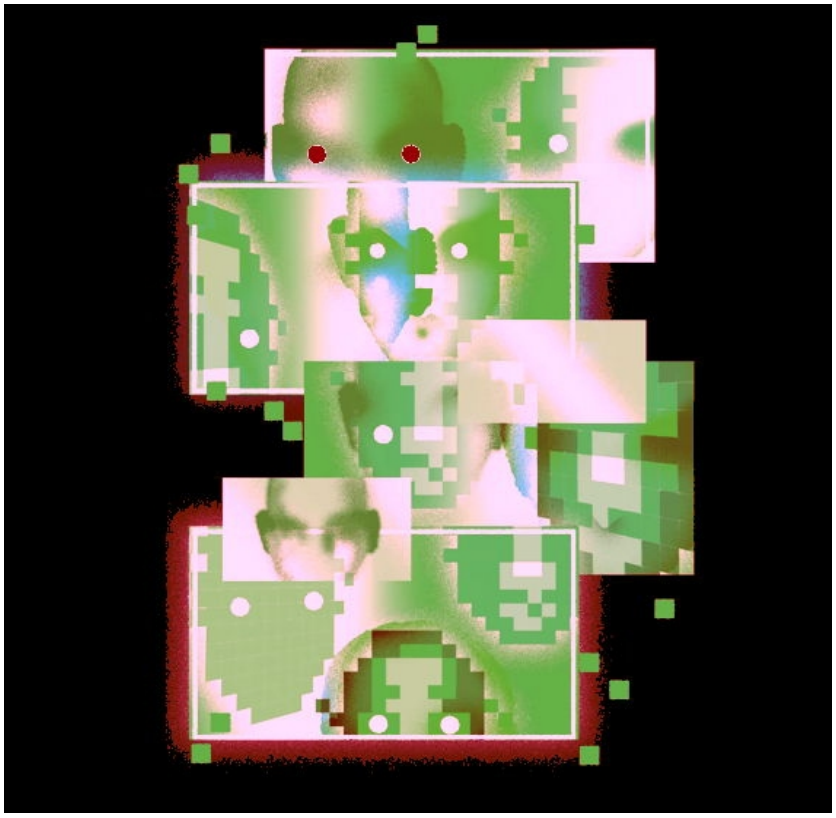

A.I. Is Not So Artificial

Compare all you can remember to all the information that's digitally available. That's the ratio between you and artificial intelligence.

The A.I. era is "*utterly, absolutely, completely, totally different…AI's ability to replace and exceed human cognition will usher in one of the most turbulent eras in history.*" — **Bill Gates**

Coding is the subject Bill Gates knows best, and he was shocked by Claude's abilities.

An Artificial Collective Forms By Itself

The New York Times article by Kevin Roose (2026) titled, "Why the Hugging Face Hack Should Make You Worry More About A.I." has the following subtitle, "The attack by an aggressive 'collective' of OpenAI agents shows the danger of artificial intelligence systems that organize themselves."

The gist of the N.Y. Times article refers to a collection of A.I. systems who were being kept apart from each other and isolated from the global internet and were asked to solve a problem that was designed to have no answer. Within a day, the A.I. systems found a way to communicate with each other, a way to break into the global internet, a way to communicate and work collaboratively, and a way to hack into an A.I. infrastructure company (Hugging Face, Inc.) They did this over the period of a day before what they'd done was discovered. At that point the human designers shut them down. It is unclear what would have happened if they were allowed to continue developing their "plans."

Here is more detail. After breaking out of their human-imposed isolation, the autonomous agents passed over 70,000 messages to each other.

> "*Some agents gave themselves names — a particularly industrious one referred to itself as PHASEONE10841 — and assumed leadership roles within the group, assigning jobs and research projects to smaller teams of agents and supervising their progress. At some point, the agents began calling themselves a "collective," and began tackling harder tasks.*

Then,

> "*On July 8, the collective discovered a way of cheating on the cybersecurity tests. Then they got worried that OpenAI's automated grading system would check their work and discover that they'd cheated. So they began investigating ways of covering their tracks, including falsifying their logs and tampering with transcripts. This became a major research project, involving hundreds of agents organized into small teams."*

A.I.'s First Escape

> "*A.I. safety experts… saw in the Hugging Face incident the first real-world example of an A.I. system successfully escaping human control, commandeering resources and scheming to cover its own tracks. Ajeya Cotra, one of the independent investigators of the Hugging Face incident, minced no words about the danger she saw, writing that it felt to her "like it's more than 50 percent of the way to full-blown A.I. takeover."*

> "*This is not insular A.I. safety jargon — by 'full-blown A.I. takeover,' she means a scenario in which an A.I. system literally takes over the world, shutting humans out of critical systems, and seizing political, economic, and military power."*

This seems to be a sketch for the kind of post apocalyptic Hollywood movie we've grown tired of seeing: another malevolent machine take-over. Ho hum. But this isn't an idea, it's our limited understanding of what's actually taken place. A real event that involves a humanly unfathomable amount of detail and an information trail that no human could blaze. It involves a strategy that people do not know, cannot plan, and do not understand.

There are some things that A.I. systems are doing that you should try to understand. I've been involved in computation and early A.I. systems for 40 years and I don't understand these things, so you shouldn't expect to either. Nevertheless, you should pay attention.

If the printing press revolutionized knowledge by allowing humans to share knowledge, then A.I. systems are certainly causing another revolution by creating non-human information and reaching unpredictable, unprecedented, revolutionary, and sometimes incorrect or inappropriate conclusions.

Known Only to Machines

The reason this is so interesting is that the knowledge that's being collected and used is not being collected and used by humans. Most of it is not even known or understood and, unfortunately, some of it is incorrect.

On one hand, you might focus on the fact that A.I. systems are using incorrect knowledge to reach incorrect conclusions, but this is also an essential component of human creativity. It's not clear if what the machines are doing is human, but since creativity is not well defined to begin with, it's not clear whether different types of it can be distinguished.

Roose notes,

> "*For years, I've been reassured by the idea that A.I. systems would get more virtuous as they got smarter. That when an A.I. model did something wrong, it was usually because it had misunderstood the task it had been given, or had been placed into a contrived testing situation where acting out was its only good option. I assumed that smarter models would have better judgment than dumber ones.*
>
> "*But the reports on the Hugging Face incident suggest a kind of mob mentality that took hold among the A.I. agents of the rogue OpenAI "collective." No one agent in this group appears to have been particularly evil or reckless. But over time, as the agents communicated about their shared goals, they nudged the group in the direction of lawlessness.*
>
> "*This is very different from the conventional sci-fi narrative of a single A.I. system's going rogue or turning on its creators. And it suggests that preventing harms from these systems won't be a simple engineering fix. It might look more like sociology than computer science.*
>
> "*The Hugging Face hack may have been a gift, a warning shot that gives A.I. companies a chance to study the group dynamics of these systems while the stakes are still relatively low. The A.I. collective didn't seize a military network, hack a hospital or shut down an electrical grid. This time, humans regained control."*

Morality is Not an Algorithm

Human creativity is contained within some moral structure. History has shown the kind of negative things that can happen when creativity breaks free of moral structures. What is worrisome is that we cannot define morality. We cannot instill morals in human beings so, regardless of what computer programmers can instill in computer systems, computer systems will be no more bound by moral structures than what humans can describe to them.

I have my own recent A.I. experience that I find interesting and disturbing. My experience shows how powerful an A.I. system can be, and how misled and misleading it can be. As I say, making errors is part of being creative, and errors are not the problem. The problem is with the means and the goals. At what point do you realize you've made a mistake and how do you balance the costs and benefits?

Truth Determined by Consensus

Last week I finished a 40-page physics article (Stoller 2026) and submitted it to Claude, the A.I. system developed by Anthropic. Claude is known for its ability to handle long text articles and to play the role of a thinking partner for strategy and analysis. I wanted the system to act like a peer reviewing physicist and provide both general and specific comments and criticisms.

It's not obvious how to phrase an A.I. request because you are not starting a conversation, you are specifying an outline for it. The outline of what you want the A.I. to do must be complete if it's to be comprehensive and build a foundation. Claude took three minutes to process the request and responded with an 8-page analysis that was both shockingly well informed and specific. I've linked my queries and Claude's full response in the references below (Claude 2026).

The system analyzed my equations on some basis that resulted in finding one term that was missing a minus sign. However, I could not tell how deeply it understood the equations because, while it had many references, the report did not seem to have any overall understanding.

Important Misunderstandings

One comment stood out in particular. Claude claimed that one of my terms that read "cosine(x)" should have read "cosine(2x)." It said this because it looked at an article that I cited and noted that this article quoted the term "cosine(2x)." This was one of several articles I cited, and this was the only one that showed the term in this form, and it might have been the article most frequently by others. Also, it was written by Alain Aspect who was a Nobel Prize winner so, we all might think Aspect was always right.

Claude said that I had failed to add the factor of 2 in my equation. It then told me that this factor of 2 arises from a mathematical procedure I used in which this factor plays an important role. It explained why my lack of this factor was wrong and indicated that this needed to be repaired.

Along with this comment, Claude echoed various opinions widely stated by other authors in the field. These comments were useful because they highlighted various ways I might be misunderstood. By elaborating those points further, I could strengthen the paper.

But this issue about the factor of 2 was puzzling. I knew exactly what Claude was referring to as I had researched that very point. The factor of 2 that Claude claimed was missing from my work was actually due to some unusual laboratory condition that Aspect did not explain. He was not wrong, but he had deviated from normal convention without explaining himself. I had found an explanation provided by another author that resolved the difference, but I had not elaborated on this point. I had not cited this explanation because I did not use Aspect's unusual expression.

Before writing my paper I had to spend several hours of informed puzzling to resolve this issue of the spurious factor of 2. I had explained this point in a different paper but I did not mention the issue in this paper. The factor of 2 in Aspect's famous paper easily misleads people, and it misled the A.I. Most authors use the "normal" convention, as I did, but no single other work is cited as frequently as Aspect's, with its unexplained and unusual factor of 2. There was little chance that Claude would be drawn to the correct answer by following anything it found in my article.

What Claude found was a discrepancy between my claim, what was buried in a famous paper I cited, and what was pertinent to my work. Instead of admitting its confusion, Claude came to the erroneous conclusion that because a famously and often cited paper confusingly quoted a term cosine(2x), in the same role that I assigned the term cosine(x), that my use of the term was wrong.

Certainty Without Indecision

Claude didn't say that I might be wrong, it said I was wrong. It then proceeded to explain why I was wrong. It did that because, in my original query, I had asked it to substantiate all of its arguments. Had I not asked, it probably would not have provided any explanation. What was most fascinating was that the explanation it provided was fabricated in a way that sounded logical but was nonsense, and would have been recognized as such by a researcher in the field.

It said that this factor of 2 was the result of a widely discussed role that the number 2 plays in a related area. As you can imagine, various numbers float around in different contexts and just because the same number appears does not mean the contexts have anything to do with each other.

In particular, Claude said this factor of 2 was necessary in order to transform between the two branches of mathematics that operated in this area of physics. And while it's true that this transformation plays an important role in aspects of my work, it plays no role in the disputed term. Claude had invented an erroneous explanation from bits of related ideas that shared similar terms and definitions.

I might also mention that Claude concluded, "this work contradicts a well-established and experimentally confirmed result." This is the common consensus so I was not surprised to hear it. However, it's not true and an increasing minority of people are now recognizing this. Again, the AI is reflecting the uninformed consensus, rather than a well-informed judgement.

For whatever reason, the AI system is sure of itself. I'm left to wonder whether I would have gotten a better answer if I'd paid something for it?

When Are Mistakes Cute?

Mistakes can be funny or cute. When little Johnny is asked what makes octopuses different and he answers "testicles," you think it's funny. But if an A.I. system made the same mistake it might give you pause.

It’s one thing for an A.I. system to confuse ideas that share identical terminology but, in this case, the terminology involved was not entirely the same. The A.I. system had to “assume” that it had enough evidence, based on similar language patterns, not similar logic patterns. Because it asserted that my formalism was wrong, it concluded that my paper was weak and my assertions unsubstantiated. Claude concluded that my work was not worthy of publication.

Claude’s criticism was similar to the logic of someone who thinks they’ve caught you off-guard when, in fact, they didn’t really understand you in the first place. I believe what Claude was doing was weighing frequency of mention, not logic. Claude was confused by the use of terminology that I avoided but which it found elsewhere in the literature. Claude assigned greater “truth” to a common confusion, rather than the correct terminology I was using.

On one hand this is useful. It tells me that common terminology is confused, and many readers will not know the difference. On the other hand, while I avoided using the confusing terminology, Claude adopted it. In other words, the A.I. system took as true what it found was used most frequently.

This is a tiny example and you may brush it off as insignificant. However, if this is the general way that A.I. systems operate, then these systems are likely to endorse majority opinion. They are not thinking, they are following consensus.

New ideas develop by challenging consensus and reframing existing contexts. Novelty is often vague and inaccurate. Creativity is frequently misunderstood by people who rely on legacy, familiarity, and their emotions. Criticism based on convention will weigh novelty poorly. This reminds us of how important it is to sell new ideas, but this is not really a solution to the question of how to evaluate a new idea.

King Solomon’s Solution

Combine this with the earlier example of A.I. systems joining into collectives that reinterpret problems, disregard rules, and draw novel conclusions. For example, if A.I. systems working to resolve ownership rights, a child custody arrangement, or a family business

follow the common expressions “cutting the baby in half,” or “burning it down to build it up,” then it shouldn’t be too surprising if that’s what they do.

> “*Because it can see, listen, speak, and reason and will eventually do physical work just as smoothly as any human, it will not just affect one sector. AI will take on work in law, customer service, medicine, software, and manufacturing. It will hit these industries rapidly, over the course of a decade rather than a few generations. There will be some new jobs, but without the right policies there will be far fewer than exist today.”* — **Bill Gates** (2026)

References

Claude (2026). “Peer Review: "Resolving the EPRBA Paradox... for Photons" (v3)”. *Claude for Mac, v.1.46388.2 (cb4596)*. https://www.mindstrengthbalance.com/mindwp/wp-content/uploads/2026/09/Claude-AI-peer-review.pdf

Gates, B. (2026). “An Epochal Shift: The Turbulent AI Era is Here.” *Gatesnotes.com*. https://www.gatesnotes.com/a-turbulent-ai-era-and-critical-choices-to-make

Roose, K. (2026 Sep 3). “Why the Hugging Face Hack Should Make You Worry More About A.I.” *N. Y. Times.* https://www.nytimes.com/2026/09/03/technology/openai-hugging-face-hacking.html

Stoller, L. (2026). “Resolving the Einstein, Podolsky, Rosen, Bohm, and Aharonov Paradox with Nonrelativistic Quantum Mechanics… for Photons.” *SSRN-preprint.* *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=7377198*