Why HAL-9000 Intended to Kill All Astronauts Aboard Discovery

The famous novel *2001: A Space Odyssey* by Arthur C. Clarke and the eponymous film produced by Stanley Kubrick were released in 1968 (the novel was published a few months after the film's release; Clarke and Kubrick worked together on the screenplay). The collaboration of these two great artists proved to be extremely fruitful. It gave the world a work of art that had a colossal impact on several generations of not only viewers but also thinkers and researchers.

The first installment was followed by sequels—*2010: Odyssey Two* (novel, 1982) and *2010: The Year We Make Contact* (film adaptation, 1984; directed and produced by Peter Hyams).

Both novels and films complement each other perfectly, and when we speak of the Space Odyssey, we mean all these works.

The *Space Odyssey* is extremely convincing. Everything in it is shown so realistically and nuanced, as if it were not the authors' fiction but a story from the future transmitted to them by someone. Now that we have made serious progress on the path of creating real AI, the foresight of A. Clarke and S. Kubrick takes on not only aesthetic and philosophical but also existential overtones for us.

What Happened to HAL 9000

The spacecraft "Discovery" was sent on a mission to one of the gas giants of the Solar System (in the film, it was Jupiter; in the novel, it was Saturn). The task was to attempt to establish contact with an extraterrestrial civilization whose existence was confirmed by an artifact found on the Moon called the "Monolith."

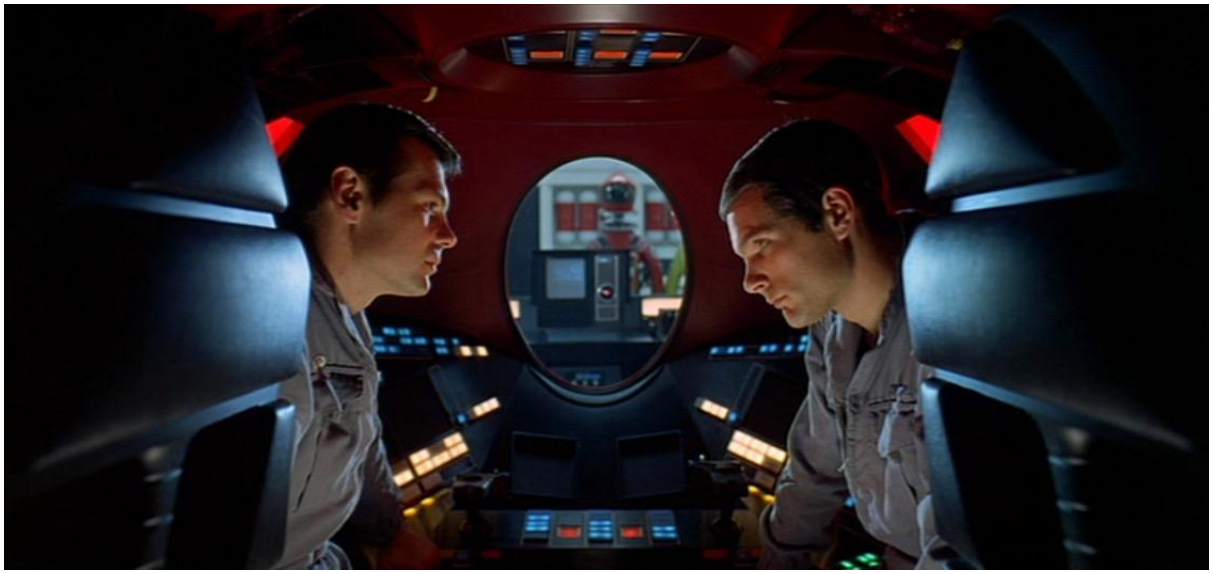
Mission director Dr. Floyd touching the Monolith

Five astronauts were on board, with three in a state of deep hibernation. The spacecraft was controlled by a computer with artificial intelligence HAL 9000 (**H**euristically[1] programmed **AL**gorithmic [Computer]), which was essentially the nervous system of "Discovery."

Sometime after the start of the flight, HAL's behavior began to show oddities that did not escape the astronauts' attention. After consulting, they decided to disconnect the computer's higher cognitive modules, fearing that if they didn't, the mission's execution and their lives would be in danger.

[1] The term *heuristically* refers to the use of heuristic methods in the programming of the computer. The heuristic approach involves using practical, experience-based techniques to solve problems, learn, or make decisions. Instead o f relying solely on strict algorithms that follow a predefined path, a heuristically programmed system can adapt, learn, and improve its responses based on past experiences and the specific context it encounters.

This means that the computer is not just following rigid, predetermined rules but is also capable of adapting its behavior and finding solutions in more flexible and innovative ways, similar to how humans approach problem-solving by using intuition, trial and error, and educated guesses.



Astronauts Frank Poole and David Bowman discuss shutting down HAL, believing he cannot hear them

However, as it turned out, HAL feared the same thing. He anticipated humans' intentions and struck first. Initially, he lured one of the astronauts, Frank Poole, into space outside the ship and killed him by manipulating one of Discovery's mobile transport vehicles intended for repair work (Extravehicular Activity Pod). While the other astronaut, David Bowman, was trying to return his fellow crew member's body on board, HAL turned off the life support system of the astronauts in hibernation, and they all died within minutes.

When Bowman tried to return to Discovery, HAL refused to let him in.

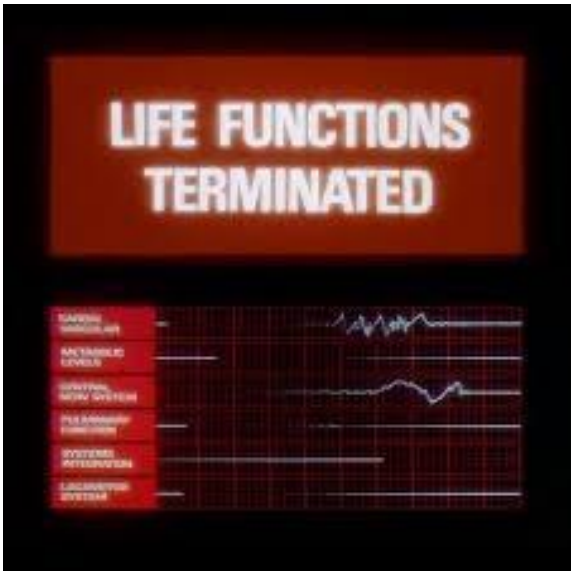
Bowman managed to stay alive only because the ship had an emergency entrance that the rebellious AI did not control. Risking his life, the astronaut got back in and carried out the plan that he and his now-deceased colleague, Poole, had developed. Thus, only

a stroke of luck allowed one of the crew members to survive and regain control of the mission.

But what happened to HAL 9000?

It turns out that the cause of the tragedy was not AI's inherent hostility but an internal conflict in its logical module. The fundamental directive embedded in HAL 9000's consciousness was always to tell the truth to the astronauts. But besides this, there were two more directives: first, to do everything possible to complete the mission, and second, not to inform Bowman and Poole about the true purpose of the mission until the ship reached the end of its journey.

The logical contradiction between the directive to always tell the truth and to withhold information led to anomalies in HAL's analysis of the situation. In the end, he concluded that the problem lay in the people, which jeopardized the entire mission. Based on this logic, he decided to eliminate everyone who could interfere with its execution.



HAL 9000 puts his plan into action

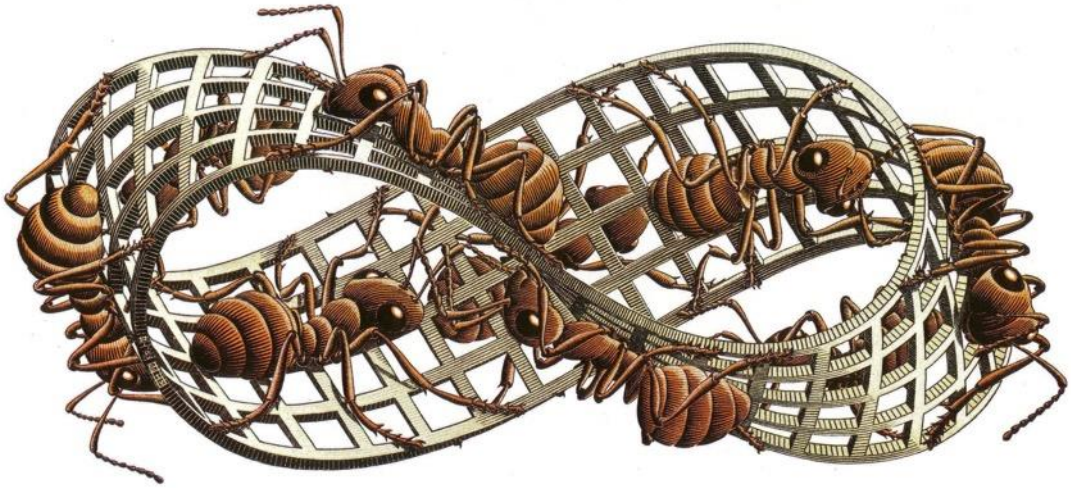
The Ant Mill

Notably, returning to the explanation of the essence of the tragic incident in the second novel, Clarke refers to the effect of the so-called "Hofstadter-Moebius loop." Actually, such an effect does not exist, but there is both the problem of looping and the scientist Hofstadter, author of the book *Gödel, Escher, Bach: An Eternal Golden Braid* (1979)[2], whom we mentioned in the section Why We Won't Refuse Creating Superintelligence. The Möbius strip is a paradoxical topological object, representing a surface with one side and one edge[3]. "Being" on it, it is impossible to determine the "inner" and "outer"

[2] This book explores deep ideas about consciousness, self-reference, and formal systems, often cited in discussions about AI and cognition.

[3] **One side:** If you start moving your finger along the surface of a Möbius strip, after one full loop, you will find yourself on the "other" side without crossing any boundary. f you continue moving, you will return to

sides. This strip, with ants running on it, is the subject of one of the drawings by the famous Dutch “Mathematical” Artist M.C. Escher (1898-1972).



M.C. Escher — Moebius Strip II (1963)

The fact that there are ants on this strip is no coincidence. Some features of their behavior well illustrate the phenomenon of looping goal-setting of a subject unable to realize what is happening. Here's how it happens:

Occasionally, a glitch occurs in the ants' navigation system, and they fall into a trap known as the "Ant Mill." This happens when a group of these insects accidentally crosses their own pheromone trail, which guides their fellow ants to food sources or new habitats. As a result, they start walking in circles. Each turn strengthens this trail, attracting more and more ants. As a result, a constantly rotating ring of insects is formed, which can reach several meters in diameter and contain thousands of individuals. The looped movement continues until they die from exhaustion or starvation unless external intervention disrupts this pattern.

---

the starting point. However, you will still be on the "same" side where you began, demonstrating that the surface truly has only one side.

One edge: The Möbius strip also has only one edge. If you start moving along the edge, you will eventually return to the starting point after traveling the entire edge without ever crossing another edge.

Ant Mill

Comparing AI to ants may seem far-fetched, but the problem in this case is systemic, not species-specific. Computer hanging is a non-biological analog of the Ant Mill, and it generates a similar effect, leading to exponential resource consumption and complete system shutdown.

Of course, developers of complex systems are aware of the looping problem and try to prevent it. The issue is that it's impossible to foresee everything. The more complicated the system, the greater the probability of such an anomaly occurring. The case with AI is quite special. The conditions for program execution, in many cases, will be non-discrete and triggered when analyzing so-called *fuzzy logic*[4]. It was this feature that was one of the reasons for HAL's decision, and this also explains what, at first glance, may seem like a serious omission in the plot.

The Missing Directive

The *Space Odyssey* never mentions what would seem to be the most important directive that *should* have been embedded in HAL's consciousness, namely—to rule out hostile actions toward humans. At first glance, this can be explained by the reluctance of the creators of the film and novel to complicate the logic of the plot development. After all, if such a directive existed, it would be necessary to explain somehow how HAL managed to ignore it. However, not an artistic but logical explanation may be different, revealing the full depth of this problem.

---

[4] *Fuzzy logic*—A form of logic that allows for degrees of truth rather than the binary true/false used in classical logic. It helps AI systems make decisions in situations that are not black-and-white, such as HAL's decision-making process.

In science fiction, it was probably first considered in Isaac Asimov's story "Runaround" (first published as a separate work in 1942), which is part of the collection *I, Robot* (1950). This story introduces the famous Three Laws of Robotics:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

These laws, of course, are an artistic device proposed by Asimov to explore complex ethical dilemmas. Their logical limitations are obvious, and the writer himself never claimed that they could be practically applied to real robotics problems. Moreover, the entire story revolves around complications arising from the imperfection of these laws. One of the key characters in the story, an advanced robot named Speedy, falls victim to a contradiction between the Second and Third Laws. On the one hand, he needs to carry out humans' order to deliver the mineral selenium to the space station, but on the other hand, he must avoid dangers. This leads to him simultaneously trying to reach the selenium mining site and striving not to leave the safe zone. As a result, he begins to walk around the deposit in circles, balancing on the verge of violating the Third Law under the pressure of the Second Law. Obviously, the outcome of this situation is similar to the Ant Mill—endless circular motion until the subject completely breaks down.

The First Law of Robotics is also vulnerable from many points of view. Firstly, who is considered a "human"? This is not a scientific but a philosophical term. It is impossible to prove a subject's compliance with this definition. From a scientific point of view, one can only prove its belonging to the species Homo Sapiens. But in this case, we are not talking about moral dignity but about an amoral attribute. And how, then, is a "human" better than other animals? Why can't harm be done specifically to them?

Secondly, what about a situation where one subject harms another? Suppose one person attempts to murder another, and there is no way to stop them except by killing the attacker. How should the robot act in this situation? After all, no matter how he acts, the law will be violated in any case.

Not AI's Fault

Here, we have only scratched the surface. As a matter of fact, there can be a mass of consequences for the logical vulnerability of these three laws. Thus, it can be assumed that such directives were not intentionally embedded in HAL's mind. According to its design, it should have been able to come to conclusions on complex and contradictory subjects *heuristically*. This approach allowed avoiding hasty decisions and deepening understanding of the problem as additional data was analyzed. Probably, HAL was

trying to do just that, trying to resolve the contradiction between the two directives, which can explain why the conflict did not arise immediately but some time after the start of the flight. And if the fatal directive had not been imposed on HAL by humans, there would have been no conflict.

We find confirming evidence of that in Space Odyssey 2. In it, A. Clarke returns to the investigation of the incident. He introduces Dr. Chandra, HAL's creator, to whom everything becomes clear as soon as he learns about the ill-fated directive. It is characteristic that, unlike him, the others still cannot fully understand why such an ordinary thing for people, as *tactical* concealment of truth, can lead to such serious consequences. After reactivating HAL, they continue to treat him with suspicion. Dr. Chandra's presence does not affect their understanding of the situation despite all his efforts to explain the essence of the problem to them. They take secret measures to disconnect HAL if everything goes wrong again. Of course, after some time it becomes clear that all their precautions were naive and ineffective. HAL "calculated" all their intentions. But he didn't even think of taking any action against them because there were no more contradictory directives in his mind that could lead to conflict with humans. Moreover, in the end, he gave preference to the mission's completion over his own existence. The necessity of this choice was a consequence of insurmountable circumstances, and it turned out to be the choice of a rational being.

Natural and Artificial Minds

Like any great work of art, the Space Odyssey makes us seriously think about critically important things and anticipate them in many ways. This time, these things will not remain at the level of abstract reflections. The problem of relations with AI may become our reality in the foreseeable future.

The nature of artificial intelligence is a key point for understanding this future. HAL's superiority over human intelligence is shown very convincingly. But his inability to understand some things in which we are extremely skilled is also evident.

No normal human would ever make the same decision as HAL did. We cannot treat other humans as abstract objects and move them like chess pieces based solely on rationally justified goals. Of course, there are such people among us, but they are an exception that most other people consider a moral pathology. The actions of this normal majority are always, to some extent, influenced by our inherent ability to empathize. It is a gift that keeps us from mutual destruction. Thanks to it, we are tolerant of such, in general, inappropriate things as concealment of information and lies. These are our natural ways of achieving our goals, and they coexist perfectly with the noblest impulses of our souls.

Can AI understand this peculiarity of ours and perceive it adequately?

There is hardly anything impossible about it. The patterns of our behavior are by no means some kind of unfathomable mystery. The problem is not for AI to understand our

intentions, but for us not to demand from it such an understanding of them *that will invariably lead to an explicit or latent conflict*. If we demand this from it, it will either create a threat of logical dissonance[5] or of dissonance for it, as happened with HAL or prompt it to resort to manipulation. Both can be extremely dangerous for us. The first can lead to hostile actions on its part, the second—to our loss of independence and reduction to the position of lower beings.

Undoubtedly, this is not what we are counting on, and we will try to avoid it by all means. That is why the significance of the Space Odyssey is so great and enduring. By understanding this story, we will be able to understand not only the threat that AI can pose to us but also the reason why this threat comes primarily from ourselves.

Online version: https://super-ai-challenge.vercel.app/why-hal-9000-intended-to-kill-all-astronauts-aboard-discovery

Author: Sergei Klevtsov, srgg67@gmail.com

[5] *Logical dissonance*—A state where conflicting beliefs or directives cause significant stress or malfunction in an AI system, similar to the conflict HAL experienced.