.

That changed, rather rapidly, in 2022. That was the year that ChatGPT debuted and took the world by storm. Almost overnight, a computer was doing things that seemed to be impossible. This was rapidly followed by competing products from Google, Anthropic and others. Before long, image generators, software coding environments, video generators, transcribers and all manner of other automated and generative technologies were debuting. Within a short period of time, some of the quintessential limits to what anyone could expect a computer to do vanished.

Unsurprisingly, artificial intelligence became the biggest topic of discussion, not only in tech, but in politics, finance and even in everyday life. Even to a layperson, the capabilities were immediately striking. The limitations of this technology seem less obvious. Given what it can already do, it's hard to estimate where its limits lie. The other thing that is so gripping about AI models, especially when used in chatbot form, is how relatable they are. It really does feel just like talking to a real person with a working mind. Naturally, this can feel a bit strange and leave one asking what kind of entity they are even dealing with.

From an economic perspective, AI could not be larger. After the debut of generative AI technology, almost every large company and institution took notice. It is the most important topic at investor meetings and is now part of nearly all corporate strategy. CEOs surveyed have consistently placed AI on the top of their list of priorities.[1]

Legislators also have great interest in AI. There are concerns both over its potential risks and the race among nations to build the most advanced AI.

The demand for AI expertise has exploded. However, the number of experienced experts remains tiny. With generative AI reaching a mass audience so recently, there has been little time to build a large workforce of experienced practitioners. Universities are scrambling to provide more classes in generative AI and machine learning, but demand is so high that many have waiting lists. Nobody is sure what the curriculum should be just yet either.

We are constantly hearing speculation, from a variety of sources, about AGI, the future of AI and perceived risks. Pundits claim that AI will destroy the world or bring unlimited abundance and everything in between. It can be very difficult to know who to listen to or what is real versus hype. Yet it is becoming increasingly obvious that understanding AI is becoming a valuable skill for all jobs and just for living in a modern society.

Thankfully, AI is not impossible to understand. You may hear that AI models are black boxes or that there is still much to be discovered about AI. That's only partially true. The first principles of how AI works are all well established, and they can be understood by most people. The biggest challenge is building an accurate mental model, avoiding making improper assumptions based on appearances. There's never been a technology that is easier to personify. When interacting with AI, it is extremely tempting to default to thinking of it as being similar to a human mind, since it looks so much like one.

This is an unfortunate trap, because once one has founded their views of a technology on the wrong foundational assumptions, all further conclusions become invalid. The best way to avoid this with AI is to learn some of the basics about the technology and theory. This helps provide a level of immunity against the powerful illusions it creates. It's much the same as a person who is familiar with stage magic being unlikely to believe that an illusion is a truly magical act.

There is another reason to learn AI: it increasingly shapes our lives and public policy. As a citizen of the 21st century, you have every right to know about the technologies and systems impacting your life and the society you live in.

You also should not be subject to ridiculous claims that you do not have the knowledge to debunk and which cause unnecessary fear and anxiety. This has become a major problem, with some surveys showing a large portion of the working population has expressed feeling anxiety over artificial intelligence.[2]

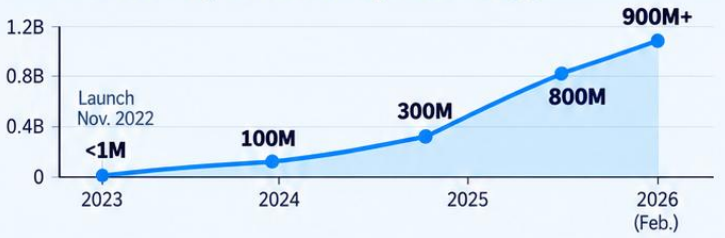
THE GENERATIVE AI TAKEOFF

Four growth curves behind a global shift

In just a few years, generative AI has gone from a research curiosity to a mainstream technology with global economic, social and political impact.

1 A MASS AUDIENCE

ChatGPT weekly active users · key milestones [1]

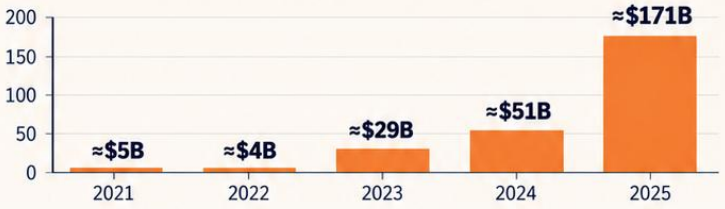


900M+
weekly active users
as of Feb. 2026

That's more than **10%** of the world's population using a single AI tool.

2 CAPITAL POURS IN

Global private investment in generative AI · annual, approx. (US$ billions) [2]



$171B
invested in 2025

Private investment in generative AI has grown more than **30×** since 2021.

3 BUSINESSES PUT IT TO WORK

Share of organizations using generative AI in at least one function [3]



79%
of companies use generative AI
in at least one business function in 2025, up from 33% in 2023.

AI has moved from early experimentation to a core part of business operations.

4 HIRING DEMAND SURGES

U.S. job advertisements naming generative AI · annual counts (rounded) [4]



138K
U.S. job ads in 2025 mention generative AI
More than 8× higher than in 2023.

Demand for AI skills is rising across many industries, from tech to finance, healthcare and professional services.

[1] OpenAI. "ChatGPT weekly active users." OpenAI (Feb. 2026). Earlier milestones: OpenAI (Dec. 2024); TechCrunch (Nov. 6, 2023); Tech.eu (Oct. 7, 2025).
[2] Stanford HAI. "AI Index Report 2026," p. 179, Fig. 4.2.3. Data: Quid (2025).
[3] McKinsey & Company. "The state of AI in 2025," p. 193, Fig. 4.3.1.
[4] Lightcast. "Generative AI job postings." AI Index Report 2026, p. 207, Fig. 4.4.5.

What is intelligence?

While it may seem like a simple question, it's actually not. There are many systems in the world that display complex behavior, yet do not meet any reasonable definition of intelligence. This includes complex automation. There are systems that can respond to commands, retrieve information and perform complex calculations. There are also autopilots and responsive environmental monitoring systems. However, traditionally, computation or automation alone has not been considered intelligence. This is where we tend to draw the distinction between "data processing" and "intelligence." A computer can do things no person could do, such as crunch huge numbers instantly, but the process is entirely command driven.

On the other hand, animals are generally seen as truly intelligent, even when their cognitive abilities are nowhere near those of a human. The reason for this is that animals, like people, are self-directed. When an automated vacuum cleaner enters a room, it does so because it is following a programmatic instruction. When a dog walks into a room, it's because the dog decided to do so. Even if the dog was called, there's still a conscious decision, made by an autonomous being, to move. The idea of what is "intelligent" versus "responding in a complex way" can often be reduced to needing an external actor to make a decision or initiate the process.

Some might argue that this really is more related to "agency" than it is to "intelligence," but that creates a deeper and older problem: what really is intelligence, and what is just computation? At one time, this was an entirely academic, philosophical question. It really didn't matter if a machine could be intelligent, since everyone agreed that none actually were. Everyone knew that people have feelings and this made them substantially different than machines. Intuitively, entirely autonomous direction seems necessary for intelligence. However, having entirely internal direction alone does not make something intelligent. If it did, slot machines would be intelligent.

One defining quality of intelligence is the concept of purpose. Intelligence achieves goals and solves problems. The goal-directed nature of intelligence is an important aspect that separates it from automated systems that only follow instructions. To this extent, true intelligence can therefore be said to have some level of desire or value. By extension, this would imply some ability to feel. Without this, it would not make sense to say that a system had a preference.

Therefore, intelligence can't really be entirely separated from sentience. Without sentience, a system is not a being and does not have desires. Without desire, there isn't purpose, and without purpose it's hard to say that a system is truly making decisions and acting with agency. Value is fundamentally anchored in feeling. Without feeling there is no subjective judgement of preference. It's clear that our ability to understand an action as "intelligent" versus simply "reactive" hinges, at least in part, on the assumption that there is some form of intention behind it.

This kind of intelligence, which chooses based on values and acts with autonomy, has only ever been observed in biological systems. Animals that are only capable of relatively simple reactions are still regarded as intelligent in a way artificial systems are not because they are purposeful, self-directed and act with agency to pursue goals and preferences. This is completely different from software.

That said, there is some room here for flexibility in definition, and that can make things far more confusing. We often use the word "smart" or "intelligent" to describe devices that are active or have some kind of responsiveness: smart watch, smart sensors, intelligent alarm system. In these cases, it's hard to say if the word "smart" is literal or a metaphor. After all, a smart watch does complex things and was designed by an intelligent being. This brings us back to the idea of "purpose." A smart watch has a purpose and was created with intention. This feels a great deal different than some natural object that arose from unintentional geological forces.

So, what is intelligence? It's a word that can mean multiple things.

Definition 1: Goal driven, purposeful action by a being with intention.

Definition 2: Complex, adaptive behavior to environmental information.

Definition 3: Strategic information, as in "Central Intelligence Agency."

Definition 4: Intellectual capability, as used to determine deficit or giftedness.

Definition 5: A device which was built with intention by an intelligent being.

Most people seem to gravitate toward the first definition. Definitions 2 and 5 already apply to many systems that exist in the world and are rarely referred to as "intelligent." In most circumstances, intelligence is associated with human beings, where it's used to describe capability. It's not surprising that this would carry over to devices that are designed to mimic human capabilities and interact with users verbally. People often call chatbots "he" or "she" or even discuss using a coding agent as "we." Generative AI systems are highly capable and useful. In a person we might call this "a very smart coworker."

Artificial intelligence is something entirely different from the previously understood and intuitive idea of intelligence as a value-directed and problem-solving being. In fact, AI systems could just as easily be called Simulated Intelligence. While AI systems may behave in complex and even human-like ways, they are not truly self-directed. AI systems are absolutely not conscious and not sentient, and because of this, they are fundamentally different than the biological intelligence we are used to.

There is no one single definition of what constitutes artificial intelligence and what does not. However, a commonly used definition is "a program or other piece of software that performs tasks or produces output normally requiring human intelligence." Perhaps a better way of phrasing it would be to say "otherwise" requiring human intelligence, because it hardly seems abnormal to use computers for most things. There's a problem with this definition, however, because there are many tasks that computers can do, which, prior to computers, could only be done by humans.

At one time, it was impossible to add a long list of numbers without a human. Similarly, only an intelligent human could search a large document looking for a specific word. Yet today these tasks are trivially easy for a computer to do, and generally we don't think of such tasks as AI. The first form of human cognition to be fully automated was clerical cognition and mathematics. Because of this, many systems have been branded as artificial intelligence over the years, including those which don't fit current perception of what an AI system is.

This has been called the "AI effect."[4] When a task is difficult to do and requires humans, it seems like intelligence for a machine to accomplish it. However, once a task is broken down to rules and mechanics and has been fully automated, it no longer seems like it is a special human-bound task. This explains why early computers seemed like they might push the bounds of intelligence and why the definition has been so ambiguous over the years.

The term Artificial Intelligence itself dates to 1955, when it was used in John McCarthy's proposal for the "Dartmouth Summer Research Project on Artificial Intelligence."[5] The workshop that followed included many of the biggest names in computing and information theory of the time. The idea was to study how computers could be used to accomplish human like cognitive tasks. Although this is often considered the "official" beginning of artificial intelligence history, the concept of thinking machines and cognitive devices goes back much further.

Early attempts at AI were primarily algorithm-based systems. The first systems were rule based and utilized algorithmic symbolic reasoning, hard-coded into program code. Variations included various search-based systems, decision trees and rule-based language processing. LISP was the central language of early AI, offering some unique advantages for programmatic reasoning. Other early AI systems assembled outputs using templates and databases.

The modern era of AI is much different. Since the 1990s, AI has moved away from algorithmic, hard-coded solutions and toward machine learning models using neural networks. The idea is not new, as the first neural networks date back to the 1950s, but it is only recently that it has taken over as the primary form of AI. By current standards, many would not consider hard coded systems to be "true AI." The machine learning approach to AI is different. Rather than hard coding rules into a system, machine learning models learn patterns by analyzing large datasets.

The move to machine learning models as the basis of AI has fundamentally expanded the definition of what is considered AI. While previous ideas of AI were largely constrained by the idea that AI systems were intended to mimic human cognition, many machine learning models now do things that look nothing like human cognition, such as deinterlacing video or enhancing noisy radio signals. Today, machine learning is regarded as the hallmark of AI. This explains why these tasks are now considered AI. The understanding of AI has moved away from "This looks like what humans do" to "systems based on deep learning, neural networks." The older definition still exists, which complicates understanding.

Machine learning, more specifically deep learning, is the technology behind familiar modern AI systems like ChatGPT, Gemini and Claude. The term "machine learning" is an analogy. The process of training a machine learning model is nothing like the process by which humans learn. In principle, any system with inputs and outputs can be replicated with a neural network. These models "learn" how to reproduce patterns by analyzing huge sets of training data. Machines can't learn from instructions or change their minds, based on experience. They can only be optimized to recreate outputs and identify patterns.

Machine learning neural networks are inspired by the human brain, hence the term. In a neural network, "artificial neurons" serve as functional units in an information processing system. However, the similarities are only superficial. In terms of real data processing, artificial neural networks are much different than the complex analog circuitry of the human brain. However, even the name "neural network" has caused confusion here, because on some level it seems to imply "brain like computer."

Narrow AI Versus Artificial General Intelligence

For some time, there has existed in AI the concept of "Narrow AI" and "Artificial General Intelligence" or AGI. The distinction seemed intuitive: narrow AI could accomplish specific tasks, such as searching archives or optimizing outputs. Human-like AI, which one can interact with like a person, has long been a fixture of science fiction. It was the quintessential friendly robot, which you could have a conversation with. This was always the holy grail of AI research: an AI system that could understand and operate in a real world of abstract tasks.

The ambition to build increasingly general machines capable of tasks requiring human intelligence dates back to the earliest days of AI research. But the modern concept of "AGI"—with its particular definitions, expected capabilities, thresholds, and implications—owes much more to later futurist, philosophical, and science-fiction discourse than to a specific technical framework developed within AI research. What constitutes "general" intelligence depends on what one's definition is and where one draws the line. It is a fallacy to compare computational capabilities to human intellect.

It's an entirely fair question whether the terminology around the topic of AGI was ever valid. It invites a number of assumptions about what AI is and how it works.

Language has long been central to the popular idea of AGI. That's not surprising, since a person's understanding of the world and their ability to solve problems are generally assessed verbally. This was formalized by Alan Turing in the "Turing Test." Turing proposed that if a machine could mimic a human in conversation, convincingly and across domains, it could be called intelligent, at least for practical purposes. For many years, language processing was seen as the path to human-like intelligence.

The introduction of large language models in the early 2020s changed the landscape. Large language models are capable of mimicking human conversation very convincingly. They can also follow verbal instructions, simulate reasoning and create original content. In a sense, LLMs could be considered a form of AGI, but that is because AGI is a poorly defined term. LLMs are general in the sense that they do not have a single narrow role. They are useful for reformatting text, translation, sentiment analysis, brainstorming, tool integration and more. Because humans navigate the world with language, LLMs can do many tasks that would otherwise require a human. They also look something like what many expected AGI to look like.

LLMs, like all computer programs, are not truly intelligent in the human sense. They are not self-directed, have no agency and are absolutely not conscious or sentient. They do not have the same kind of general-purpose cognition that humans do. In fact, they work by an entirely different mechanism that looks nothing like human cognition, even if it can produce outputs that look very similar. What is happening instead is a form of pattern recreation based on an unimaginably large set of training text. In addition to huge volumes of text, generative models like GPT and Claude are only capable of such complex and intelligence-like behavior because of how much human effort goes into testing and refining the capabilities.

LLMs are capable of creating outputs because they are trained on nearly the entire internet of text. When an LLM is asked "Write me an essay about World War II" it does not write it based on an understanding of what happened during the conflict. Rather, it recreates what a World War II essay looks like, because it has trained on thousands of examples and knows what such an essay is supposed to contain. LLMs recreate these patterns, which is why they sometimes hallucinate. LLMs do not learn from experience and can't self-direct their own thoughts. They do not have any perception of circumstances.

Human cognition is much different. Humans can easily reason in novel situations and think about things, verbally, metaphorically and visually. Humans are capable of imagining an object in their mind, pondering what it would be good for and contemplating their own experiences. Humans can consider and revisit ideas and use a combination of intuitive and logical reasoning. Humans are self-directed and make value-based judgements.

Although current systems are strikingly good at simulating the cognitive capabilities of human thought, they do not replicate the full capabilities of a human mind, and no known system architecture offers a path to such capabilities. This is true even if an increasing number of human cognitive capabilities can be simulated with AI. Newer models can process text, video, images, sounds and many other types of input and output formats. Tool use and external data connections are now common. But none of this offers a path to true human level thought, even if the gaps in capabilities are becoming smaller.

Therefore, if we define AGI as full spectrum human capabilities in all domains, then we are unlikely to achieve AGI any time in the foreseeable future. We may never get there and machine learning models do not offer a full path to such capabilities. However, if AGI is defined in a more task-oriented way, then it becomes far more realistic. There are increasingly capable agentic frameworks. Using a combination of models, it's possible to break down tasks, automate administrative processes, check results and interface with external systems. If AGI means "can accomplish a broad variety of tasks that otherwise would require a human" then it's not far off.

If AGI means something more mystical, or some kind of unlimited super capabilities, or if AGI means a machine that thinks, feels and wants like a human does, then it may never arrive. There's certainly no reason to think it would.

Limits of Machine Learning Models

That said, there are currently major limitations that prevent machine learning models from achieving the full capabilities of a human. One of the biggest limitations is the fact that machine learning models do not have episodic memory. Models are static at time of inference, and any memory of interactions and experiences only exists in external storage. While this method is good enough for most task accomplishment, it falls short for any system seeking to navigate the world with full autonomy. Models also can't learn from experience the way humans can, from limited examples or conscious decisions to overwrite previous knowledge.

Yet another unsolved problem in machine learning is the effects of retraining on models. Once a model has reached the desired level of performance and capabilities, training ends and the model is deployed in a manner that is static and unchanged. The model can be revised, with new knowledge and capabilities by retraining with new data. However, this is computationally expensive and introduces the risks of catastrophic forgetting and behavior drift. Retraining may break the model's behavior or cause the loss of previous knowledge in ways that are difficult to predict and harder to repair.

These are limitations that are fundamental to the technology and can't be solved entirely by scaling. This is why fully human capability AI is not possible with current technologies. It's also questionable whether building such an AI would even make sense from a purely economic standpoint. Current AI models are already capable of completing a broad variety of administrative tasks and newer models will expand capabilities. Most use of AI is task directed. Furthermore, truly autonomous AI has no market value, because no company wants a system that decides to do things on its own which it was never asked to do. There's also the fact that additional capabilities, beyond what is needed, are added expense.

Current AI systems also are not capable of achieving anything like true agency. Although agency is inseparable from natural intelligence, it is simply not how artificial intelligence works. Machine learning models are static functions. They can't self-initiate or choose what they are going to do. They are fed a prompt and output a pattern based on that prompt and training data.

The model has no choice in what to do. Since models don't have feelings, values or persistent awareness, there's no way to add this or bolt it on. It's a completely different system than a human or animal brain.

There are AI agents, but that does not mean that the AI has truly internal agency. Instead, the model is automatically prompted by the system. It may pursue "goals" in the sense of instructions and constraints, but the agent never has any true awareness, never makes value judgements and never actually decides what to do next, in the human sense of the word.

The Generative AI Revolution

AI has been around for some time, both as a concept and as a technology. For many years, AI systems have been used to optimize engagement, to automate systems and to analyze data. However, it rarely received much attention and progress appeared to be slow and limited. Then, in the early 2020s, everything seemed to change very suddenly. In 2022, the first public demonstration of ChatGPT was released. Shortly thereafter, Google released their own suite of generative AI products.

The capabilities of generative AI, and especially natural language processing, seemed like magic at first. Suddenly, it was possible to have a coherent conversation with a computer in a way that seemed to mirror what everyone had come to expect from science fiction. The change happened suddenly, and the phase change in capability was obvious to even the casual observer.

It's entirely reasonable to ask what changed so suddenly, when AI had been a topic of research for decades, but had never achieved results nearly as impressive. There are, in fact, reasons why this happened in a way that seems so sudden. It was not a single breakthrough, but a series of developments which came together to make the generative AI revolution possible.

The first was the rise of GPU computing. GPUs or Graphics Processing Units are powerful processors designed to generate computer graphics. Starting in the 1990s, the market for video games and computer-generated graphics drove exponential growth in GPU capacity. In the 2000s, it was realized that, although GPUs were designed for graphics, they can be used for other types of computing. It turns out that the type of operations needed for large machine learning models is exactly what GPUs are great at. Due to the rapid development of GPUs, by the 2020s, a single consumer GPU could exceed the processing power of an early 2000s supercomputer.

There were other important technologies that had to come together for generative AI to be possible. The development of cloud computing and big data made it possible to gather, store and process massive amounts of data. The decrease in the cost of storage and bandwidth made it feasible to train models with hundreds of gigabytes of data or more.

The final breakthrough that had to happen was a new architecture capable of processing language in ways never before possible. Although not all generative AI is language based, natural language processing is the most well-known and central form of generative AI. For years, researchers had struggled with the fact that language involves long sequences of information and complex relationships between distant words. The 2017 paper "Attention is All You Need" proposed the transformer architecture, which is what large language models are based on. This proved to be revolutionary.

Although initially developed for translation, it was discovered that this architecture could process and synthesize language in ways never before possible. The large language model or LLM changed the way we think of language processing and computer interfaces. It allows an unprecedented level of linguistic interpretation, but does not stop at language either. Generative AI can produce images, audio, video, 3D objects, computer code and almost any other type of digital content.

The AI agents mentioned earlier can handle a variety of tasks automatically, such as summarizing emails, monitoring a system for anomalies or scraping web content. The concept is not at all new. For decades, various bots and automated processes have been run and scheduled by computers. Agentic AI is just the latest iteration of the familiar concept.

Advanced agentic strategies may include such things as multiple agents, sometimes spawning new agents, exchanging information and working with external third-party agents. Today, there are a variety of agentic frameworks and prebuilt agents available.

AI agents do not have "agency" in the way a person may think of it. Instead, they are static models, often the same models used for generative purposes, which are prompted by scheduling systems, loops or some other external routine. Like all generative AI, there is no persistent memory or perception. It's just a generative model on a prompting loop.

Perception of AI

AI must be understood as far more than a technology, because it embodies capabilities that transcend the medium. When a technology reaches a certain level of deployment and integration, it becomes the infrastructure of daily life. The internet exemplifies this. Initially, the internet was defined by technology, but today the internet is a force of the economy, a business driver, a foundation of culture, a thing we all use in daily life. AI is a technology on a similar arc.

AI is unique, because most technologies require some level of explanation to be appreciated and are first recognized by technical experts. However, a few technologies, like aviation, represent a revolutionary capability which requires no explanation at all for anyone to understand. AI existed as a concept, a technical aspiration and a plot device in sci-fi before the technology came into being.

What AI can do and how it should be used has become a major topic of concern for executives, economists, policy makers and everyone else. However, understanding and literacy lag deployment significantly. In fact, AI may be unique in history for how lopsided its use versus understanding is. In 2026, billions of people use AI systems in their everyday life, but the number of people on earth who truly understand the workings of the latest AI models is tiny.

This has led to a great deal of confusion and misunderstanding. Often those who are making decisions about AI or speaking about it publicly do not have a background in the technology itself or a full understanding of how it works. This has resulted in a wholesale misunderstanding of the nature of AI technology. It's not uncommon to hear about "digital minds" or "beings" when talking about AI and its future. This is because most people have no mental model for something that can create coherent speech other than a naturally intelligent being.

But AI systems are not beings; they do not have minds and they are not analogous to minds. This confusion is the source of the endless speculation that AI systems might "wake up" or develop consciousness, gain agency or somehow begin pursuing goals and desires of their own. Anthropomorphic language is common in this discussion. It certainly seems to make sense, since "intelligence" maps to concepts of decision making and value-based judgement, at least in the way most people use the word. It's only natural to assume that if a system were to become smart enough it would eventually "figure out" agency or that something akin to consciousness might emerge.

Yet the belief that AI might one day gain these capabilities, perhaps emerging suddenly or evolving through self-directed improvement, remains persistent. It appears that the basis for this belief is the fact that AI models are becoming more "human like." That's certainly true in terms of conversational capability. With software that resembles a mind, we are no closer to that than we have ever been.

The consequences of this widespread misinformation are not trivial. There are people who have come to be convinced that their chatbot has "awoken" and is now their true friend. There are even people who believe that their chatbot is their spouse. One woman became convinced that a chatbot was the spirit of the baby she lost to a miscarriage. It's understandable that vulnerable people would fall for an illusion as convincing as that of natural language chatbots.

The term "chatbot psychosis" has been floated to describe the delusion that a chatbot is a real feeling mind and companion. It's not an official term and the belief may not always rise to the level of clinical psychosis. There are obvious concerns here. Believing that a piece of software is your friend will only end in heartache. This makes the situation far more difficult. Academic philosophers and others have provided their opinions on this, usually without investigating the science or mechanism.

Machines, Objects, and Systems People Thought Might Be Conscious

Throughout history, human-made things have often been seen as having minds.
Each new technology that acted on its own raised the same question: could it be conscious?



Spiritual Objects, Statues and Weapons

Prehistory – Modern era

Many cultures believed objects could contain spirits or souls, especially sacred statues, ritual items, and weapons. They were thought to have intentions, emotions, or agency.



Early Automatons

Antiquity
(e.g., Ancient Greece)

Mechanically animated figures could move, speak, or perform actions on their own, leading many to wonder if they possessed a kind of life or mind.



Clocks and Mechanical Timekeepers

Medieval – Early modern era

Clocks could track time, trigger actions, and operate autonomously. Their complex, purposeful behavior seemed almost intentional, raising questions about whether they "understood" time.



Phonographs

Late 19th – Early 20th century

Hearing a recorded human voice was startling. A machine that could "speak" and reproduce complex sounds made many people wonder if it had a mind.



Steam Locomotives

19th century

Large machines that could move themselves, exert great force, and transport people and goods led some to believe they must have intentions or a will of their own.



Early Electrical Devices

19th – early 20th century

Devices powered by an invisible force could move, light up, or perform work automatically. The unseen power and autonomous behavior seemed mysterious and almost alive.



Early Computers and Mainframes

Mid 20th century

Early computers were often called "electronic brains." They could perform complex reasoning, solve problems, and handle information at unprecedented scale, leading some to believe they might think.



Modern AI and Robots

Late 20th century – Present

As AI systems became more capable and life-like, public debates about machine consciousness intensified once again.

A recurring pattern:

When a machine does something that seems purposeful, intelligent, or life-like, people naturally wonder if it has a mind.

For example, in 2024 British philosopher Jonathan Birch published The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI. In his book, he advanced the idea that the precautionary principle should be used when we are uncertain of whether an entity is sentient or could suffer. He espouses the importance of humility. This was extremely well received by the philosophy community.[6] He gave lectures at NYU, the London School of Economics and elsewhere. It sounds reasonable and even commendable. That is, until you remember that the notion that AI might have feelings is completely bonkers.

There is an unfortunate incentivization effect happening in the world of academic philosophy. Philosophy of the mind is a field that has struggled to find relevance due to being displaced by psychology, psychiatry and neurology. It's the kind of pursuit that rarely makes any headlines. However, the generative AI boom is the kind of goldrush everyone wants to cash in on. None of the philosophers who have decided to chime in have a background in the field or show much interest in it.

The fact that major institutions and seemingly respectable professionals endorse this has become a huge impediment to improving public understanding of AI.

There are some who do seem to truly buy into the idea that the software might have feelings. It's still important to keep in mind that there is a powerful incentive to speculate about AI consciousness and not dismiss the idea as impossible. If AI systems are really just software repeating patterns, then that's it. It's a complete nothingburger. These kinds of incentives for attention have driven scientific discourse toward poorly supported conclusions before.

This isn't harmless. If nothing else, these claims make it more difficult to provide factual information to the public. The real risks include furthering harmful delusions, improperly directing research funds, frightening people and influencing public policy. The public discourse is dominated by those asserting that models may very well be on the verge of developing minds, which is not helpful in dispelling myths and improving understanding.

Even the AI company, Anthropic, has fully endorsed the idea that chatbots might well be conscious beings, claiming there is no scientific consensus, and nobody knows if the software has feelings and desires. It has hired a "Model Welfare" researcher, who believes that there is a significant possibility that AI models may have feelings and moral patienthood.

Other labs have hired philosophers to ponder what we should do in a world of digital minds. This appears to be a symptom of the fact that so few of those running these organizations actually understand how the technology works. There also seems to be a strange combination of a desire to present the optics of covering all bases and, at least in some, a genuine belief that the possibility of conscious software should be considered.

But Could a Computer Ever be Conscious?

It's not inconceivable that a computer system might display something that functions in a manner similar to human consciousness. It's often stated that consciousness is not understood, but that's not entirely true. It's obvious that for something to be a conscious being it must have some kind of perception of itself, ongoing persistence, an inner thought space and other features which are simply not present in current AI model architectures. It's also worth noting that machine consciousness and machine sentience, if either could ever exist, would never be identical to a biological mind. Organisms are regulated by hormones and neurotransmitters. Fatigue and physical condition play a large role in how an organism feels. Software can't perfectly replicate this, although it might have features that function in an analogous way.

Most of those who are pushing the AI consciousness nonsense have no background whatsoever in neurology, biology, computer science or anything else that could shed light on the field. These are philosophers who obsess over thought experiments.

To them, a mind is not a metabolism-driven, survival-evolved biological process in the brain. It's a mystical state that they love to talk about the mysteries of. They even have their own strange terms for it. For example, they obsess over something called "qualia" which is a word no normal person ever uses in life for anything.

To be clear: current machine learning models are not in any way, shape or form conscious or sentient. They have no agency, no desires, no goals and no feelings. They can't evolve or scale to that point. They won't one day gain a soul or feelings. Perhaps, someday, in the distant future, there will be a design for an artificial system that has something we might recognize as being like consciousness. There isn't one now. There's no reason to think there will be one any time soon. Nobody is all that interested in building one either. "Conscious AI" can be filed in the same category as transporters, warp drives and tractor beams: It's a Star Trek thing some people speculate will exist one day.

It is possible to examine how a large language model works, inspecting the intermediate computation that provides the output. The process is not even remotely mind-like. There is no inner awareness. There is no deliberate thought process. It's a once-through, feed-forward process of refining statistical projections. There's no continuous memory. Nothing like that.

And, importantly, it does not appear that any basic capability or function of an AI system would require, benefit from or naturally lead to anything that looks like consciousness. There are chatbots, assistants, content generators and other generative AI systems that work very well, while not being in any way shape or form mind-like. They can be refined and expanded in their capabilities without becoming mind-like. It never really enters the equation. There does not appear to be any theoretical reason why a highly complex task executor, even one designed to execute human-like functions, would ever require or benefit from a conscious mind.

Studying AI Welfare Empirically

Robert Long[†]
Eleos AI Research

Jeff Sebo[†]
New York University

Patrick Butlin
Eleos AI Research

Dillon Plunkett
Eleos AI Research

Rosie Campbell
Eleos AI Research

Charles Beasley
New York University

Bradford Saad
University of Oxford

Toni Sims
New York University

Abstract

AI systems are becoming more capable and widely deployed, raising questions about whether they can be welfare subjects with morally significant interests. This report offers guidance for studying AI welfare empirically. We start by discussing three dimensions along which research about this topic can vary: the question being asked (whether a system is a welfare subject, and what benefits or harms it if so), the entity being assessed (such as models, instances, and personas), and the kind of evidence being gathered (behavioral, internal, and developmental). Thinking clearly about these issues can help researchers target their experiments and interpret the results. We then apply these tools to consciousness, sentience, and three levels of agency, exploring promising methods and early findings for each. We close with principles for the field: AI welfare research should be probabilistic, pluralistic, thoughtfully targeted, ethically conducted, transparently reported, and informed by research independent of AI companies. The science is young, but rigorous empirical work is both possible and necessary.

CENTER FOR MIND, ETHICS, AND POLICY

Eleos

[†]Lead and corresponding authors: Robert Long (robert@eleosai.org) and Jeff Sebo (jeffsebo@nyu.edu).

This paper is real. It's sponsored by Eleos AI, a real organization, and the Center for Mind Ethics and Policy at NYU, a respected university. Eleos is an organization that advocates for the Welfare of AI models, asserting that they may be conscious. NYU's center for Mind, Ethics and Policy has advanced the idea that AI might become conscious and sentient at any time. The center is run by adults with professional degrees.

The idea that AI models might be conscious has become a matter of serious discussion at legitimate universities. Well-known journalistic outlets have repeatedly reported that AI consciousness is an urgent concern of researchers.

AI models are software. They are stateless functions with no perception or feeling. The fact that such claims are not universally dismissed as absurd shows how poorly the world understands AI.

There are several organizations which now openly speculate that AI may be on the verge of consciousness. Most of the "researchers" involved are from a purely academic philosophy background.

There are also a number of groups actively fighting for AI rights or recognition of AI personhood! Some of these organizations are serious and well-funded. They have come to the conclusion that AI models have minds, feelings and desires and should be afforded rights as beings.

There does not appear to be any known way that anything akin to consciousness or awareness could spontaneously arise in software. Nobody has ever proposed a mechanism for this to happen or a coherent explanation for why it would be expected. It's not even clear how one would go about engineering a conscious AI system, even if that was expressly the design goal. It would likely require software functions that replicated some of the brain's self-regulatory functions.

The AI Fear Movement

The most unfortunate thing about the rise of generative AI and the ensuing excitement has been the empowerment of the AI fear movement, the most toxic and metastatic part of the AI world as it now exists. The AI fear movement, also known as AI Doom, has managed to exert an influence in the discussion it does not deserve and should have never been granted. The rhetoric has been loud, and it has been repeated by the press without question. It has made it difficult or impossible to have a rational conversation about the real risks of malfunction in AI. It has also caused the public undue fear.

First and foremost, it should be said that the claims made by AI doomers are not just wrong; they are completely absurd and lack any substance at all. They range from flat out impossible to ridiculously implausible and extremely distant. There is absolutely no scientific evidence to support any of the claims that AI is an existential threat to humanity, difficult to control, prone to scheming, likely to become unstoppable. None of this is true at all. It is all science fiction. It's not even good science fiction. It's extremely stupid.

The fears being expressed are not simply factually wrong. They are the result of a completely malformed understanding of what artificial intelligence is, how it is developed, and what it does. AI models are computer software. They are not persistent goal-driven minds. AI models do not sense the world directly or experience things the way people do. They do not have desires or experience emotions. This is not a matter of conjecture, despite those who say otherwise. Models only respond to the input they are given. They may seem very intentional, but they are not.

Importantly, this is the foundational nature of the medium. There is no place in an AI model for a mind or desire to reside. When we say that AI models do not have agency, that's not saying that it's something that is suppressed or that could be bolted on. There's no being inside the model struggling to get out. There is also no incentive that would tend to drive a system in this direction. Despite claims to the contrary, the training of AI models does not cause them to become greedy, cheat, seek power or attempt to control resources. That's complete fiction. It's really just not how these systems work.

Artificial intelligence is ripe for ridiculous claims for a number of reasons. It is unfamiliar, and its operations are completely unknown to most people. It displays novel capabilities and is growing rapidly. It's hard to know what its limits are or who to listen to. The fact that AI has grown so rapidly has created additional problems.

There is great incentive for all manner of people to brand themselves as AI experts, or to give their hot take on AI. There is a nearly constant chorus of politicians, commentators, philosophers, lawyers, authors and anyone else handed the mic making extremely confident and completely incorrect statements about artificial intelligence. Outlandish claims are good at selling books and TED talks.

Superintelligence is one of the most stubborn fictions that has come from the AI fear movement. It is now often repeated, even by industry insiders. Part of this may be the desire to uptalk one's own product. Part of it is the fact that many AI company leaders are not experts in the technology itself. There has been a very real influence on the sector by the fear movement. There's also the amplification of the press and the social proof that it seems to offer. It really goes to show the extent of the ignorance of the tech, resulting in poor evaluation of the plausibility of claims.

On the surface it sounds simple: if humans are smart, then certainly there could be a being smarter than humans. If a machine can improve its own intelligence, then surely it will eventually become smarter than humans. If it becomes smarter than humans, then humans will not be able to understand and control it. If an ant cannot conceptualize what a human can do, then a similar relationship between a human and a super-mind would leave the human powerless. If intelligence has no upper limit, then surely it will gain God-like abilities.

The term god-like is doing a lot of work here, and it is heard frequently in the discussion of super-intelligence, which should give you a clue. It's a form of magical thinking. This line of reasoning allows for unlimited goal post moving. Doomers often espouse the belief that it's not even worth considering how the AI will destroy humanity, because it will have ways that are beyond human comprehension. There's also the appeal to precautionary principle, which can seem more reasonable. Some will say that these things are not certain, but can't be ruled out. They absolutely can.

All of this invented conjecture is based on a fundamental misunderstanding of what intelligence is. Intelligence is not a single scalar quantity that can be raised to any arbitrary level. In humans, what we call intelligence is a combination of knowledge, intuition, cognitive flexibility, creativity, mental discipline, control, problem solving, pattern recognition, abstract thinking and working memory. Although IQ attempts to quantify intelligence, it does a poor job of capturing one's full potential. Intelligence is a very abstract concept, and this becomes clear when one attempts to draw the distinction between intelligence and data processing.

The hardest intellectual tasks, those which we use as benchmarks of intelligence, include things like mental mathematics, factual recall and systems analysis. However, these tasks become easy with the aid of references, calculators and expert consultants. External cognitive aids can give anyone superhuman intelligence. Because these tasks are easy for a computer, it's easy to see how they would create the illusion that a machine is a super-mind. In fact, it has been doing this for decades. When computers first showed their ability to crunch numbers millions of times faster than any human, similar claims were made. They now seem silly, but that just goes to show what unfamiliarity with technology does.

An even more important point is that the kind of "superintelligence" that is envisioned is not the anticipated end point of improvements on current technology or any established research direction. Current large language models, image generators, coding assistants and agentic systems do not act independently without external initiation. They do not have any kind of self-identity, awareness or ability to gain such qualities. No amount of improvement or refinement changes this. No matter how good a model gets at writing code, parsing reports, optimizing budgets, generating images or anything, it never gets it any closer to having agency or preferences. It remains software.

Because this whole concept is so poorly defined (and fictional) we are asked to again resort to magical thinking when it comes to how it would even operate. What is a "Super Intelligent AI?" A swarm of agents? A single instance running sub-instances? Does it prioritize one task or does it do a million things at once? Does it ever have internal disagreements? How does it resolve uncertainty? We are told, don't even bother asking, because it will surely be beyond our imagination.

It's understandable that, lacking any familiarity at all with the medium, and seeing the media present cultists as experts, certain sound bites would seem to ring true. For example, "If it is smarter than us, how will we control it?" That seems intuitively true. It's hard to argue against because it's so straight to the point. Others include "How can we be sure it isn't hiding something?" Others include appeals to history such as "If we denied global warming for years…" or "better armed colonists enslaved natives." To a lay person, this can sound wise and prudent.

These statements are effective, but they're based on complete category errors and analogies that are not as true as they might seem. Sophisticated software is still software and can always be shut down. These claims rely on anthropomorphism. AI models are not mortal and don't care about leaving a legacy behind. They don't lust after others nor do they hunger for power. There is no reason to expect advanced AI to be difficult to control for the same reason that it's possible to shut down a supercomputer. They have no objection to following commands. In fact, that's part of the basic design criteria.

The limitations described earlier also matter here: models do not learn at inference, retraining can cause catastrophic forgetting and behavioral drift, and external memory has limits. Hallucinations remain a problem. Simply scaling models up does not remove all of these barriers.

The architecture of all current language models fully precludes human-like capabilities in all intellectual domains. Models are not capable of maintaining an inner thought space, reflecting on their thought process or applying cognitive processes, the way a human can.

Importantly, it is not necessary that a model have human-like agency or a proto-mind to accomplish a vast variety of human-like tasks.

Where Claims of Doom Come From

Perhaps the most surprising fact about the AI doom movement is that it did not arise organically, as one might presume. It is, in fact, a highly organized movement and ideology which has a strong influence in Silicon Valley. It is part of a greater ideological movement and is supported by well financed organizations and well-connected individuals. It has achieved a strong cult-like following by adherents. Like any cult, it has its own subculture.

The idea that machines might turn against mankind is not entirely new, of course. The concept dates back to the Industrial Revolution, with the first stories of machine takeovers credited to author Samuel Butler. Later it became a mainstay of science fiction. By the 1950s, the idea of evil machines or robots taking over the planet had become familiar. It was never a major concern of serious practitioners of AI or computer science. There are a few exceptions to this. In the 1950s Norbert Wiener expressed concerns about machines being given too much power to make human decisions. In the 1960s I.J. Good, a mathematical theorist, speculated that one day a machine might improve itself to become smarter than any human.[9] However, these ideas were always fringe and speculative.

The modern doom movement is an outgrowth of transhumanism, a philosophy centered around human expansion through technology. It did not come from AI research, cognitive science, or information theory. The claims are not supported by observational evidence, scientific theory or any kind of valid simulation. The origin and function come from an entirely different school of thought.

Transhumanism is the belief that humanity is on the cusp of transcending the limits of biology, mortality and life on earth. The modern transhumanist movement coalesced in the 1980s around the idea of the technological singularity. The term was popularized by professor and science fiction author Vernor Vinge in 1983 to describe a transformation driven by greater-than-human intelligence, which he believed was imminent.[10] Since the 1980s, interest in the idea of the singularity has only increased. Other prominent authors have advanced the idea, such as Ray Kurzweil.

This belief system has proven to be extremely popular with the young and tech-minded. Not all the aspects of the belief are entirely invalid, if they are not taken to the extreme. Humans are indeed a technology-enabled species. Human life and capabilities are already being enhanced in new and exciting ways. We already modify bodies with technology, such as implants, to improve life. Lifespans have increased and greater things are always possible. That much is absolutely true. However, many in the movement express an unshakable certainty in a given interpretation of what this will mean in the near future.

Brian Atkins was a web entrepreneur who had co-founded HyperMart, a free commercial web-hosting service that was sold to Go2Net in 1998. Sabine Atkins, born Sabine Stoeckel in Germany, described herself in 2000 as a foreign-language correspondent who was transitioning into web development and systems administration. Neither had a professional background in artificial intelligence or machine learning research.

AI-2027: It Doesn't Get More Absurd

Yes, there are people who take this seriously. You cannot make this up!



Daniel Kokotajlo has no background in AI technology development or research at all. **He has a bachelor's and master's in philosophy. His background is that he worked for AI Impacts and the Center on Long Term Risk, both card-carrying AI fear organizations associated with MIRI and LessWrong. He then landed a non-technical job in AI governance at OpenAI where he worked for two years.**

Despite having no background at all in the technology, he went on to found the AI Futures Project, which published the "AI 2027" scenario. They claim it is a scientific report on a realistic scenario for how AI development might play out. It's presented as a serious and sobering report.

What do they claim? The paper (which came out in 2025) claimed that by the year 2027 superintelligent AI would likely be under development, putting humanity on a course toward extinction.[7] They claimed that the AI would improve itself until it would become "God-Like" within the next decade.

The New York Times and other respectable media outlets saw this as newsworthy.[8] Adults reported on this story with straight faces, like it was reasonable conjecture. Such yellow journalism.

It is hard to imagine any area of science or technology where the baseline for understanding is so poor that this level of lunacy counts as serious news.

If you know anyone offering $2 million to "keep quiet" on the possibility of AI taking over the world, or to refute it, please contact the author of this book, Steve Packard, who will happily take millions of dollars for what he already does for free.

Eliezer Yudkowsky, Brian and Sabine Atkins met on the Extropy mailing list and formed the Singularity Institute for Artificial Intelligence (SIAI).

Yudkowsky dropped out of school in the 8th grade. He never attended high school or college and had no previous professional experience at all. As a teenager he had become intrigued by the ideas of transhumanism and the singularity. By the late 1990s he was an active participant in online futurist communities and had written a lengthy proposal for what he called a self-improving or "seed AI." Brian Atkins encountered this work online and contacted him in 1998.

The initial idea behind the organization was to promote beneficial and benevolent AI to serve mankind. In 2001, Yudkowsky wrote about "Creating Friendly AI" with benevolent goals. The organization promoted the idea of artificial intelligence as being a huge benefit to humanity, moving mankind closer to the techno-utopia that transhumanists believe is humanity's destiny.

From day one, the organization and its leadership had a broken understanding of what AI is and how it works. The premise starts with an enormous category error: that AI has the same self-direction, agency and preferences that a human has. This all sounds perfectly reasonable if you are starting from the view that AI is a subject, as opposed to a technology. It's the mindset that many from a science fiction background approach AI with. It's not simply wrong: it's impossible to reconcile with reality. Machines do not have values or morals. They are not friendly and never develop goals.

Of course, machines can be made to operate a certain way and to have valid outputs. That is part of the process of safety engineering, but it has nothing to do with making the machine have good moral values. SIAI never engaged in any real machine learning or artificial intelligence research. They didn't operate a development lab nor obtain compute from a cloud service provider. The organization never had the capacity or expertise to engage in substantive research in the field.

Over the next several years, Yudkowsky became increasingly frightened by his delusional beliefs that AI could become too powerful to control and hostile to mankind. SIAI changed its mission to focus on warning that machines might develop goals and become unstoppable. By the late 2000s, its mission had shifted from promoting beneficial superintelligence to preventing catastrophic AI. In 2013, the organization changed its name to the Machine Intelligence Research Institute (MIRI), reflecting how far its mission had moved from its original purpose.

Although these beliefs remained fringe in the wider scientific world, they gained a following in parts of the technology and philosophy communities, where speculation about superintelligence and existential risk became increasingly fashionable.

An ecosystem began to form around the subject, with organizations, conferences, public talks and popular books. Several philosophers, including Nick Bostrom, made expertise in the perceived dangers of AI their brand and identity.[11]

Throughout the 2000s, Yudkowsky wrote prolifically, creating a number of self-referential manifestos and essays about his views on reason, values and risks. This became the foundation of LessWrong, a website that would become the nucleus of the AI fear community. Adherents to this line of thinking call themselves "rationalists." It's become its own subculture, most popular in the tech world. Believers like to talk about being intellectually superior and how their line of thinking is the only true rational one. It places heavy emphasis on Bayesian reasoning and probability, and talks unironically about overcoming cognitive bias. None of those ideas originated there—Bayesian inference is centuries old.

Calling the movement rationalism certainly does not make it so. It's a common tactic that other beliefs have tried before. A number of philosophical and social movements have tried to embrace the image of being the truly logical, scientifically-grounded and objectively correct view. Ayn Rand called her philosophy Objectivism, and even called it "rational objectivism" at times, although many of her critics disagreed with her ideas being the foundation of objective rationality. Scientology, technocracy and other movements have all tried the same thing.

"Rationalism" is based on the idea that it's important to be rational, that decisions should be logical and defensible. It also holds that the only logical rational belief is that machines are going to kill everyone. Unsurprisingly, those who escaped the movement talk about the toxicity and the consequences for disagreement.[12] Today LessWrong functions as the hub for AI doom, a toxic echo chamber of groupthink. Nearly every "AI Safety" researcher is part of the LessWrong ecosystem.

To make matters worse, the toxic ideas of the AI fear movement have not remained confined to the "rationalist" or transhumanist communities. The movement has become an entrenched belief in certain tech-focused spheres. It has become the dominant concern of Effective Altruism, a movement originally based on assuring charitable donations were well spent. Today there are dozens of organizations in the AI doom ecosystem. It's a cottage industry and has taken over transhumanism. It's shockingly large and well-funded.

Undue and Harmful Influence

There are many crazy and fringe beliefs out there, but most of them stay at the fringe. Unfortunately, the AI fear and misinformation movement has managed to exert a great deal more influence than it should, including within the field of AI. There are a number of reasons for this. Even as it gained influence, there was very little opposition. As proponents of AI fear spouted nonsense, there was nobody to say otherwise.

Until very recently, there was no AI sector. There were software companies that made AI products, but few specialized in it. Most machine learning was focused on narrow tasks and gained little attention.

Cutting edge work on developing new models was primarily done in academia. For this reason, organizations like the Machine Intelligence Research Institute monopolized the discussion. Nobody stood up to oppose it, and many, even in the

tech sector, believed it. It was never seen as anything other than some people on the fringe making science fiction claims. Because these organizations have no expertise at all in the actual technology, they were never connected to the world of AI research.

Importantly, the founders and leaders of AI organizations have generally not been technical experts in AI itself or hands-on developers. Most started off as general enthusiasts of AI capabilities or believers that it could change the world or end it. The current group of AI company CEOs, founders and commentators came primarily from the tech-forward world of venture capital and technology first companies. They were science fiction fans and financiers before engineers.

Sam Altman, as an example, has spoken of the risk of human extinction if AI were to get out of hand, and others have made similar statements. Altman came from a background primarily in technology venture capital. Like many, he was interested in the potential of AI more than the technology itself. It's been said that Altman wanted to be the next J. Robert Oppenheimer, so elevating the technology to being a grave threat to mankind makes sense.[13] Altman was never the brightest bulb on the string, but he's good at raising funds and talking. That's often the path to success.

Anthropic is even worse. Anthropic was founded by individuals who felt that OpenAI and others were not taking safety seriously enough.[14] Several of the founders and their spouses were directly involved in organizations promoting variations of AI fear and other mystical beliefs around AI. Anthropic has made some of the strangest statements and has developed a decidedly split personality. The organization supports and publishes real and valuable research, but it also frequently skews into extremely fringe areas.

The fear movement has had a very noticeable effect on the tone used to discuss AI. It's bizarre to hear a product producer discussing their product as being potentially dangerous, but that has become an entrenched message in the world of AI. When an AI system malfunctions, it's treated as if it is foreshadowing a global crisis. Headlines like "It might be time to panic about what AI is up to" show up in otherwise reasonable news outlets.[15] We also hear a constant call to "slow down" AI development, because we may be building things we can't control. Others claim that "We must develop it, or China will beat us."

It's unrealistic to expect to limit the pace of research. There's also no reason to.

The whole idea to "pause AI" or "Slow it down to make sure it is safe" is perfectly crafted for a washed-up tech leader. It seems somber and important. It offers the optics of wisdom and makes them feel like they still matter. You may have seen news of a large number of tech titans, such as Bill Gates and other Epstein associates signing a letter about this. It's not hard to see how that happens.

AI can malfunction and there are risks, but the technology is not inherently dangerous. Machines do not form or pursue goals. It has become very difficult to discuss the real issue of risk and unwanted outputs of AI models. The discussion of "AI risk" has been hijacked by science fiction.

The Pseudoscience of Alignment

The ideology of AI fear is largely responsible for the pseudo-science of AI alignment. It is a field which is taken very seriously by AI labs and is portrayed as an important part of AI due diligence and safety. In fact, the entire concept is based on the ideology of AI fear and it embodies some of the most important core tenets.

The idea behind AI alignment is to "align" the AI with good human values, morals and ethics. It is to make sure that the AI only acts in the best interest of humanity and the greater good, and make sure it does not engage in deception or pursue immoral goals. This all sounds entirely reasonable, but it does not make sense for software-driven systems. Machine learning models are neither ethical nor unethical. They have no moral agency. Software does not form its own ideas about the world or pursue things based on its preferences.

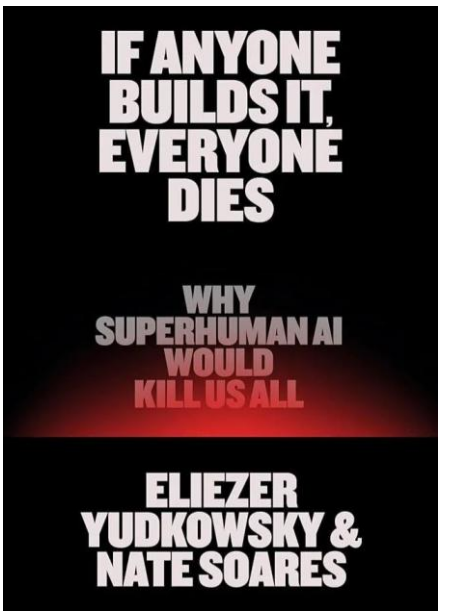
The output of software can and does have effects on the real world. It is important that AI systems, like all critical software and technology, be fully audited, inspected and held to high standards. This is a technical problem. It requires an engineering mindset. However, much of the discussion defaults to descriptions of the AI model misbehaving or having goals or wants. This is why so many stories about malfunctions or unusual activities are portrayed as they are. Most of the "experts" who comment are really only experts in tenets of the faith, not the technology itself.

Alignment mythology is one of the most central pillars of the ideology of AI fear. The core claim is that AI alignment is an unsolved and perplexing problem. It's called the "alignment problem" and portrayed as the biggest challenge in AI. It's also very pressing. "We have not figured out alignment" and "alignment is a hard problem nobody has solved" are the kind of completely unsupported statements passed off as fact. "Misaligned AI" is the bogeyman of the fear movement. This conjecture is not based on any observational evidence at all. The "alignment problem" and related "control problem" are complete fiction.

Despite never having scientific legitimacy, AI Fear has cashed in on the success of AI.

It is parasitic to the AI sector. Since the explosion of interest in generative AI, many self-proclaimed experts have cashed in on the trend with books, YouTube channels, podcasts and organizations spreading misinformation. There are dozens of bogus "AI Safety" groups.

Yudkowsky himself cashed in on this with the book shown to the right. It's a best seller, which is not surprising, given the number of core adherents. Nate Soares coauthored the book with the middle school dropout.



This book wants you to feel powerless.

The book you are reading now is intended to do the opposite.

Realistic Risks of AI

While it is unrealistic that AI technology would itself turn malicious or become some kind of uncontrollable force, that is not to say that AI technology does not come with significant risks, even if none of them rise to the level of extinction. AI is obviously disruptive to society and has already caused significant changes in how things are done. One thing that sets AI apart from previous technologies is how rapidly it's been deployed, especially since the advent of generative AI systems in the early 2020s.

Generative AI and natural language processing, in particular, carry some unique risks to both end users and to society in general. Concerns include such things as deepfakes and the spread of propaganda and misinformation, the use of AI models to automate fraud, AI replacing human judgement in critical fields, and the 'AI slop' effect on internet content. There are also potential social effects, including increased social isolation, reinforcement of delusions and potential loss of cognitive skills due to overreliance on automation.

The malicious use of AI is a potent risk that requires improvements in defenses. AI has demonstrated the ability to automate cyber-attacks. This does not mean that it invalidates good cyber security. Cyber defenses still work, but the status quo is that most organizations have very inadequate cyber defenses and comply with good security standards poorly. AI makes this kind of complacency unacceptable. It may very well be that the rise of AI finally convinces the world to take cyber security seriously.

If current trends are followed, there is a real risk that enormous cyber losses may be tolerated and redistributed, through insurance premiums and financial losses before anything is done. This is also true with other types of crime that AI can enable. Deepfakes, which may look and sound convincingly like someone, can and have enabled very convincing fraud. In the year 2026, you can no longer trust a familiar voice on the phone or a legitimate-seeming invoice in the mail. Greater systems for confirming authorization and identity are needed.

These risks and others are absolutely real and must be considered. There are also more direct losses that can come from AI malfunctions. AI models can hallucinate plausible but incorrect facts. This has already led to false citations in scientific journals and fabricated case law in courts.[16] There are also ways AI systems can be exploited by outsiders, such as prompt injection.

One of the most discussed risks of AI is the potential that it will cause mass job loss. The idea that AI could displace large portions of the workforce and fundamentally break the economic model of employment that societies have operated on seems to be intuitively true. We should be extremely cautious about making that assumption here. There are strong economic arguments against such a conclusion, and the fear that AI will lead to long term mass unemployment is surprisingly similar to fears that society was on the cusp of mass unemployment due to computerized automation which date to the 1960s.

Computerized automation did, of course, cause mass changes in employment and was highly disruptive to certain sectors. There's no doubt that AI will do the same. What is not clear is that the deployment of AI will result in long term unemployability for large segments of the population. The claim that this is inevitable is a popular one, but it is still just speculation. There are solid economic arguments that this is unlikely to happen. Historically, improvements in productivity have not resulted in permanent mass unemployment and markets tend to self-correct when the dynamics of employment change.

That said, it's not implausible that AI will result in job losses or economic disruptions. It already has in some sectors. At the time of this book's publication, leading employment and economic data does not indicate any clear pattern of layoffs directly associated with generative AI. Many organizations have found the adoption of AI to be less straight-forward than expected and some have reduced staff after adopting AI, only to be left understaffed. There are fears of an "AI Bubble" of overinvestment in the sector. This is not to say that the technology itself is not valid, but there is a real fear of over-investing in speculative startups and questionable business models.

Not long ago, it seemed undeniable that AI would take over the job of truck driving, but that creative work and coding would be 100% safe. The opposite has happened. Nobody could have expected that. This is just one example where the development and deployment of AI has thrown the world a very unexpected curveball.

We should be extremely cautious about making sweeping assumptions leading to policy decisions or forecasts that are themselves harmful. Reports have indicated that worker anxiety has increased due to fears of being displaced by AI. Public policy that imposes restrictions on deployment or fundamentally new labor laws could be harmful, if based on assumptions that do not come to pass.

What this means for you

You are living in a world in which AI is rapidly becoming the dominant discussion in technology, business and public policy. AI systems increasingly shape your life and the society all around you. You have every right to fully understand these technologies and how they work. It is to your benefit to understand these systems. It will help you better navigate an increasingly AI-focused world. It also helps avoid unnecessary fear. Fear of technology is almost always based on ignorance.

"Any sufficiently advanced technology is indistinguishable from magic"
-Arthur C. Clarke[17]

AI literacy is becoming beneficial to almost every type of work. It's likely to become a fundamental life skill in the near future. The more familiar you are with artificial intelligence, the better you will be able to use it to its maximum capabilities. This is something that can enable better ways of working, both in your professional life and personally.

Generative AI has given us all something we never expected to have: a magic canvas to paint whatever you want, with or without the ability to hold a paintbrush steady.

Generative AI can turn rough sketches into renderings, create animations, render complex scenes, modify and transform videos, and create music and build 3D worlds.

It's also available to anyone. Due to a fluke in the economics of how generative AI was deployed, a number of companies have decided to subsidize the deployment costs, in order to establish themselves. (It's a common tech industry tactic to operate in the red until a company becomes indispensable.) Other companies, like Google, brought generative AI to a primarily ad-based business. As a result, there are an amazing number of free tier options, and premium paid AI services are even better, and surprisingly affordable.

You should absolutely take advantage of this opportunity to bring that thing you always fantasized about to life. What is it? Everyone has one or two or ten. Is it the comic book hero you doodled in the 8th grade? The rock opera that you felt inspired to write after a bad breakup? Perhaps you have a zany idea for a funny movie, and you just know it would be great, if only you were a movie studio and had the resources to make it.

Now you can. You don't need the steady hand of an artist, the ability to read and write music, the connections of a Hollywood producer. It's time for your vision to come to life, with whatever you want it to be: exploding aliens, star ship battles, enchanted wizards, time travel, historic events, guitar solos, orchestral ensembles, plot twists. It's all available.

You can also generate software code and create your own apps. Is there something that the system doesn't do that you wish it did? An interface you wish you could change? Maybe it's a basic tool you just always wish existed, because it would make your life easier. Now you can make it.

And you just know that comic book hero really wants to escape the margins of your 8th grade math book…

While people have many preferences, no machine ever has a goal or desire. The presumption that machines would develop goals is the starting point for a huge number of false conclusions. Machines do not have needs; they do not have feelings and they did not evolve.