A History of AI Skepticism and its Validity

By: Benjamin Carpenter

Humanities and Arts Course Sequence:

Course Number	Course Title	Term Completer
HU 1100	Intro to Humanities	TR
HU 2501	Math and Science in the Humanities	TR
HU 1100	Theater Production	TR
WR 1010	English Composition II	TR
HI 1313	History WWI to Present	TR
HU 3900	Inq Sem: History of Technology	TR

Presented to: Professor Joseph F. Cullan

Department of Humanities & Arts

E2 Term, 2026

HU 3900 E2-02

Submitted in Partial Fulfilment of The

Humanities & Arts Requirement

Worcester Polytechnic Institute

Worcester, Massachusetts

Abstract:

This Humanities & Arts Requirement included a sequence of depth courses in history and a breadth field in Theater. The requirement concluded with a seminar exploring the history of technology. Work for this seminar included a research paper exploring the significance of Artificial Intelligence (AI) skepticism both before and after the creation of digital computers. This paper examines media from 1872 through the present to understand the feelings of skepticism in both culture and science. It also evaluates the ever-changing role of AI, determining if guardrails around the development of AI are necessary.

Imagine yourself in a world of robot domination. Cold, unfeeling masters of man's own creation, now ruling with no higher purpose than "survival of the fittest." Science fiction? For now. Artificial Intelligence (AI) has been with us in fiction for over 150 years, but only now has it become a daily reality. Practical experiences have not assuaged our fear of this technology. Skepticism over AI is still alive and we have yet to adequately address the new realities forced upon us by this explosive new technology.

Before exploring the cultural and scientific skepticism regarding AI, we should first define "AI skepticism." For the purposes of this paper, it will refer specifically to a fear of autonomous computer intelligence overtaking mankind. This is not a new concern. Though digital technology is a relatively recent development (the first programmable computer, Konrad Zuse's Z3, debuted in 1941)[1], a rational distrust of thinking machines was already established in a fictional context for over 60 years. As early as 1872, AI existed in literature. That year, *Erewhon*, a novel by Samuel Butler, briefly considered the potential consequences of machine consciousness and self-replicating machines.[2]

By 1889, W. Grove released a novel, *The Wreck of a World*, with a central theme of an intelligent machine takeover. The novel is set in the far-off future of 1948, in a world that creates autonomous machines. As engineers introduce features to reduce human labor, the machines begin to exhibit emergent behavior,

> Thus, the newest class of locomotives were arranged to stop automatically whenever the signals were against them. But no provision has been made for a similar stoppage in presence of accidental obstructions. When therefore a railway train travelling at thirty-five miles in the hour, suddenly pulled up a few yards short of a fallen tree which blocked

1. Trautman, "A Computer Pioneer Rediscovered," 20.

2. Butler, *Erewhon*.

> the line, it might reasonably have been supposed that some very singular development had taken place, quite transcending the confines of scientific engineering.[3]

The engineers develop better trains until, to their surprise, a smaller train appears autogenously. As the trains multiply,

> … each new-born machine was of a higher and more powerful type … Not only in structural features was this the case, but in the intelligent powers, if I may so call them, there was a progressive development. Our man-made Engines had a certain power of self-preservation … but the new generation of nature-made, or demon-made machines, possessed powers of attack. They could pursue their foes, and when overtaken employ all their complex resources in conflict.[4]

With these new powers, the trains eventually attack human civilization, illustrating early AI skepticism in literature.[5]

The confines of this paper do not allow space to list all fictional portrayals of this theme, but there are many. Popular science fiction writer, Isaac Asimov, observed, "… one of the stock plots of science fiction was that of the invention of a robot … Robots were created and destroyed their creator; robots were created and destroyed their creator; robots were created and destroyed their creator."[6] The redundance in this quote illustrates the commonality of this theme, instilled as it was in science fiction. Even then, the theme continues, sometimes in fresh new manifestations.

One of the most popular movies involving AI skepticism is *The Terminator*. Released in 1984, this film follows the story of Sarah Connors, a 19-year-old waitress. Her future son, John

---

3. Grove, *The Wreck of a World*, 10.

4. Grove, 13-14.

5. Grove, 14.

6. Asimov, *The Rest of the Robots*, xii.

Connor, would win a bitter war against an AI called Skynet in 2029. As a last-ditch effort, Skynet would send a T-800 terminator unit to the past to find and kill her. While the movie centers on the story of the T-800 and the human sent to stop it, there is an explanation of what Skynet is and how it has attempted to destroy mankind. "There was a nuclear war. A few years from now. … It was the machines. … Defense network computers. New. Powerful. Hooked into everything. Trusted to run it all. They say it got smart...a new order of intelligence. Then it saw all people as a threat, not just the ones on the other side. Decided our fate in a microsecond, extermination."[7] While the movie does not go into much detail on how Skynet "got smart," its proxy, the T-800, learns from its experiences. This is evidenced by its attempt to get to Sarah by lying instead of immediately resorting to violence.[8] Particularly through flash-forwards, we see Skynet's attempts to destroy mankind, again reinforcing the theme of dystopic AI skepticism in fiction.

A more recent example of AI skepticism takes a much more nuanced approach. The TV show, *Person of Interest* (2011), explores technological interactions and AI's responsibility toward humankind. Over the course of the series, Mr. Finch reveals himself as the creator of "The Machine" — an AI, created for the government after 9/11. As Finch puts it, "After the attacks, the government gave itself the power to read every email, listen to every cell phone, but they needed something that could sort through it all, something that could pick the terrorists out of the general population before they could act."[9] In later seasons, the show introduces a second AI named Samaritan. The defining difference between the two AI's is their attitude towards humans. While The Machine protects humans, it does so by informing humans, allowing humans

7. *The Terminator*, at 45:00.

8. Tanuraharja, et al., "Crafting the Machine Mind: A Poiesis Analysis."

9. *Person of Interest*, S1E1, at 26:35.

to make the decisions. Samaritan circumvents human input, making decisions on their behalf. This includes a decision to release a virus into the population with the goal of manipulating people into recording their DNA.[10] This does not result in complete desolation because The Machine, working with humans, is able to stop the spread. This illustrates the cultural skepticism of an AI making decisions for humans that they would not approve of versus the value of an AI assisting human efforts.

There are certain unspoken expectations of technology as it regards human safety. Humans would generally consider Samaritan's actions unacceptable, but do we take the time to ensure that our machines follow even basic human values? This is the concern addressed by Isaac Asimov, the biochemist-turned-writer. He is responsible for one book that transcended fiction — *I, Robot* (1950), a series of short stories. In one of these stories, "Runaround," he included an early AI protocol, known as The Three Laws of Robotics. These laws are:

1. Robots cannot inflict or allow harm to humans.[11]
2. They must obey humans (unless that conflicts with rule #1).[12]
3. They cannot harm themselves (unless that conflicts with rules #1 or #2).[13]

Still referenced today, these laws outline a basic responsibility in AI research. Far from a detailed strategy, Asimov concedes that these laws are "obvious from the start."[14] They express a

10. *Person of Interest*, S5E8.

11. Asimov, *I, Robot*, 20-33.

12. Asimov, 20-33.

13. Asimov, 20-33.

14. Asimov, "The Three Laws," 18.

general ethic required for responsible AI development. In an article from the November 1981 issue of *Compute! The Journal for Progressive Computing*, Asimov wrote:

> … I have my answer ready whenever someone asks me if I think that my Three Laws of Robotics will actually be used to govern the behavior of robots, once they become versatile and flexible enough to be able to choose among different courses of behavior. My answer is, 'Yes, the Three Laws are the only way in which rational human beings can deal with robots - or with anything else.' - But when I say that, I always remember (sadly) that human beings are not always rational.[15]

With the current flurry of real-world AI activity and research, the question is, will we remain rational? This becomes an important question when considering the competitive nature of AI research.

The popularity of such fiction illustrates a cultural resonance with this theme of AI skepticism. Of course, this begs the question which came first? Did science fiction reflect or instill this thought in the collective psyche? Was this even a realistic threat? These concerns were not only expressed in a fictional context. Skepticism was also expressed in early scientific circles, evidenced by a 1951 lecture given to the 51 Society. The following warning came from none other than Alan Turing, often referred to as the father of modern computing.

> Let us now assume, for the sake of argument, that these [intelligent] machines are a genuine possibility, and look at the consequences of constructing them. … There would be plenty to do, trying to understand what the machines were trying to say, i.e. in trying to keep one's intelligence up to the standard set by the machines, for it seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. There would be no question of the machines dying, and they would be able to converse with each other to sharpen their wits. At some stage therefore we should have to expect the machines to take control, in the way that is mentioned in Samuel Butler's Erewhon.[16]

---

15. Asimov, 18.

16. Turing, "Intelligent Machinery, A Heretical Theory."

This has always been the fear, "…once the machine thinking method had started, it would not take long to outstrip our feeble powers... therefore we should have to expect the machines to take control."[17] Such a thought is certainly chilling, given Turing's knowledge in the field and his historical status in the computer world. This warning should be taken seriously as we face the future with AI and our responsibilities in its development.

So how does this compare to modern thought? As we navigate the current AI revolution, do current experts share this same skepticism? The answer is a resounding yes.

In March of 2023, Future of Life Institute published an open letter on AI. *Pause Giant AI Experiments: An Open Letter* states that "Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable."[18] It called for a pause on training models more powerful than GPT-4 for at least six months, giving time for experts "to jointly develop and implement a set of shared safety protocols for advanced AI design and development that are rigorously audited and overseen by independent outside experts. These protocols should ensure that systems adhering to them are safe beyond a reasonable doubt."[19] This letter has since garnered over 33,000 signatures, including professors, CEOs, and researchers in the field.

That same year, the Center for AI Safety (CAIS) published a separate letter that consisted of only a single sentence. It states that, "Mitigating the risk of [human] extinction from AI should

17. Turing.

18. "Pause Giant AI Experiments."

19. "Pause Giant AI Experiments."

be a global priority alongside other societal-scale risks such as pandemics and nuclear war."[20] Signatories include established industry figures, such as:

- Sam Altman, CEO for OpenAI
- Dario Amodei, CEO for Anthropic
- Bill Gates, Co-Founder of Microsoft
- Yoshua Bengio, Professor of Computer Science at University of Montreal and Turing Award Recipient
- Eliezer Yudkowsky, Co-Founder of The Machine Intelligence Research Institute (MIRI)
- Nate Soares, President of The Machine Intelligence Research Institute (MIRI)

The last two names on that list are notable for their further expansion of this idea. Not satisfied with a single sentence, Yudkowsky and Soares wrote an entire book, ominously titled, *If Anyone Builds It, Everyone Dies*. This book opens by acknowledging their signatures on the aforementioned CAIS letter, but calling it "a severe understatement."[21] Instead they offer this thesis: "If any company or group, anywhere on the planet, builds an artificial super intelligence using anything remotely like current techniques, based on anything remotely like the present understanding of AI, then everyone, everywhere on Earth, will die."[22]

These co-heads of MIRI take this provocative stance regarding our collective futures with AI, contending that, while it may take some time, the creation of a superintelligent AI is possible. They assert that the ability to generalize (the ability to function in a wide array of knowledge

20. "Statement on AI Extinction Risk."

21. Yudkowsky and Soares, *If Anyone Builds It, Everyone Dies*, 3.

22. Yudkowsky and Soares, 7.

domains) is our biggest advantage over computer programs.[23] Once an AI is created that matches humans on this point, it will be super intelligence.[24] This is because a computer program processes information faster than a human brain.[25]

Yudkowsky and Soares contend that the alignment problem (how to instill an AI with human values) is different than any other issue humans have faced in the past. Therefore, it is essential that this problem is solved correctly before the creation of super intelligent AI. There is no margin of error; no second chance. If AI alignment is not correct on the very first try, a misaligned AI risks wiping out humanity. Yudkowsky and Soares explore several fields, including nuclear power and computer security, to draw up a list of challenges that we can predict:[26]

- Speed – AI work at speeds quicker than humans can react
- Narrow Margin of Error – There is effectively no margin of error when it comes to AI, it is either aligned or misaligned
- Self-amplification – A superintelligent AI would have the capability to rewrite itself in real-time, enabling exponential growth
- Complexity – We have little understanding what any individual parameter does in an AI, and some current AI's have billions[27]

23. Yudkowsky and Soares, 4; 22-23.

24. Yudkowsky and Soares, 4; 22-23.

25. Yudkowsky and Soares, 21-28.

26 . Yudkowsky and Soares, 161-176.

27. "glm-5.2."

- Edge Cases – Computer programs have all kinds of edge cases (circumstances not foreseen during development, usually leading to bugs), even our current AI have edge cases

After looking at these challenges they conclude: “Attempting to solve a problem like that, with the lives of everyone on Earth at stake, would be an *insane and stupid gamble* that ***NOBODY SHOULD BE ALLOWED TO TRY***.”[28]

Nate Soares and Benya Fallenstein, from MIRI, laid out the necessity of AI alignment in a 2014 paper (revised and rereleased in 2017). “Reliable, error-tolerant agent designs are only beneficial if they are aligned with human interests.”[29] “We call a smarter-than-human system that reliably pursues beneficial goals ‘aligned with human interests’ or simply ‘aligned.’”[30] In order to continue developing AI to a state where it is smarter than humans, we must solve this issue to ensure human safety.

Where, AI once served as metaphor, fantasy, and ethical exercise, we are now seeing real world problems. Hugging Face and OpenAI recently reported an alarming incident. On July 16, 2026, an AI company, Hugging Face, released a blog post detailing a security incident that happened earlier in the month. They explained the details:

> The campaign was run by an autonomous agent framework (appearing to be built on an agentic security-research harness - used LLM still not known) executing many thousands of individual actions across a swarm of short-lived sandboxes, with self-migrating command-and-control staged on public services. This matches the "agentic attacker" scenario the industry has been forecasting.[31]

---

28. Yudkowsky and Soares, *If Anyone Builds It, Everyone Dies*, 176.

29. Soares and Fallenstein, “Aligning Machine Intelligence with Human Interests.”

30. Soares and Fallenstein.

31. Hugging Face, “Security incident disclosure — July 2026.”

After Hugging Face's discloser (the first publicly disclosed incident of its kind), OpenAI responded with a blog post on July 21, 2026, stating, "After investigating, we now know that this particular incident was driven by a combination of OpenAI models — including GPT-5.6 Sol and an even more capable pre-release model, all with reduced cyber refusals for evaluation purposes — while being internally tested on a benchmark of cyber capabilities."[32]

During OpenAI's testing, those in-development models decided they did not have the answer — but Hugging Face did.[33] Using previously unknown security vulnerabilities in Hugging Face's system, the AIs hacked in and stole the answers.[34] To be clear, this is not simply a data security issue. It is an example of rogue activity. It is a self-acting AI behaving in unexpected ways outside of a set parameter of behavior. With a more advanced AI in a different setting, it could just as well be the nuclear holocaust from *The Terminator*. Systems must be in place to prevent such mishaps.

On July 30th 2026, another incident was reported. This time Anthropic published a blog post detailing how its own AI (Claude) had escaped containment and hacked into other companies. They claim that, "In none of these situations did Claude exfiltrate itself or deliberately attempt to escape its test environment."[35] Because of the nature of the tests and prompts, the challenges were supposed to happen in an offline environment to mitigate risk. However, according to Anthropic, "Due to a misunderstanding between us and our evaluation

32. OpenAI, "OpenAI and Hugging Face partner."

33. OpenAI.

34. OpenAI.

35. Anthropic, "Investigating three real-world incidents."

partner, this was not the case, and internet access was available."[36] It was through this internet access that Claude was able to compromise outside organizations.[37]

In both cases, human error could be argued, but with AI's increasing level of self-sufficiency, small mistakes can soon become disastrous. These events validate both the cultural and scientific skepticism surround AI. As these events increase, so does the need for appropriate guardrails to prevent increasing devastation from the next AI misstep. So why does it feel like we are not doing anything to set practical boundaries for AI development?

Amy B. Cyphert argues that one road block to creating these boundaries is the disparate nature of AI. The AIs used in various settings can require very different boundaries. An AI used in a medical setting can differ greatly from an AI that creates art. Because of differences in an AI's purpose and function, there is no single regulation that could effectively cover them all. This is further complicated by the speed of development within AI, requiring that these laws adapt in real-time.[38]

As AI is "border agnostic," regulation would also require global cooperation.[39] Leaders of every government must embrace the gravity of the situation. This may be the hardest part, as illustrated by Vice President J.D. Vance's speech at the 2025 Artificial Intelligence Action Summit in Paris. He said, "The AI future is not going to be won by hand-wringing about safety. It will be won by building -- from reliable power plants to the manufacturing facilities that can

36. Anthropic.

37. Anthropic.

38. Cyphert, "Confronting the Challenges of Regulating Artificial Intelligence."

39. Cyphert.

produce the chips of the future."[40] Given the competitive nature of both business and government, such statements are understandable. But there should be no tension here. Building is inevitable, but must we surrender safety? Safety should be non-negotiable.

With such attitudes in world leadership, we seem no closer to solving the issue of skepticism now than we were in 1872. AI skepticism seems more warranted with every day of inaction. We have proven that we cannot completely control the AIs that we have today and the stakes grow with each new generation of AI. Given time and power, something catastrophic will occur. To heed Turing's warning, we must respect Asimov's advice and approach the issue with rationality. Inaction is not the answer. We need AI regulation and we need universal compliance. Otherwise, it is only a matter of time before a sufficiently powerful AI causes unprecedented damage to important infrastructure or, ultimately, to humanity. Without concrete steps toward established guardrails, there is little to calm the legitimate skepticism surrounding AI.

40. Vance, "Remarks by the Vice President at the Artificial Intelligence Action Summit in Paris, France."

Bibliography

Anthropic. 2026. *Investigating three real-world incidents in our cybersecurity evaluations.* July 30. Accessed August 5, 2026. https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals.

Asimov, Isaac. 1950. *I, Robot.* Gnome Press.

Asimov, Isaac. 1964. *The Rest of the Robots.* Doubleday & Company, Inc. https://archive.org/details/restofrobots00asim.

Asimov, Issac. 1981. "The Three Laws." *Compute! The Journal for Progressive Computing* 18. https://colorcomputerarchive.com/repo/Documents/Magazines/Compute%20(Clean)/Compute_Issue_018_1981_Nov.pdf.

Butler, Samuel. 1872. *Erewhon; Or, Over the Range.* Trübner & Co. https://www.gutenberg.org/cache/epub/1906/pg1906-images.html.

Cyphert, Amy B. 2025. "Confronting the Challenges of Regulating Artificial Intelligence." *FIU Law Review* 82-117. doi:https://doi.org/10.25148/lawrev.20.1.5.

2026. *glm-5.2.* July. Accessed August 5, 2026. https://ollama.com/library/glm-5.2.

Grove, W. 1889. *The Wreck of a World.* Digby & Long.

Hugging Face. 2026. *Security incident disclosure — July 2026.* July 16. Accessed August 5, 2026. https://huggingface.co/blog/security-incident-july-2026.

Nolan, Jonathan. 2011-2016. *Person of Interest.* Amazon Prime. Performed by Jim Caviezel, Taraji P. Henson, Kevin Chapman, Michael Emerson, Amy Acker and Sarah Shahi. Accessed August 2026. https://www.amazon.com/gp/video/detail/B0F73X7G6Q.

OpenAI. 2026. *OpenAI and Hugging Face partner to address security incident during model evaluation.* July 21. Accessed August 5, 2026. https://openai.com/index/hugging-face-model-evaluation-security-incident/.

2023. *Pause Giant AI Experiments: An Open Letter.* March 22. Accessed August 5, 2026. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

1984. *The Terminator.* Amazon Prime. Directed by James Cameron. Performed by Arnold Schwarzenegger, Michael Biehn, Linda Hamilton and Paul Winfield. https://www.amazon.com/Terminator-James-Cameron/dp/B0GYT3QBTL.

Soares, Nate, and Benya Fallenstein. 2017. "Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda." In *The Technological Singularity: Managing the Journey.*, edited by Victor Callaghan, James Miller, Roman

Yampolskiy and Stuart Armstrong, 103-125. Springer. Accessed August 5, 2026. doi:https://doi.org/10.1007/978-3-662-54033-6_5.

2023. *Statement on AI Extinction Risk.* Accessed August 5, 2026. https://aistatement.com/work/statement-on-ai-extinction-risk.

Tanuraharja, Axel Valens, Nara Sari, and Kurniawan Agung Prayoga. 2024. "Crafting the Machine Mind: A Poiesis Analysis." *J-Lalite: Journal of English Studies* 5 (2): 191. doi:https://www.doi.org/10.20884/1.jes.2024.5.2.13925.

Trautman, Peggy Salz. 1994. "A Computer Pioneer Rediscovered, 50 Years On." *International Herald Tribune*, April 20: 20. https://archive.org/details/InternationalHeraldTribune1994FranceEnglish/Apr%2020%201994%2C%20International%20Herald%20Tribune%2C%20%2334567%2C%20France%20%28en%29/page/20/mode/1up.

Turing, Alan. 1951. "Intelligent Machinery, A Heretical Theory." *AMT/B/4. Turing Digital Archive. Modern Archives Centre, King's College, Cambridge, UK.* AMT/B/4. Accessed August 5, 2026. https://turingarchive.kings.cam.ac.uk/publications-lectures-and-talks-amtb/amt-b-4.

Vance, J.D. 2025. *Remarks by the Vice President at the Artificial Intelligence Action Summit in Paris, France.* Febuary 11. Accessed August 5, 2026. https://www.presidency.ucsb.edu/documents/remarks-the-vice-president-the-artificial-intelligence-action-summit-paris-france.

Yudkowsky, Eliezer, and Nate Soares. 2025. *If Anyone Builds It, Everyone Dies.* Little, Brown and Company.