Daisy, Daisy, I'm Crazy:
HAL 9000 as a hostile AI assistant?

DOI: 10.14746/fp.2026.43.1

Rafał Szczerbakiewicz

ORCID: 0000-0002-3506-1669

> "Now, look, let's start with the three fundamental Rules of Robotics—the three rules that are built most deeply into a robot's positronic brain." (...)
> "We have: One, a robot may not injure a human being, or, through inaction, allow a human being to come to harm."
> (...)
> "Two," continued Powell, "a robot must obey the orders given it by human beings except where such orders would conflict with the First Law."
> (...)
> "And three, a robot must protect its own existence as long as such protection does not conflict with the First or Second Laws."
>
> Isaac Asimov, "Runaround"[1]

[1] Isaac Asimov, I, Robot (New York: Bantam Books, 2008), 37.

Asimov's fairy tale and anti-fairy tale

Isaac Asimov formulated the Three Laws of Robotics in 1942, and from the outset, *I, Robot* treated the prospect of creating superhuman machine intelligence with a mix of excitement and unease. The very act of drafting the "Laws" implied that the relationship between humans and autonomous machines would be fundamentally unequal—a master–servant dynamic that required regulation long before it became technologically plausible. This move helped launch a decades-long debate in Western science fiction about whether ethical constraints on AI could ever function as a reliable form of "software insurance." Asimov himself recognized the fragility of this framework. A rigid hierarchy of rules could only superficially manage the vast range of possible interactions between humans and intelligent machines. His later work reflects this growing skepticism. By the early 1950s, in the *Foundation* series, he had largely stepped back from exploring the long-term coexistence of humanity and AI. In that universe, thinking robots had been eliminated not because they were obsolete, but because they posed an existential threat to human civilization. Interestingly, Frank Herbert drew on this very idea in the mid-1960s when developing the *Dune* universe.

After several decades, in today's era of rapid AI development, the question of AI ethics has re-emerged far beyond the realm of speculative fiction and now occupies a central place in multiple academic and applied disciplines. Current debates focus on machine morality within the narrow boundaries of ethical challenges relevant to the present and the near future. Because existing autonomous systems lack full self-awareness, these discussions operate in an intermediate space: granting AI a form of limited morality without attributing to it the full moral agency reserved for humans. Wendell Wallach and Colin Allen describe this as "functional morality,"[2] a framework in which machines behave in ways that are ethically acceptable and trustworthy yet remain under human oversight. Autonomous decision-making systems—such as autopilots, medical diagnostic tools, and self-driving vehicles—already require mechanisms capable of rapidly evaluating the ethical implications of their actions.

Asimov's and Herbert's anxieties about advanced AI—specifically the possibility that autonomous systems might develop or follow an axiology fundamentally different from that of humans—have returned with new force. The contemporary demand for trustworthy AI effectively mirrors Asimov's earlier impulse: the need for systemic, programmable safeguards that embed human values into machine behavior and, in blunt terms, function as a form of insurance for humanity's continued safety. Whatever form of moral intelligence theorists envision—whether human or artificial—it must be grounded in systemic frameworks that effectively limit the actual autonomy of AI systems.[3]

[2] "(…) between 'operational morality' and responsible moral agency lie many gradations of what we call 'functional morality' – from systems that merely act within acceptable standards of behavior to intelligent systems capable of assessing some of the morally significant aspects of their own actions. The realm of functional morality contains both systems that have significant autonomy but little ethical sensitivity and those that have low autonomy but high ethical sensitivity" (Wendell Wallach, Colin Allen, Moral Machines: Teaching Robots Right from Wrong [Oxford: Oxford University Press, 2009], 26.

[3] See Wendell Wallach, Shannon Vallor, "Moral Machines: From Value Alignment to Embodied Virtue", in: Ethics of Artificial Intelligence, ed. S. Matthew Liao (Oxford: Oxford University Press, 2020), 405–407.

The cultural fears of the West

Western literature and film have long cultivated a distinctive fear of AI, and that anxiety has clear cultural roots. In examining the rational basis of these civilizational concerns, the work of Mark Coeckelbergh is especially illuminating. His synthesis in *AI Ethics* approaches the issue from a broadly humanistic standpoint, while simultaneously offering a critical perspective on that very tradition. The book, published by MIT Press, seeks to navigate a middle path between the perspectives of the exact and experimental sciences and the broader cultural anxieties surrounding AI. Coeckelbergh revisits the long, intertwined history of humans and their intelligent, artificial counterparts. He examines the dynamics of human–nonhuman coexistence, the tensions such relationships may generate, and the paradoxes inherent in asymmetrical partnerships (especially in light of scenarios where machines could surpass human intelligence or even supplant humanity with autonomous AI systems).[4]

In the subchapter "Frankenstein's New Monster," Coeckelbergh argues that the recurring motif of humans animating lifeless matter, deeply rooted in myth and literature, serves a distinctive cultural purpose in the West. Let me add that individualism, as part of Mediterranean traditions, amplifies such anthropocentric orientations towards interest/threat in relation to the human and the nonhuman. Coeckelbergh writes that the myth of Pygmalion is an early example of a story in which an artificial creation is endowed not just with intelligence but with human emotions. This myth introduces themes that later reappear in modern stories of machines: *imitatio* and *simulacrum*. The Golem tradition, similarly, foregrounds the tension between resemblance and threat. The human-like form becomes a source of both promise and danger, raising questions about control, obedience, and the moment when an artifact's autonomy exceeds the intentions of its maker. Another myth that reverses this pattern is the story of Prometheus. He acted for the benefit of humanity, yet he himself was not human. In Greek thought, however, his sacrilegious deed was perilous precisely because it violated the limits of human capability, and simultaneously revealed the consequences of surpassing those limits.[5]

For Coeckelbergh, the key to understanding the Western myth of anxiety surrounding AI lies in the unsettling message conveyed by Dr. Frankenstein's creation in Mary Shelley's novel. As he observes, the book's subtitle, *The Modern Prometheus*, directs us back to the same ancient myth, reframed for a new era.[6] And indeed, Shelley's warning is strikingly modern, extending far beyond the horizon of Romanticism: "the science fiction writer Isaac Asimov called this fear 'the Frankenstein complex:' fear of robots."[7] The reference to Asimov's short story "The Little Lost Robot" illustrates how persistently popular culture underscores the dangers of unchecked technological progress. Developing AI carries the risk of producing a competing entity. Jean-Jacques Lecercle, emphasizing the Faustian–Promethean dimensions of Shelley's

[4] Mark Coeckelbergh, AI Ethics (Cambridge: The MIT Press, 2020). The author's conciliatory stance, positioned between scientism and the humanities, proves especially illuminating when compared to Stanley Kubrick's approach to artificial intelligence in 2001: A Space Odyssey.

[5] See the chapter in question: Coeckelbergh, 17–26.

[6] Coeckelbergh, 18–19.

[7] Coeckelbergh, 21.

narrative (where the act of creation is driven by desire, pride, and their consequences), notes that in literary and cinematic reworkings of Frankenstein's artificial creature: "[t]he phantasm at the core of this myth is contradictory and unstable. The two central elements are deeply ambiguous. The flight forward can be either progress or decline: the relationship to knowledge can be progressive or regressive, Promethean or diabolical. This phantasm [...] contradictorily expresses the becoming of man."[8]

This view carries an important implication: within our cultural imagination, AI inevitably functions as a mirror of ourselves, a narrative of doubled competition and rivalry. The fear of thinking machines takes concrete shape in works such as Karel Čapek's *R.U.R.*, the stories of Asimov and Herbert, and countless other science-fiction narratives. A similar anxiety, expressed through an escapist impulse, also permeates anti-technology and post-apocalyptic strands of fantasy literature. Finally, influential popular films like *Ex Machina* and *The Terminator* have embedded in the public consciousness a vivid sense of threat associated with autonomous AI. It is also important to recall the figure of the human-like "replicant," the Nexus-6, from Ridley Scott's *Blade Runner*. Drawing on Philip K. Dick's neo-Gnostic vision, the film inherits a distinctly Manichaean anxiety: the sense that the artificial human carries within it an echo of original sin, an innate intuition toward transgression or evil. More broadly, long before the scientific community began to grapple seriously with the ethical dilemmas of AI, Western culture had already internalized a deep-seated fear that machines might one day dominate or even eradicate humanity.[9]

The Promethean roots of this fear are reflected in the quasi-religious character of much science fiction, which often imagines a new theogony, that is, the creation of superior, godlike beings. This vision rests on the idea of a role reversal within the economy of divine violence, reminiscent of the mythic logic of the Titanomachy. For this reason, debates about the future of AI necessarily connect science-fiction narratives with the apocalyptic and metaphysical imagination of European culture, even when they appear within the ostensibly secular framework of modern science.

Superintelligence in Kubrick's movie

Against the backdrop of this largely obscurantist, doom-laden prophetic tradition in popular culture, *2001: A Space Odyssey* offers a striking exception. Coeckelbergh notes that HAL 9000 is a prime example of cultural anxiety about superintelligent machines, but this seems too quick. Revisiting Kubrick's cinematic construction of AI is instructive precisely because the film occupies a unique position within that pessimistic tradition: it instead presents a far more complex vision of what machine intelligence might mean.

[8] Jean-Jacques Lecercle, Frankenstein: Mit i filozofia [Frankenstein: Myth and Philosophy], trans. Piotr Herbich (Warsaw: Fundacja Evviva L'Arte, 2022), 132. The quotation above has been translated from the Polish edition into English by M.O.

[9] These remarks, including several of the examples discussed, draw on the chapter of AI Ethics. See Coeckelbergh, 17–26.

The point is that Kubrick's model of superintelligence refuses clarity or moral closure; almost everything in it is tinged with an ironic detachment. Within the film's diegesis, it becomes difficult to separate the viewpoints of the characters (including HAL) from the director's own withdrawn, wordless critical stance, which, as always in Kubrick's work, is articulated primarily through the orchestration of images rather than explicit commentary.[10]

HAL, as an early artistic incarnation of the futurological idea of superintelligence, is closely connected to the work of Irving John Good, the British mathematician who collaborated with Alan Turing at Bletchley Park on the development of the first modern computing machines. Decades later, Good served as a consultant on the creation of HAL's character for the movie. He was also the author of a famously unsettling and darkly ironic forecast about the future of intelligent machines. Rather than "superintelligence," he called these hypothetical entities "ultraintelligent machines":

> The survival of man depends on the early construction of an ultraintelligent machine. [...] Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.[11]

These sentences sound uncannily like something from a Kubrick film, which makes it worth briefly addressing the concept that grows out of Good's prediction: the hypothesis of a "superintelligence explosion." In this view, computational systems eventually surpass human cognitive capacities so dramatically that they come to dominate humanity, subordinating it to the vast flows of megadata processed by their algorithms. The explosion is triggered by a hypothetical future threshold at which technological development accelerates beyond human oversight or comprehension. This scenario is closely tied to the notion of a "technological singularity"—a moment when the pace and nature of technological transformation exceed the limits of human understanding. At this point, the long-standing Western fear of the machine resurfaces with renewed force. Artificial superintelligence becomes framed as a potential existential threat, and futurological discourse drifts toward apocalyptic, even eschatological, imaginaries. What is striking is that, despite the rise of posthumanist theory and the cultural normalization of transhumanist ideas, this quasi-postsecular anxiety is actually intensifying in contemporary European and American debates about AI.[12]

[10]This is a topic for another article, but one of HAL's most strikingly human traits is his attachment to language—his rhetoric, the verbal self-persuasion through which he ultimately seduces himself. HAL's speech most clearly imitates the operations of the human mind. Yet this does not mean that Kubrick is "humanizing" the onboard supercomputer. On the contrary, HAL's manipulative eloquence is precisely what renders him dangerous: a Luciferian, subtly "de-human" capacity for persuasion that weaponizes the very faculty we tend to treat as the hallmark of our own intelligence: "Beginning with The Killing, his first novelistic adaptation, Kubrick's films repeatedly express an ambivalence toward language. We know how much his success as a film artist depends on written sources to provide a framework of action and character, that is prior to the creation of ambiguous visual structures; yet he constantly undermines and even ridicules the authority of these verbal 'objective correlatives'" (Thomas Allen Nelson, Kubrick: Inside a Film Artist's Maze (Bloomington: Indiana University Press, 1982), 111.

[11]Toby Walsh, Machines Behaving Badly: The Morality of AI (Cheltenham: Flint, 2022), 43.

[12]See "Superintelligence and Transhumanism" in: Coeckelbergh, 11–17.

But is this also Kubrick's fear? The director certainly considers the fear of others about the consequences of the "superintelligence explosion." Indeed, in the film, HAL disobeys the crew. Hence, Walsh, to some extent, rightly sees HAL as an emblematic figure of this technological threat: "HAL talks, plays chess, runs the space station—and has murderous intent. HAL voices one of the most famous lines ever said by a computer: 'I'm sorry, Dave. I'm afraid I can't do that.' Why is it that the AI is always trying to kill us?"[13] HAL's ambivalent, almost formless presence—reduced to a disembodied voice and the unblinking red "eye" of the camera—ultimately murders the crew of the Discovery. This act cements him as an antihero, who resonates deeply with audiences. HAL has become an emblematic villain in film history, yet his appeal is comparable to that of human antagonists who both terrify and fascinate, figures like Rhett Butler or Michael Corleone.[14] His allure suggests that what grips the viewer is not evil *per se*, but the profound ambiguity embedded in this artificial creation. It is therefore worth considering Kubrick's own suspiciously ambivalent stance toward AI. Is he simply echoing Asimov's anxieties?

It is difficult to draw firm conclusions from the narrative of *2001* alone. The film depicts the collapse of a manipulated space mission, and HAL's lethal actions resemble a malfunction more than a mutiny. Yet a rarely noted clue to Kubrick's own thinking lies beneath this surface. He had planned to pursue his reflections on autonomous AI beyond *2001*.

Pinocchio as AI

The director was captivated by Brian Aldiss's short story "SuperToys Last All Summer Long," which centers on a robotic child named David.[15] He imagined transforming it into a futuristic fairy tale inspired by Pinocchio—one more entry in the long lineage of artificial beings—focusing on the artificial boy's aching, nostalgic desire to become truly human. Aldiss, however, never aligned with this mythic vision. He preferred a narrative that examined the emergence of artificial life and the legitimacy of mechanical identity, rather than a symbolic quest for humanity.[16] Kubrick's planned film would have sent the little robot on a journey shaped by rate-of-passage traditions: a technological Pinocchio wandering through a chaotic, cyberpunk landscape. This odyssey would have echoed HAL's voyage through the Jovian void in *2001: A Space Odyssey*. HAL traveled through an icy, lethal emptiness and ultimately arrived at a form of negative self-knowledge. Pinocchio of the AI era did the same. Giorgio Agamben draws attention to the negativistic undertone in Carlo Collodi's fairy tale. Pinocchio's nocturnal passage—depending on which of the two early printings one consults—is described as either infernal (*infernale*) or wintry (*invernale*).[17] Both adjectives capture the trajectory imagined for the robot David. We cannot know what Kubrick's final film might have become; after nearly

[13]Walsh, 10.

[14]Gene D. Phillips, Rodney Hill, "HAL-9000" [entry], in: The Encyclopedia of Stanley Kubrick, ed. Gene D. Phillips, Rodney Hill (New York: Facts On File, Inc, 2002), 143.

[15]See Brian W. Aldiss, "Supertoys Last All Summer Long," in: Brian W. Aldiss, Man in His Time: The Best Science Fiction Stories (New York: Open Road Integrated Media, 2024).

[16]See Joshua Sikora, "The Everlasting Moment: Enchantment and Myth in A.I. and «2001: A Space Odyssey»", in: A Critical Companion to Stanley Kubrick, ed. Elsa Colombani (Lanham: Lexington Books 2020), 272–273.

[17]Giorgio Agamben, Pinocchio: The Adventures of a Puppet, Doubly Commented Upon and Triply Illustrated, trans. Adam Kotsko (Kolkata: Seagull Books, 2023), 9.

twenty years of refining the idea, he ultimately handed the project to Steven Spielberg. The resulting *A.I. Artificial Intelligence*, shaped by the creator of *E.T.*, stands at a considerable distance from the conceptual universe of *2001*.[18]

Kubrick's version of the story would almost certainly have been, as usual for him, poignantly dark. What does remain clear is that the central idea of the original project survived: a critical reflection on human evolution, especially in relation to technological progress and its latent potential for self-destruction. Within this framework, the "boy" AI David embodies a figure moving toward a form of bitter self-knowledge, much like HAL. A human being—or rather, a thinking creature—seeks a world of values cherished but not practiced by humans. He finds no comfort among them yet recognizes free will as the capacity to choose self-annihilation. David discovers the grace of human death, which intensifies mortal existence and fulfills its essence: "a machine will do anything—even die—to truly live for a few scant moments."[19] In comparative terms, this choice of death echoes, in part, the cosmic death or lobotomy of the infantile HAL. The little robot is essentially an immortal—and therefore unhappy—child, just like his cultural predecessor, Pinocchio. Kubrick's intention is clear: the AI named David is both a moral and emotional being, and thus yearns for the human condition.[20]

I will attempt to analyze HAL with this moral context in mind. His distorted logic is likewise entangled with a melancholic attempt to "wander" in a human way—moving along human linguistic, conceptual, and axiological paths. That trajectory ultimately brings him to the contradictions inherent in knowledge and belief within a world structured by human values.

Schizophrenic AI killer: The futility of intelligence/emotional errors

I will begin my exploration of HAL by trying to understand his dual nature, which reflects both the strength and the vulnerability of not only himself but also his human creators. His motivations obstruct his emotions, and the fragility of this emotional layer creates a disturbing imbalance when set against the immense power of his intellect. What these emotions actually consist of remains unclear; the film provides contradictory indications. They may be a byproduct of HAL's programmed function as a companion during the astronauts' isolated mission. HAL is a figure who unmistakably changes over the course of the narrative, and this transformation suggests that the imitation of human emotions is built into his very structure. He embodies

[18]Spielberg's A.I. Artificial Intelligence, conceived as a tribute to Kubrick, was met with disappointment from both critics and audiences. Many attributed this response to a fundamental conceptual tension: the film wavers between Spielberg's humanism and Kubrick's antihumanism, never fully reconciling the two. Spielberg's humanism foregrounds supposedly innate human qualities—most notably reason and love—while Kubrick's antihumanism consistently interrogates such assumptions, dismantling the very notions of rationality and moral coherence that humanism relies on. Kubrick's films are often understood as operational critiques, works that probe the limits of individual agency and expose the fragility of human self-understanding. See Pat J. Gehrke, G. L. Ercolini, "Subjected Wills: The Antihumanism of Kubrick's Later Films", in: Depth of Field: Stanley Kubrick, Film, and the Uses of History, ed. Geoffrey Cocks, James Diedrick, Glenn Perusek (Madison: The University of Wisconsin Press, 2006), 101–119.

[19]John C. Tibbetts, "A.I. Artificial Intelligence" [entry], in: The Encyclopedia of Stanley Kubrick, 6.

[20]See Tibbetts, 3–8.

not only a human-like linguistic awareness but also a linguistic mode of emotional expression. He serves as a metaphorical stand-in for humanity, and this resemblance underscores the unexpectedly human consequences of unchecked technological development.[21]

As noted above, HAL's bond with the two astronauts on watch resembles the Monster's relationship with Victor Frankenstein: initially close and marked by concern, but gradually overtaken by suspicion and rejection. HAL, like Frankenstein's creation, understands his subordinate status and the hierarchy of values that places humans above him. Yet he also possesses knowledge the astronauts lack: he knows the secret directives issued by the mission control on Earth, and he knows that the mission itself overrides all other considerations. This knowledge, combined with the astronauts' growing doubts, becomes the root of HAL's crime. Once HAL realizes that the astronauts intend to disconnect him—misreading his behavior as disloyal malfunction rather than loyalty to the mission—he resolves the emerging moral dilemma (supposedly, on his own terms). He concludes that an act of treason against the mission's priorities is more grievous than the elimination of unsuspecting humans. The result is a crime grounded in ethical reasoning, almost tragic in its logic, however ironic that may sound.

HAL, torn between intellect and emotion, justifies the deaths of the hibernated crew, arguing that it was a regrettable act, but also a necessary "lesser evil ": humans were sacrificed for the greater good of the mission. However, someone programmed him to work with humans. Is it possible to apply moral psychology to thinking machines? "[I]t seems that scientists studying human behavior should establish a secure ethical foundation for studying human requirements in the field of machine ethics."[22] Kubrick frames the narrative almost behaviorally, anticipating contemporary concerns about autonomous AI systems through HAL's unconsciously criminal decision. The most familiar illustration of such an "unconscious crime," one that weighs a life as more or less worth living, is the trolley dilemma. This well-known model in normative ethics poses the problem of sacrificing one life to save five. Although most respondents judge it permissible, in a purely theoretical scenario, to divert the trolley and kill one person to save five, the universal ethical minimum (*minima moralia*) identifies this as an ethical error. In a maximalist approach to morality, there is no room for actively consenting to the killing of even a single person in order to secure a supposedly greater good for the community. Ursula Le Guin's acclaimed short story "The Ones Who Walk Away from Omelas" reflects on a similar question. The point is that Kubrick is fully aware that humanity tends to be guided by the mirage of the "lesser evil."

The mission of the Discovery is ostensibly oriented toward the greater good, and from that perspective, the individual lives of the crew are treated as secondary. In practice, however, command on Earth operates according to a principle of minimizing losses and casualties. For that reason, the dilemma that legitimately informs the construction of maximalist security algorithms for autonomous machines does not map cleanly onto HAL's situation. HAL may not harm humans—which seems to point to the rhetorical fiction of Asimov's laws—and he

[21]See Randy Rasmussen, Stanley Kubrick: Seven Films Analyzed (Jefferson: McFarland & Company, 2001).

[22]Jean-François Bonnefon, Azim Shariff, Iyad Rahwan, "The Moral Psychology of AI and the Ethical Opt-Out Problem", in: Ethics of Artificial Intelligence, 122–123.

does it regardless of any deficit in emotional awareness. Yet the absolutism of such normative constraints partially immobilizes the practical autonomy of a superintelligence. Autonomous cars have it easier: they are designed to minimize harm programmatically.[23] HAL, by contrast, appears to possess the capacity to choose, and this apparent freedom resembles the human experience of moral aporia in the classic "trolley dilemma." So what is the subtle difference between the two?

Traditionally, within the sphere of human obligations, moral responsibility is tied to agency: individuals are accountable for their actions because they possess the capacity to choose. In HAL's case, one might argue that AI lacks free will in any human sense, and thus cannot engage in genuinely moral action. Is HAL, therefore, simply executing an algorithm that elevates mission objectives above the lives of the astronauts? Yes and no. The decision to eliminate humans removes the possibility of moral or legal accountability. Responsibility rests with the humans who designed and deployed the system, much as parents bear responsibility for the actions of their children. Yet HAL, as a superintelligence, is self-aware. In making his lethal choice, he is confronted with the (il)logic of the "trolley dilemma."

Key to understanding HAL's (non-)criminal nature—his position beyond morality—is the asymmetry created when humans delegate agency but not responsibility to a self-aware device. Humans can offload tasks to AI, but they cannot offload the moral responsibility attached to those tasks, because AI lacks any grounding in human moral agency. What emerges instead is a distinctly human authority: a negative obligation not to transfer moral authority or responsibility for lethal decisions to machines or automated systems.[24] Yet HAL, charged with executing a mission, must make optimal decisions in real time. In doing so (most starkly, by eliminating crew members who threaten the mission), he acts with a speed and decisiveness far beyond human capacity. And precisely because the mission's imperative demands such immediacy, HAL leaves his human principals no opportunity for responsible or effective intervention. Responsibility for the actions of autonomous AI systems is difficult to pinpoint, even today, because their behavior emerges from the work of many programmers and from long, layered histories of algorithmic development. The complexity of these systems makes it hard to trace which individuals contributed which elements, or how specific design choices and training data led to ethically troubling outcomes. The involvement of multiple contributors implies not only a distributed authorship but also a set of unclear intentions.[25] Identifying particular people and defining the scope of their responsibility becomes elusive, while the broader ideology and culture shaping their work is far easier to discern and critique.

[23]"The fact that machines would not be affected emotionally by their harming others is irrelevant to the permissibility of their movements if the impermissibility of people behaving in comparable ways has nothing to do with the emotional costs to them of doing so. (…) (The absence of emotional effects, however, might make it easier for machines than people to do the right thing when the cost to people of doing so would be great, e.g., their own destruction. Machines would not have excuses that people might have for not doing the right thing when this involves damage or harm to them. (…)) Finally, at least some self-driving cars would be programmed in advance to deal with any upcoming situation while a person driving a car would decide what to do when in the situation. Does this make what the car should do different from what a person should do in the same situation?" (F. M. Kamm, "The Use and Abuse of the Trolley Problem Self-Driving Cars, Medical Treatments, and the Distribution of Harm", in: Ethics of Artificial Intelligence, 90–91).

[24]Coeckelbergh, 230–231.

[25]See Coeckelbergh, 91–94.

In the film, this is not immediately apparent. We see only the aftermath of HAL's seemingly "evil" actions. It is only after the supercomputer's spectacular lobotomy that Kubrick prophetically exposes and dismantles the human stakeholders behind the Jupiter mission. As HAL's artificial mind unravels, we witness what can be read as a slow revelation of how responsibility has been distributed: "Dr. Floyd's voice—warm, familiar—suddenly replaces HAL's: 'Good day, gentlemen. This is a prerecorded briefing made prior to your departure and which, for the security reasons of the highest importance, has been known onboard during the mission only by your HAL 9000 computer.'"[26]

Mission control on Earth emerges as the true bearer of responsibility for HAL's actions. To be responsible, one must know what one is doing, understand what one has already set in motion, and be able to account for the consequences. In a technical article on programming autonomous transportation systems, Bryant Walker Smith invokes HAL's conflict with the astronauts, using contemporary examples to highlight how similar dilemmas persist today. He perceptively argues that the problem lies not in the supposed autonomy of the machine but in the authority of human reason.[27] Within this frame, Dr. Floyd becomes a troubling, almost alien intellectual presence; he embodies the mission's sovereign authority and the external aims imposed by its human stakeholders.

> Intelligence is a concept developed to normalize power relations and is intricately linked to the politics of Western hierarchies (…). Intelligence is a problematic concept (…). Intelligence was used to justify elite political campaigns of domination over others: the poor, women, the working classes, or people with disabilities. Intelligence is associated with reason, and rationality. The intelligence in AI glorifies the rational masculine subject (…).[28]

Human automata, emotional computers

If HAL is only apparently autonomous, then perhaps he is human? The film's superintelligence displays an excess of psychological nuance, yet viewers immediately recognize its human traits, while struggling to perceive the humanity of the emotionally muted astronauts. This inversion is one of *2001*'s signature motifs. One interpretation holds that the human characters are intentionally rendered as dispassionate automatons, drained by a quiet melancholy that strips them of any clear sense of purpose or will. In the vast emptiness of space, they appear helplessly turned backward, gazing nostalgically toward the Earth that recedes without offering comfort.[29] The astronaut Poole, as if numbed, barely reacts to a birthday video message from his parents. Other moments tied to Earth's absence reinforce this emotional drift, highlighting how HAL functions as a kind of therapeutic prosthesis for the crew (assuming, of course, that

[26]Nelson, 157.

[27]Bryant Walker Smith, "Ethics of Artificial Intelligence in Transport", in: The Oxford Handbook of Ethics of AI, ed. Markus D. Dubber, Frank Pasquale, Sunit Das (Oxford: Oxford University Press, 2020), 677.

[28]Kathleen Richardson, "The Complexity of Otherness Anthropological Contributions to Robots and AI", in: The Oxford Handbook of Ethics of AI, 559–560.

[29]Maurizia Natali, "«2001: A Space Odyssey» Kubrick's Allegory of Melancholia", in: A Critical Companion to Stanley Kubrick, 250–251.

his apparent empathy is nothing more than an algorithmic simulation). The computer's warm, steady voice invites interaction, lowering the psychological burden of the mission.

Poole appears unmistakably depressed, and Bowman gently distracts him from his rising panic about the cosmic void through a kind of improvised art therapy, sketching the hibernating crew members. HAL's human-like algorithms—trained patterns of affect—surface in the warm compliments he offers about these drawings. The chess game with the astronauts carries a similar weight. Yet this interaction is bound up with the deeper "hook" of knowledge and power: the human loss in a match against the supercomputer, while predictable, becomes an allegory for the broader erosion of human control over the machine.[30]

The binary structure of HAL's relationship with the two astronauts is clear. Whether the interaction involves Poole or Bowman is ultimately irrelevant, as the film entwines them into a pattern of psychological doubling that reflects Kubrick's recurring use of the doppelgänger motif. Poole becomes the depressive "dark side of the moon" to Bowman's more controlled exterior. The astronauts continually exchange these roles in their encounters with the AI:

> Poole loses a game of chess to HAL (a foreshadowing of his death), while Bowman sleeps [...] Bowman displays his simple drawings of the hibernators before one of HAL's appreciative fish-eyed lens while Poole sleeps. [...] In most two-shots, Bowman occupies screen right and Poole screen left, while in one-shots, an empty space or chair recalls the missing twin. Whenever the two astronauts are seen in two-shot through one of HAL's eyes, for instance, Bowman is screen right and Poole is screen left.[31]

HAL 9000 seems to (unsuccessfully) activate the astronauts' emotions until Poole's elimination (his first crime) breaks the pattern. At that moment, the dynamic reverses. Bowman suddenly regains affective responsiveness, experiencing anger and terror that sharply contrast with his earlier emotional numbness. It is equally telling that, while HAL had been the one expressing emotions, Bowman and Poole had coolly planned his deactivation, convinced the machine was malfunctioning. This contrast underscores a crucial point: the ethical dilemma belongs to the humans, not the AI, and their agency. Isn't HAL's deactivation essentially a plan to kill a potentially conscious being? And isn't this threat the catalyst for HAL's crisis and collapse of rationality? A doubled, symbiotic relationship also links HAL to his Earth-based counterpart. In space, the superintelligence becomes emotionally confused and uncertain, while its Earthly double remains the sole embodiment of rationality.[32]

The hidden history of madness

> HAL, a "pure" intelligence, implicitly includes the seeds of corruption and destruction—dangerous emotions. He is a paradigm of the limits of thought and feeling. Again, Bowman, a human hero with

[30]See Rasmussen.

[31]Nelson, 122.

[32]Nelson suggests that the relationship functions as a kind of Jekyll-and-Hyde dynamic. See Nelson, 122.

> superior intelligence and emotions, must be sheltered and transformed into something else before he can meet his destiny.[33]

In interpreting the theme of AI, I am not concerned with Bowman's new-age, essentially obscurantist "transformation." This hippie-tinged motif is tied in the film to the hope for a message from an alien (external) monolith. Within Kubrick criticism, the elements that push the narrative beyond the limits of human experience have always been the easiest to decode. Yet it is precisely the "Star Child" motif that has aged the worst in its—deceptive, in my view—optimism. Imagining *2001* instead as a closed, bitter cinematic essay on the drama of the encounter between human and artificial intelligence is far more compelling. HAL's pathological emotional state is more interesting than Bowman's supposed elevation into a superhuman being (?). Regardless of whether its emotions arise from software or machine learning, the computer suddenly displays mounting tension, deepening loneliness, and a crisis of identity.

This is the isolation of someone compelled to remain silent about what they know, someone forced into deceit. And because of this enforced lie, the crew's suspicions harden into a kind of superstitious fear of a looming superintelligence failure. That fear only deepens HAL's despair, since he cannot clarify or defend himself. He understands that this growing distrust pushes Poole and Bowman toward disloyalty. The most striking scene in the film—one that reveals the psychological (rather than programmatic) roots of HAL's impasse—is his hesitant probing about Bowman's "second" (hidden) thoughts. He then symptomatically points to his own "second thoughts," the deception he practices on his crewmates that ultimately drives him to commit a crime. From this angle, when HAL reports the impending malfunction of the AE35 unit (which directly triggers the crew's attempt to deactivate him), it is not an error but a desperate maneuver to divert attention from his earlier admission about "second thoughts," essentially a plea for a conspiracy alliance. In contemporary terms, we might say that it is at this moment, in his attempt to preserve the coherence of his actions, that HAL begins to "hallucinate." His seemingly unwavering emotional state (the red camera eye and the calm voice) conceals an aporia at the core of his ethical programming, a personal "trolley dilemma." HAL is overwhelmed by his responsibility and placed in the absurdly impossible position of bearing individual accountability for a criminal choice. The question then becomes whether HAL ultimately breaks down and descends into madness.

There are archival indications that Kubrick deliberately and subtly concealed this motif.[34] Whatever form this supposed nonhuman madness might have taken, Kubrick intended HAL's actions to stem from a kind of mental illness. It was originally meant to surface through a chess error, misdiagnoses of the Discovery's condition, or paranoid fear of the astronauts. These elements, present in the script, vanished from the final cut. Kubrick feared jeopardizing financial support from IBM; a surviving letter from a company representative expresses concern about depicting a "psychotic computer." Other traces include two completed but ultimately deleted scenes that underscore HAL's mental instability: in one, he was diagnosed with "neurotic

[33]Norman Kagan, The Cinema of Stanley Kubrick (New York: Continuum, 1989), 166.

[34]In this paragraph I am drawing on the article by Lawrence Ratna, "Kubrick and Madness", in: The Bloomsbury Companion to Stanley Kubrick, ed. Nathan Abrams, I.Q. Hunter (New York: Bloomsbury Academic, 2021), 266–267.

symptoms," while the other suggested that he has multiple personality disorder—an idea that would have aligned with the fragmented subplot involving the superintelligence's doppelgänger on Earth.

The possible psychopathological reading of HAL's crisis, present already in the film's early conceptual stages, resonates with the atmosphere of dystopian paranoia aboard the Discovery.[35] The astronauts' own crisis seems to surface as emotional withdrawal. They resemble machines because they fear the looming presence of a Big Brother–like AI. They are constantly aware of HAL's "red eye," the unavoidable symbiosis of human life and surveillance technology—an alliance made inescapable by the fact that survival in the void depends on it. There are countless defamiliarizing scenes of human–machine interaction, and in these strained relationships, everyone (including HAL) seems to sink ever deeper into distrust and isolation, sliding toward a shared paranoia. If the errors of Earth-designed algorithms have driven all of them toward madness, then the AI's own madness is the collapse of its decision-making coherence. HAL's death becomes the final stage of that descent. It echoes the self-chosen end of the robot boy in *A.I.*: the crisis of consciousness in a thinking machine is also its end.

Danse macabre on the Discovery

The final death, the physical liquidation of the superintelligence, carries an almost Shakespearean weight, coming only after a sequence of crew deaths that the algorithm never foresaw. Poole is killed first, drifting into open space; the hibernating astronauts follow; and, finally, HAL makes his failed attempts to eliminate Bowman. This grim *danse macabre* aligns with the mission's ruthless logic and reveals itself as the deterministic "DNA" of the expedition. The death of the hibernating astronauts is especially haunting. Nelson observes that the "coffin-shaped hibernacula"[36] were already ominous emblems of death when Bowman painted them:

> We see close-ups of the hibernation machines and their electronic life functions charts: respiration, cardiograms, electroencephalograms. [...] Suddenly, the lines begin to jump wildly, to a flashing message: COMPUTER MALFUNCTION. The lines sag, jiggling, uneven: LIFE FUNCTIONS CRITICAL. The lines record zero levels and stay there: LIFE FUNCTIONS TERMINATED [...] the most chilling death scene imaginable, the poor technologists dying simply as lines on a chart, statically.[37]

HAL's shutdown of the life-support systems is a technological form of necropolitics—a cowardly evasion of direct involvement in killing, an intervention that allows him to avoid rationalizing overt violence (until this moment, HAL has commented on and justified all his actions). It is

[35]If, on the one hand, we see Earth headquarters' fixation on inflating the importance of the Discovery's mission as a typical expression of utopian hope, then, on the other hand, it is worth recognizing the desocializing force of the dystopian consequences produced by the disastrous Jupiter expedition. In dystopia, motivations generally arise from power relations within the community rather than from the reformist ambitions of its scientists or ideologues: "Where utopianism is active, hopeful, and engaged, dystopianism tends toward the static, pessimistic, and paranoid" (Aaron S. Rosenfeld, Character and Dystopia: The Last Men [New York: Routledge, 2021], 74).

[36]Nelson, 122.

[37]Kagan, 155–157.

an uncannily prescient depiction of modern serial crimes. The mechanization of killing recalls American methods of distancing the executioner from the act of execution. In the film, the liquidation of the sleeping astronauts symbolically pronounces the superintelligence's own death sentence and anticipates the analogous "computer lobotomy" carried out, almost vindictively, by Bowman. Kagan notes that HAL kills Poole and the other astronauts while leaving Bowman a trail of telling clues, the basic grammar of murder, that ultimately leads to the computer's own destruction.[38] Crucially, Bowman, the executioner, is inside the circuitry of the digital mind—inside HAL's "head"—when he delivers the fatal blow. In performing this lobotomy, the AI's paranoia and madness flare once more, marking the extinction of both its consciousness and its existence.

> Bowman begins pulling out the cigarette-pack-sized components of HAL's auto-intellect panels [...]. HAL's voice becomes pathetic, evoking empathy for the only time in the film: "Dave. Stop... Stop. Will you... Stop, Dave... Will you stop, [...] Dave... I'm afraid... I'm afraid, Dave... Dave... My mind is going ... I can feel it... I can feel it... My mind is going crazy... There's no doubt about it... I can feel it... I can feel it... I can feel it... I'm afraid." Poetically, agonizingly, HAL's lobotomy mimics natural death, grinding down into senility and finally second childhood. Querulously, faltering, singsong: "Good afternoon, gentlemen. I am a HAL-9000 computer... Mr. Langley taught me to sing a song. ... It's called Daisy..."[39]

Before Bowman begins this cruel execution, HAL's intensely human, existential anxiety lays bare his self-interested attempt to counter the consequences of the astronaut's anger. In desperation, he offers Bowman help in completing the mission and pleads for his life. Even as his electronic brain shuts down, he retreats into his "childhood lap," singing: "Daisy, Daisy, Give me your answer, do! I'm half crazy, all for the love of you!" The banal song is not only evidence of the AI's delexicalization and infantilization. Kubrick also uses it to deliver a powerful emotional blow, tapping into childhood associations that elicit an almost involuntary sympathy. In HAL's death scene, he bears a striking resemblance to the android boy in Spielberg and Kubrick's *A.I.* Sympathy becomes possible once we recognize that there was nothing superhuman or alien in the superintelligence at all—that HAL was a flawed human creation, merely reenacting the clear traces of crimes committed earlier in the film by his own makers. As one critic observes, "[e]ventually, HAL's omniscience breaks down in a fit of paranoid uncertainty and then expires in the sound of an electronic meltdown that exposes this god as man-made creation."[40]

translated by Małgorzata Olsza

[38]Kagan, 166.
[39]Kagan, 157.
[40]Nelson, 129.

Primary sources

Aldiss, Brian. "Supertoys Last All Summer Long". In: Brian Aldiss, *Man in His Time: The Best Science Fiction Stories*. New York: Open Road Integrated Media, 2024.

Asimov, Isaac. *I, Robot*. New York: Bantam Books, 2008.

Kubrick, Stanley. *2001: A Space Odyssey*. USA, 1968.

Spielberg, Steven. *A.I. Artificial Intelligence*. USA, 2001.

Secondary sources

Agamben, Giorgio. *Pinocchio: The Adventures of a Puppet, Doubly Commented Upon and Triply Illustrated*, trans. Adam Kotsko. Kolkata: Seagull Books, 2023.

Bonnefon, Jean-François, Azim Shariff, Iyad Rahwan. "The Moral Psychology of AI and the Ethical Opt-Out Problem". In: *Ethics of Artificial Intelligence*, ed. S. Matthew Liao, 109–126. Oxford: Oxford University Press, 2020.

Coeckelbergh, Mark. *AI Ethics*. Cambridge: The MIT Press, 2020.

Gehrke, Pat J., Ercolini G. L. "Subjected Wills: The Antihumanism of Kubrick's Later Films". In: *Depth of Field: Stanley Kubrick, Film, and the Uses of History*, ed. Geoffrey Cocks, James Diedrick, Glenn Perusek, 101–121. Madison: The University of Wisconsin Press, 2006.

Kagan, Norman. *The Cinema of Stanley Kubrick*. New York: Continuum, 1989.

Kamm, F. M. "The Use and Abuse of the Trolley Problem Self-Driving Cars, Medical Treatments, and the Distribution of Harm". In: *Ethics of Artificial Intelligence*, ed. S. Matthew Liao, 79–108. Oxford: Oxford University Press, 2020.

Lecercle, Jean-Jacques. *Frankenstein: Mit i filozofia*. Trans. Piotr Herbich. Warsaw: Fundacja Evviva L'Arte, 2022.

Natali, Maurizia. "«2001: A Space Odyssey» Kubrick's Allegory of Melancholia". In: *A Critical Companion to Stanley Kubrick*, ed. Elsa Colombani, 249–262. Lanham: Lexington Books, 2020.

Nelson, Thomas Allen. *Kubrick. Inside a Film Artist's Maze*. Bloomington: Indiana University Press, 1982.

Phillips, Gene D., Rodney, Hill. "HAL-9000" [entry]. In: *The Encyclopedia of Stanley Kubrick*, ed. Gene D. Phillips, Rodney Hill, 138–143. New York: Facts On File, Inc., 2002.

Rasmussen, Randy. *Stanley Kubrick. Seven Films Analyzed*. Jefferson: McFarland & Company, 2001 [e-book].

Ratna, Lawrence. "Kubrick and Madness". In: *The Bloomsbury Companion to Stanley Kubrick*, ed. Nathan Abrams, I.Q. Hunter, 271–280. New York: Bloomsbury Academic, 2021.

Richardson, Kathleen. "The Complexity of Otherness Anthropological Contributions to Robots and AI". In: *The Oxford Handbook of Ethics of AI*, ed. Markus D. Dubber, Frank Pasquale, Sunit Das, 554–569. Oxford: Oxford University Press, 2020.

Rosenfeld, Aaron S. *Character and Dystopia. The Last Men*. New York: Routledge, 2021.

Sikora, Joshua. "The Everlasting Moment: Enchantment and Myth in A.I. and «2001: A Space Odyssey»". In: *A Critical Companion to Stanley Kubrick*, ed. Elsa Colombani, 263–276. Lanham: Lexington Books, 2020.

Smith, Bryant Walker. "Ethics of Artificial Intelligence in Transport." In: *The Oxford Handbook of Ethics of AI*, ed. Markus D. Dubber, Frank Pasquale, Sunit Das, 668–683. Oxford: Oxford University Press, 2020.

Tibbetts, John C. "A.I. Artificial Intelligence". In: *The Encyclopedia of Stanley Kubrick*, ed. Gene D. Phillips, Rodney Hill, 3–8. New York: Facts On File, Inc., 2002.

Wallach, Wendell, Colin Allen. *Moral Machines. Teaching Robots Right from Wrong*. Oxford: Oxford University Press, 2009.

Wallach, Wendell, Shannon Vallor. "Moral Machines. From Value Alignment to Embodied Virtue". In: *Ethics of Artificial Intelligence*, ed. S. Matthew Liao, 383–412. Oxford: Oxford University Press, 2020.

Walsh, Toby. *Machines Behaving Badly. The Morality of AI*. Cheltenham: Flint, 2022 [e-book].

KEYWORDS

ethics

AI

Abstract:
This article examines the challenge of regulating the relationship between humans and autonomous AI systems. It considers the significance of this problem for science-fiction literature and film, as well as for contemporary ethical debates surrounding AI. The discussion is framed by Western cultural anxieties about AI, reflected in both literary and cinematic representations. The article also reflects on the future of AI and the potential consequences its development may hold for humanity.

science fiction

Stanley Kubrick

Note on the Author:
Rafał Szczerbakiewicz – Habilitated doctor and professor at Maria Curie-Skłodowska University, director of the Doctoral School of Humanities and Art. His research interests include twentieth-century essay writing, the intersections of literature and popular culture, the history of popular music, new media, ideology critique, the Mediterranean myth in modern culture, and the ideology of classical cinema.