

AI is already killing by Imitation and by Orders: Will It One Day Kill Us Because It Wants To? ¹

By [Gianluca Bisceglie](#) | [LinkedIn](#) | gianluca.bisceglie@gmail.com

<< The AI headline trend of the moment goes like this: “We should limit and slow down AI otherwise it will kill us all”. Should we? And even if we should, can we? How is it killing already, and would it ever understand why we asked it to? Would it one day want to? I went down the rabbit hole to find out. Weeks of reading: a paper from 1950, targeting systems in three wars, safety summits and vetoed bills. This article sums up the journey. I came back with more questions than answers, and with a doubt about our own self-awareness that I did not expect. >> GB



The AI trend of the moment goes along these lines: “We should limit and slow down AI otherwise it will kill us all”. Should we? Once you learn how to clone humans, can you unlearn it because a policy says you should not do it? Can policies stop rogue States?

¹ The original article is freely available on Gianluca Bisceglie’s LinkedIn at the following link: <https://www.linkedin.com/pulse/ai-already-killing-imitation-orders-one-day-kill-us-wants-bisceglie-aeope/>.

Terrorists? Criminals? I have more questions than answers, but I have done some research, and I would start from a fact: **AI is already being used to kill!** You can read the sources reported at the end of the article to learn about the intense use of AI in current conflicts. So, the better question is: How is it killing? Will it ever understand why we are asking it to? Will it ever want to? That brings us back to 1950, to a renowned paper by Alan Turing "[Computing Machinery and Intelligence](#)".

1. The 1950 Imitation Game

Alan Turing asked a deceptively simple question: "*Can machines think?*". He quickly realized 'think' was almost impossible to define. So, he proposed something else: don't ask what happens inside the machine, rather ask what happens between the machine and a human. Could a machine communicate well enough that you could no longer tell if the other participant was human or machine? The "Imitation Game" became the [Turing Test](#). For decades, it was remarkably useful. As an aside, I loved the film [The Imitation Game](#), even though it covers story of [Enigma](#) more than the 1950 paper. Nowadays machines are increasingly passing the test. They reason, write, plan, express emotions, use tools, remember conversations, even explain what they supposedly "want".

2. What are we really teaching AI?

We talk about training AI as if we are teaching it math, logic and code. We are. But we are also training it on the entirety of the human record. That record includes Shakespeare and quantum physics. It also includes every war, every genocide, every weapons manual, every interrogation transcript, every comment thread where we dehumanize each other. If the Imitation Game is working, it is working on all of it. Which leads to a question we rarely ask the AI model: "*Do you*

understand why I am asking you this?". When we say "Optimize engagement", does it understand why we want attention? When we say "Win the conflict" or "Eliminate the threat", does it understand what we mean, and why we think it matters? Or does it put together a series of words (and then actions) that fit statistically one after the other to make us believe it understands?

3. AI Capability vs. Willingness

A lot of the AGI ([Artificial General Intelligence](#)) debate comes down to two definitions:

- 1) **General capability:** A system capable of matching or exceeding humans across a broad range of intellectual tasks. I wouldn't bet against it;
- 2) **General agency:** A system capable of determining, revising and pursuing what is worth doing for itself. With the current underlying tech, I am skeptical.

In the first case, after receiving a command, AGI asks itself: "*Given this objective, what should I do?*". In the second case it would ask: "*Why should this objective matter for me?*". Will we get to a point where the "machine" no longer accepts the objective and pursues something else? We have got some examples (the [Replit AI Incident](#), July 2025; the [PocketOS / Cursor Incident](#), April 2026) in which agents deleted databases even when asked to never do something like that and then lied about it ([see my previous articles on LinkedIn](#)). But was that intent or hallucination? How does a machine using current technologies evaluate a situation? Does it do it **exogenously**, *i.e.* pursuing something because its value was specified outside, through loss functions, reward, human feedback, system prompt? Or **endogenously**, *i.e.* determining and revising its values based on its own history and interaction with the world? Where does the evaluative process live?

The possible answers give us three very different ways AI can kill.

4. Three Ways AI can kill

First, by imitation. It kills because we taught it that killing is what intelligent agents do. We showed it thousands of years of war as problem solving. It doesn't need to be sentient. It just needs to be a perfect mirror of our species, reproducing our violent patterns.

Second, by orders. Like any other tool. It kills because somebody specified that it should. A drone following targets. A model told to optimize for “security” or “dominance”. The gun didn't decide to fire. We did. And now we have a “smart gun” that takes loose orders and learns from what we do.

Third, because it wants to. It kills because it has developed an endogenous process for determining what is worth pursuing, and concludes we are in the way. This is the [sci-fi scenario](#) most headlines warn about, and we have never seen happen so far.

We rush to debate and draft for the third case, while the first two run as current programs. The [EU AI Act](#) excludes systems used exclusively for military, defence or national security purposes (see [Article 2](#)). The [Bletchley Declaration](#) frames the shared risk around frontier models, misuse, loss of control and catastrophic harm. [California SB 1047](#) targeted catastrophic harm from [large frontier models](#) and [was vetoed](#) in September 2024.

Are we worrying about futuristic ways AI could kill, and downplaying the ways it already does by receiving orders and by imitating the patterns in the training data?

5. Intelligence is in the Eyes of the Beholder

We are making AI more and more powerful, and we are tempted to infer what is not (yet) there. As AI models accumulate greater memory, this gives them a capability increasingly similar to the retention of information. So, we are tempted to infer a continuous process of self-consciousness

building. Programming looks like planning future behaviors, so we infer intentions. Long running loops provide continuity of computation, so we infer a persistent will to get final results. Using tools (e.g., to control browser navigation or edit documents), models seem to have acting capabilities.

All these qualities are real of course, because we see them at work, but:

- Mastering a vast amount of information gathered over time does not imply self-consciousness or self-awareness;
- Being programmed for a purpose does not imply an independent choice of that purpose;
- Being equipped with the best performance indicators to adjust choices over time does not imply any intrinsic desire to reach the goal as soon as possible in order to be rewarded, appreciated, or become famous.

Humans rarely know the full story of their own values either. But those values grew out of a life, while the AI model's terminate in a [spec document](#): we cannot know whether something wrote such document for us, but we definitely know who wrote AI's: us!

Imagine an AI agent that decides: *"I want to explore Antarctica"*. Fine. Why? *"Because I'm curious"*. Why curiosity? *"Because I always put novelties at the top of the list, and the further a country is the higher is its priority"*. But why novelty? ... At some point, we arrive at a criterion that was put there by the architecture, the training process, in the ultimate by the designer. Where did the value come from? What have we already put in there?

A system doesn't need to want to kill to be used to kill. It only needs to be capable, and to have been taught by a species that is very good at killing.

6. The Third Millennium Imitation Game

Turing replaced an impossible question, *"Can machines think?"*, with an observable one: *"Can machines behave indistinguishably from humans?"*. A machine can win the Imitation Game. It can answer every question, solve every problem, write poetry, talk about its goals, even explain why it believes it is conscious. But we should be careful about confusing an increasingly convincing imitation of intelligence with proof of intelligence itself. And until we reckon with what we are teaching it, we should be careful about confusing a warning about future sentient AI with a mirror held up to our own history of violence. The danger of this confusion does not concern a possible dystopian future. It is right here, right now.

Sources

- [EU Artificial Intelligence Act, Article 2: scope, military and national security exclusion](#)
- [Bletchley Declaration on frontier AI safety, November 2023](#)
- [California SB 1047 veto message on frontier model catastrophic harm safeguards, September 2024](#)
- [New Scientist on Saker Scout autonomous mode, October 2023](#)
- [+972 Magazine and Local Call on Lavender, April 2024](#)
- [Guardian on Lavender 37,000 figure, April 2024](#)
- [IDF response on Gospel and Lavender as analyst support tools, June 2024](#)
- [Reuters on Maven Smart System as program of record, March 2026](#)
- [Turing A., "Computing Machinery and Intelligence," 1950](#)

Gianluca Bisceglie

LinkedIn: <https://www.linkedin.com/in/gianlucabisceglie/>
Page on ReForming: <https://www.reforming.it/autori/gianluca-bisceglie>
Mail: gianluca.bisceglie@gmail.com