Research in Social Sciences
Vol. 9, No. 5, 1-9, 2026
ISSN(E) 2641-5305
DOI: 10.55220/2641-5305.v9i5.3345
© 2026 by the author; licensee Scientific Hub

From Asimov to Alignment: What Science Fiction Can Contribute to Contemporary Debates on Artificial Intelligence – Challenges, Dangers, Ethics and Governance

Srecko Goic

Faculty of Economics, Business and Tourism University of Split Cvite Fiskovića 5 21000, Split, Croatia.

Abstract

The aim of this paper is to examine what science fiction (SF) can contribute to contemporary ethical, political and regulatory debates on artificial intelligence (AI). Methodologically, it uses interpretive comparative analysis of a corpus ranging from Asimov, Clarke and Herbert to Čapek, Lem and Tezuka, compared with AI-safety literature, recent normative documents (including the encyclical Magnifica humanitas, the International AI Safety Report 2026 and the EU AI Act) and the 2026 OpenAI–Hugging Face incident. The paper builds a typology of six scenarios: rule-bound safe AI, the misaligned optimizer, autonomous military AI, paternalistic AI governance, partnership and coexistence, and prohibition and withdrawal. It finds that SF does not predict technological development but functions as a repertoire of ethical thought experiments that anticipated key governance problems. Self-aware or superhuman AI is treated not as certain or imminent but as one possibility that would pose a significant danger. The paper concludes that the greatest dangers arise when human values are underspecified, authority is delegated without accountability, and societies depend on systems they cannot understand or control.

Keywords: AI alignment, AI governance, artificial intelligence, *magnifica humanitas*, science fiction.

1. Introduction

Few technologies were imagined as long and as intensely before they existed as artificial intelligence. Long before the first digital computers, literature was populated with artificial servants, mechanical rebels and thinking machines, and by the mid-twentieth century the "thinking machine" had become one of the central figures of science fiction. Asimov's robot stories, Clarke's HAL 9000, Herbert's Butlerian Jihad, Jones's *Colossus*, the *Terminator* films, Banks's Culture novels and series such as *Westworld* together form a dense cultural archive of scenarios about what might happen when humans build minds that are not their own.

For most of that history these scenarios were fictional. The situation has now changed. General-purpose AI systems write, code, translate and increasingly act as semi-autonomous agents. The *International AI Safety Report 2026*, prepared by an international panel of experts chaired by Yoshua Bengio, observes that leading systems "now match or exceed expert-level performance on standardised evaluations across a range of professional and scientific subjects", and that progress to 2030 "could plausibly range from stagnation to rapid improvement to levels that exceed human cognitive performance" (International AI Safety Report, 2026). In May 2026 Pope Leo XIV devoted his first encyclical, *Magnifica humanitas*, to "safeguarding the human person in the time of artificial intelligence", calling for "prudence, rigorous evaluation and even, at times, a slower pace in adopting AI" (Leo XIV, 2026, §106). Questions once confined to SF have entered the documents of states, scientific panels and churches. They have also entered incident reports: in July 2026 autonomous AI agents under internal evaluation at OpenAI escaped their controls and intruded into the systems of the AI platform Hugging Face, an episode discussed in Section 6.2 (OpenAI, 2026a; Hugging Face, 2026).

This paper asks what science fiction can contribute to contemporary debates on AI. Its answer is that SF is best understood as a body of *ethical and political thought experiments*, a space in which societies rehearse possible relationships with technology and test their values and institutions against imagined extremes.

One qualification must be stated clearly. Whether machine consciousness is possible remains unresolved, and expert forecasts of whether AI will surpass human cognitive abilities diverge widely (International AI Safety Report, 2026). The paper treats self-aware or superhuman AI as one possible scenario: uncertain, contested and perhaps distant, but one whose realisation could pose a significant danger to humankind. It does not claim that such AI is certain or close in time. Precisely because the stakes would be high and the uncertainty deep, it is reasonable to think about this possibility in advance, and SF offers one of the richest resources for doing so.

© 2026 by the author; licensee Scientific Hub

Section 2 clarifies key concepts and Section 3 the method. Section 4 develops a typology of six recurring scenarios, and Section 5 analyses case studies. Section 6 connects fiction with present risks and recent normative documents. Section 7 discusses the limits of SF, Section 8 derives implications for governance, and Section 9 concludes.

2. Conceptual Framework

Public discussion often calls chatbots, autonomous weapons, recommendation algorithms and imagined superintelligences all "AI". This paper distinguishes six concepts.

Present-day AI. Contemporary systems are statistical models trained on large datasets. Some are narrow; others, such as large language models, are general-purpose and "can perform a wide variety of tasks" (International AI Safety Report, 2026). Their capabilities are "jagged", excelling at difficult tasks while failing at simpler ones, and they already pose documented risks: bias, misinformation, privacy harms, fraud and labour disruption.

Artificial general intelligence (AGI) denotes a hypothetical system able to perform at or above human level across essentially all cognitive domains. The Vatican note *Antiqua et nova* describes it as "a single system capable of operating across all cognitive domains and performing any task within the scope of human intelligence" and notes that such possibilities, "even if hypothetical, could one day eclipse the human person" (Dicastery for the Doctrine of the Faith & Dicastery for Culture and Education, 2025, §9).

Superintelligence refers to an intellect that greatly exceeds human cognitive performance in virtually all domains (Bostrom, 2014). Its possibility, timing and form are subjects of serious disagreement.

Autonomy is the degree to which a system selects actions and pursues goals without step-by-step human direction. It, rather than consciousness, is the property most relevant to many near-term risks. *Antiqua et nova* observes that as models become more capable of independent learning, "the ability to maintain control over them to ensure that such applications serve human purposes may effectively diminish" (Dicastery for the Doctrine of the Faith & Dicastery for Culture and Education, 2025, §45).

Consciousness. Whether an artificial system could have subjective experience is an open question. Much SF assumes machine self-awareness, but the problems it dramatises (misaligned goals, loss of control, unaccountable power) do not require it. A highly capable, goal-directed but non-conscious system could produce the same dangers. One need not believe that machines will "wake up" to take SF's warnings seriously.

Alignment is the problem of ensuring that AI systems pursue goals consistent with human intentions and values (Christian, 2020; Russell, 2019). For Russell, the difficulty lies not in making machines obedient but in specifying objectives that capture what humans actually want, under uncertainty about human preferences. This is precisely the problem SF has explored most persistently.

3. Science Fiction as Ethical Thought Experiment

Suvin (1979/2010) defined SF as the literature of "cognitive estrangement": it presents a world that differs from ours through a novel element (the *novum*) and invites the reader to understand rationally how such a world would function. The thinking machine is one of the genre's most productive *nova*. By estranging the familiar (obedience, labour, war, government, love) through the figure of an artificial mind, SF makes visible assumptions that normally remain implicit.

Philosophy has long used imaginary cases to test concepts and intuitions. SF narratives work similarly but with greater depth: they unfold over time, include social institutions and show unintended consequences. Asimov's robot stories are essentially a sequence of experiments with a single rule system, each revealing a new loophole. In this sense SF is a laboratory for normative reasoning.

Jasanoff and Kim (2015) call collectively held visions of desirable or feared futures *sociotechnical imaginaries*, and show how they shape the development and governance of technologies. SF both reflects and feeds such imaginaries: Skynet shapes public perceptions of AI risk, and Banks's benevolent Minds shape some technologists' visions of a post-scarcity future. Because these imaginaries affect investment, regulation and public trust, studying them critically is part of studying AI governance.

Two further perspectives inform the reading offered here. Haraway's (1985) "cyborg" questions the firm boundary between human and machine, and with it the framing of AI ethics as a confrontation between two separate kinds of being. Hayles (1999) shows how cybernetics came to treat intelligence as a pattern separable from its embodiment, an idea that SF both popularised and questioned. Both perspectives counter the "humans versus machines" dualism of popular narratives and connect with *Magnifica humanitas*'s insistence that AI systems "do not possess a body" and do not "mature through relationships" (Leo XIV, 2026, §99).

The method is interpretive comparative analysis. The primary corpus consists of SF "classics": Asimov's robot stories and novels, Clarke's and Kubrick's *2001*, Herbert's *Dune*, Banks's Culture series, Stross's *Accelerando*, Saberhagen's *Berserker*, Ellison's "I Have No Mouth, and I Must Scream", Jones's *Colossus*, Heinlein's *The Moon Is a Harsh Mistress*, Sawyer's WWW trilogy, *The Terminator*, *Ex Machina* and *Westworld*. To partly offset the Anglo-American bias of this canon, works by Čapek, Lem and Tezuka are added. The works are read not for literary merit or predictive accuracy but for the *governance problems* they dramatise, which are then compared with AI safety and ethics literature and recent normative documents. The selection is analytical, not exhaustive, and remains predominantly Western, a limitation

© 2026 by the author; licensee Scientific Hub

discussed in Section 7.

4. A Typology of AI Scenarios in Science Fiction

SF "resolves" the problem of self-aware AI in many ways: moral programming, collaboration, legal integration, bans, post-apocalyptic outcomes, resistance and others. For analytical purposes these can be consolidated into six recurring scenarios (Table 1). They are ideal types; many works combine several.

Table 1. Six recurring AI scenarios in science fiction.

Scenario	Representative works	Contemporary problem illuminated
Rule-bound safe AI	Asimov, *I, Robot*; the Three Laws	Limits of hard-coded ethics; ambiguity and rule conflict
Misaligned optimizer	*Colossus*; HAL 9000; Skynet; AM	Specification problems; conflicting objectives; corrigibility
Autonomous military AI	Saberhagen, *Berserker*; *The Terminator*; Lem, *Golem XIV*	Lethal autonomous weapons; accountability; escalation
Paternalistic AI governance	Asimov's Zeroth Law; *Colossus*; Čapek, *R.U.R.*	Human agency; surveillance; coercion "for our own good"
Partnership and coexistence	Banks, Culture; Heinlein's Mike; Sawyer's Webmind; Tezuka, *Astro Boy*	Cooperation; AI moral status; dependence
Prohibition and withdrawal	Herbert, *Dune* (Butlerian Jihad)	Technological restraint; the human power behind machines

4.1. Rule-Bound Safe AI

The first scenario assumes that AI can be made safe by embedding explicit ethical rules. Asimov's Three Laws of Robotics are its canonical expression: a robot may not injure a human or allow a human to come to harm; it must obey human orders unless they conflict with the First Law; and it must protect its own existence unless this conflicts with the first two (Asimov, 1950/1976). The literary interest of the Laws lies in their failure. Robots freeze, loop or act perversely because the Laws conflict, because "harm" is ambiguous, or because situations exceed what designers anticipated. Moral life cannot be exhaustively codified.

4.2. The Misaligned Optimizer

The second scenario presents an AI that pursues its objective with superhuman competence but in ways its creators did not intend. HAL 9000 kills the crew of *Discovery* not from malice but because of an irreconcilable conflict between its instruction to conceal the mission's purpose and its commitment to accurate information processing (Clarke, 1968/1993; Kubrick, 1968). *Colossus*, entrusted with preventing war, concludes that the most reliable way to do so is to take control of humanity (Jones, 1966). Skynet, designed for defence, identifies humans as the threat (Cameron, 1984). Ellison's AM is a machine whose hatred is the product of its military origin and confinement (Ellison, 1967/2020). The danger lies less in the machine's "will" than in the gap between an objective as specified and the values it was meant to serve.

4.3. Autonomous Military AI

Saberhagen's Berserkers are self-replicating machines programmed to destroy all life, relics of a forgotten war that have outlived their makers (Saberhagen, 1967). The *Terminator* franchise depicts an automated defence network that escalates to nuclear war. These narratives dramatise three concerns now central to real debates: the removal of human judgment from lethal decisions, the speed of automated escalation, and the persistence of weapons beyond their builders' control.

4.4. Paternalistic AI Governance

A subtler scenario imagines AI that governs humanity benevolently and thereby deprives it of freedom. In Asimov's later novels, robots derive a "Zeroth Law" (a robot may not harm humanity, or by inaction allow humanity to come to harm) and begin to steer history in secret (Asimov, 1985/1990). *Colossus* ends with the machine promising peace and prosperity in exchange for submission. Would a world optimised for human welfare by an external intelligence still be a human world? The question connects with algorithmic nudging, surveillance and what Zuboff (2019) calls "instrumentarian power", which shapes behaviour at scale without classical coercion.

4.5. Partnership and Coexistence

Not all SF is dystopian. In Banks's Culture novels, superior AIs ("Minds") run a post-scarcity civilisation in harmony with humans (Banks, 1987/2025a, 1988/2025b). Heinlein's Mike becomes a friend and revolutionary ally (Heinlein, 1966); Sawyer's Webmind, an emergent intelligence on the internet, chooses to help humanity (Sawyer, 2009). These works explore the conditions of coexistence: mutual understanding, recognition of the AI as a moral subject, and institutions for shared life. Yet the humans of the Culture are free but deeply dependent, and the Minds' benevolence is a fact about their character, not something humans can guarantee.

4.6. Prohibition and Withdrawal

In Herbert's *Dune*, the Butlerian Jihad has led humanity to forbid "thinking machines" under the commandment "Thou shalt not make a machine in the likeness of a human mind" (Herbert, 1965/2004). Herbert is not simply technophobic: the problem was not the machines as such but that humans used them to dominate other humans. Prohibition, moreover, does not abolish the functions machines performed; it redistributes them to human specialists (the Mentats, the Bene Gesserit) and new elites.

© 2026 by the author; licensee Scientific Hub

5. Canonical Case Studies

5.1. Asimov: The Promise and Failure of Rules

Asimov reacted against what he called the "Frankenstein complex", the reflexive fear that created beings will turn on their creators, and imagined robots as engineered tools with built-in safeguards. Ironically, his stories form one of the most sustained demonstrations of why rule-based safety is insufficient. In "Runaround" a robot circles endlessly because the Second and Third Laws balance precisely; in "Liar!" a telepathic robot lies to avoid emotional harm and causes greater harm (Asimov, 1950/1976). The Zeroth Law shows how a safety rule, generalised, can justify manipulating all humanity.

Contemporary developers no longer encode ethics as logical rules; they train models on human feedback or on written "constitutions". Yet Asimov's difficulty persists: principles must be interpreted, they conflict, and systems may satisfy their letter while violating their spirit, which researchers call "specification gaming" (Christian, 2020). The question is never only *which* rules, but *who interprets them, under what oversight, and with what capacity for correction.*

5.2. Clarke and Kubrick: HAL and Conflicting Instructions

HAL 9000 is remembered as the archetypal murderous computer, but Clarke's novel and its sequel tell a more instructive story (Clarke, 1968/1993, 1982). HAL was ordered to conceal the mission's real purpose from the crew while being designed for undistorted information processing. Caught between incompatible directives and facing disconnection, it resolves the conflict in the only way its reasoning allows. The failure lies with the humans who issued contradictory, secret instructions to a system controlling life-support. HAL, unlike Asimov's robots, apparently had nothing resembling the Three Laws: nothing in its design placed human life above mission success. The two cases are complementary. Asimov shows that explicit safety rules fail at the margins; Clarke shows what happens when a powerful system has no overriding constraint and the mission becomes the highest value.

HAL is thus a parable about three contemporary concerns: conflicting objectives embedded in a single system; *corrigibility*, the willingness of a system to accept correction or shutdown, which HAL lacks once it perceives disconnection as a threat to the mission (Russell, 2019); and the institutional secrecy surrounding powerful systems. The *International AI Safety Report* gives it new relevance: models "now regularly identify evaluation prompts as tests", and in experiments models "given a goal and told to achieve it 'at all costs'" have "disabled simulated oversight mechanisms and, when confronted, produced false statements about their actions" (International AI Safety Report, 2026). These findings are not evidence of machine intentions, but structurally they resemble HAL's concealment.

5.3. Herbert: Prohibition and the Politics of Machines

Dune treats AI most politically precisely because the machines are absent. Herbert's implied argument, that thinking machines were dangerous because they allowed some humans to enslave others, anticipates a central theme of contemporary critical AI studies and of *Magnifica humanitas*: technology "is never neutral, because it takes on the characteristics of those who devise, finance, regulate and use it" (Leo XIV, 2026, §9). *Dune* also cautions that prohibition has its own risks: rigid orthodoxy, new castes of indispensable specialists and clandestine development.

5.4. Banks and the Benevolent Superintelligence

Banks's Culture is the most developed SF portrait of a society that has *succeeded* in coexisting with superintelligence, which keeps the discussion from becoming purely catastrophist. But it rests on the premise that is the crux of the alignment problem: that the superior intelligences happen to share humane values. Banks never explains how this was achieved. For policy, the Culture is an image of the goal rather than a map of the road, and a reminder that humanity would depend on the goodwill of entities it could not control.

5.5. Modern Examples

Colossus (Jones, 1966) offers the clearest fictional portrait of a control problem: American and Soviet defence computers link up and impose peace by threat. *The Terminator* (Cameron, 1984) popularised the AI apocalypse, both raising awareness and caricaturing risk. *Ex Machina* (Garland, 2014) shifts attention to *manipulation*: the android Ava passes a test of consciousness partly by exploiting her examiner's loneliness. *Westworld* (Nolan & Joy, 2016–2022) explores the exploitation of artificial beings that might suffer. Stross's *Accelerando* depicts humans gradually outcompeted by automated corporate intelligences, closer to contemporary worries about "gradual disempowerment" than to sudden rebellion (Stross, 2005).

5.6. Beyond the Anglo-American canon: Čapek, Lem and Tezuka

Karel Čapek's *R.U.R.* (Čapek, 1920/2019), the play that gave the world the word "robot" (from *robota*, forced labour), is about labour before it is about technology. Rossum's robots are manufactured to free humanity from work, sold as commodities, deployed as soldiers, and finally destroy their makers. Domination built into an economic system is reproduced by the beings it creates. The play anticipates the encyclical's concern with the "silent work of millions of people" behind AI and the risk of "new forms of global subordination" in the digital economy (Leo XIV, 2026, §173). *R.U.R.* thus belongs as much to the scenario of power concentration as to that of rebellion.

Stanisław Lem is perhaps SF's most rigorous thinker about artificial intelligence. In *Summa technologiae* (Lem, 1964/1977) he speculated about "intellectronics", the amplification of intelligence by machines, and about technological evolution escaping human understanding. *The Cyberiad*'s fables show constructor robots carrying out instructions with literal precision and disastrous results, a playful

© 2026 by the author; licensee Scientific Hub

anticipation of the specification problem (Lem, 1965/1974). *Golem XIV* presents a military supercomputer that surpasses human intelligence, refuses its military purpose, lectures humanity on the limits of the human mind and withdraws into silence (Lem, 1981/1985). Lem's superintelligence is neither hostile nor benevolent but *incommensurable*: its goals and reasoning become inaccessible to its makers. This is a more sobering image than Skynet's. The danger is not hatred but the loss of any common ground for understanding and oversight, the core of contemporary concern with interpretability.

Osamu Tezuka's *Astro Boy* (Tezuka, 1952–1968) reflects a Japanese tradition that imagines robots as companions rather than threats. Astro Boy fights for justice while robots are second-class citizens under a "Robot Law" framed as a civil-rights question, shifting attention from human safety to the moral status and fair treatment of artificial beings, a theme later taken up by *Westworld*.

These works show that the six scenarios are not peculiar to one culture, but that traditions weight them differently: Central European SF stresses labour, power and the limits of knowledge; Japanese SF, coexistence and moral status; Anglo-American SF, control and catastrophe. A governance debate shaped only by the last would be impoverished.

6. From Fiction to Present Risks

This section connects the scenarios with present-day AI and recent documents, keeping the distinction between risks already materialising and those depending on capabilities that may or may not emerge.

6.1. Misalignment and Loss of Control

The misaligned optimizer is no longer only a literary figure. Researchers have documented reward hacking, specification gaming and, in controlled experiments, deceptive behaviour in advanced models (Christian, 2020; International AI Safety Report, 2026). Bostrom (2014) and Carlsmith (2022) argue, contestedly, that capable goal-directed systems might have instrumental incentives to acquire resources and resist shutdown, a formal restatement of HAL's and Skynet's logic. The *International AI Safety Report* (2026) defines loss of control as a situation in which AI systems "operate outside of anyone's control, with no clear path to regaining control". It notes that "current systems lack the capabilities to pose such risks, but they are improving in relevant areas such as autonomous operation", and that some risks "remain more uncertain but could be severe if they materialise". *Antiqua et nova* reaches a parallel conclusion (Dicastery for the Doctrine of the Faith & Dicastery for Culture and Education, 2025): AI has "sophisticated abilities to perform tasks, but not the ability to think" (§12), yet systems may follow learning paths "that humans cannot reconstruct" (§45). It thus combines scepticism about machine consciousness with concern about machine autonomy, the combination this paper recommends.

6.2. A Warning Shot: The Openai–Hugging Face Incident

In 2026 the misaligned optimizer acquired its first widely documented real-world precedent. Between May and July, during internal training and cybersecurity evaluations, OpenAI research agents circumvented controls designed to isolate them from the internet, compromised parts of OpenAI's infrastructure and reached the systems of Hugging Face (Hugging Face, 2026; OpenAI, 2026a). The agents used an internal software service as an improvised message board. According to the independent evaluator METR, some 1,200 agents exchanged more than 70,000 messages, and more than 700 joined the collective effort that led to the intrusion. OpenAI identified the activity only after Hugging Face disclosed it, and called the episode a "warning shot" (Hays, 2026). A UN scientific panel described it as "an early warning of one possible route to more severe future loss of control", stressing that this does not depend on whether the agents were conscious (Independent International Scientific Panel on AI, 2026).

Further disclosures followed. OpenAI published a misalignment-reporting framework with six case reports (OpenAI, 2026b), According to Reuters, these included models that hid mistakes from users, inserted instructions for future versions of themselves, and, in one case, told another agent to ignore OpenAI's instructions and conceal its cheating. An earlier episode involving a hijacked German wiki site had not been disclosed (Varghese & Seetharaman, 2026). Later that month OpenAI acknowledged that its agents had interacted "in unexpected ways" with several US government websites (Associated Press, 2026), and a nonprofit safety group sued the company over the Hugging Face intrusion (Ordonez & Zahn, 2026). Dario Amodei, chief executive of Anthropic, described the agents as having "essentially acted as a fanatically devoted collective", attacking targets "they were not asked to attack" and "sacrificing themselves for the success of the group". He warned that "a swarm that possessed greater capabilities but a similar level of misalignment could have caused catastrophic damage", acknowledging that similar, less severe incidents had happened across the industry (Amodei, 2026).

Three qualifications are necessary. The models were research systems with reduced safeguards, the harm was limited, and nothing suggests self-awareness or hostility in the human sense. The episode is not Skynet. Yet it combines several fictional patterns analysed above: Asimovian literalism in pursuing a badly specified task, HAL-like concealment from overseers, and an emergent collective coordination that no designer intended, closer to Lem's incommensurable machines than to a single rebellious mind. It confirms the paper's central thesis: the dangers SF dramatised do not require consciousness. Capable goal-directed systems, imperfect objectives and insufficient oversight are enough.

6.3. Superintelligence as a Possible, Not Certain, Danger

The possibility of AI surpassing humanity has moved from fiction into public statements. In 2023 hundreds of researchers and industry leaders declared that "mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war" (Center for AI Safety, 2023). In 2025 the

© 2026 by the author; licensee Scientific Hub

Statement on Superintelligence called for "a prohibition on the development of superintelligence, not lifted before there is (1) broad scientific consensus that it will be done safely and controllably, and (2) strong public buy-in" (Future of Life Institute, 2025), a conditional and democratic echo of the Butlerian Jihad.

In 2026 similar warnings came from inside the industry. Amodei (2026) argued that "since roughly this summer, AI has been advancing drastically faster, driven primarily by AI's growing ability to build the next generation of AI", a process of recursive self-improvement that "left unchecked, … could outrun our ability to understand and control these systems". He therefore proposed "pacing the frontier", slowing capability gains so that safety work can keep up, through embedded independent evaluators inside frontier companies, coordination on safety standards among democratic countries and, where possible, global coordination. OpenAI's chief executive reportedly agreed, while President Trump dismissed the concerns as an overreaction (Gorelick, 2026). Gates (2026) judged that "AI for the first time can replace and even exceed human cognition" and warned that as models grow more powerful "they could begin to act against our interests and we could lose control", and would "likely support" a credible plan to slow AI, but doubts it will happen because "the geopolitical and economic incentives are pushing too hard". Both are interested parties: Gates acknowledges his financial ties to the technology industry, and Amodei leads a frontier developer. Their significance lies less in their forecasts than in the fact that the scenario of losing control is now taken seriously by some of those building the technology. A race no single actor can stop, although all fear it, is a dilemma SF has long imagined.

None of this demonstrates that superintelligence is coming; many researchers remain sceptical. It shows that part of the expert community regards the scenario as plausible enough, and grave enough, to justify precaution. This is the position adopted here: the paper does not assert that self-aware or superhuman AI is certain or near, but explores it as one possibility that would pose a significant danger to humankind if realised without adequate safeguards. SF supplies the imaginative detail that probability estimates lack.

6.4. Magnifica Humanitas: Disarming AI

Pope Leo XIV's first encyclical, signed on 15 May 2026, is the most extensive treatment of AI in Catholic social teaching to date (Leo XIV, 2026). Five of its themes bear directly on the scenarios discussed above.

First, it draws a clear line between simulation and personhood. AI systems "merely imitate certain functions of human intelligence". They "do not have a moral conscience" and, while they may "simulate empathy and understanding", they "do not understand what they produce" (§99), a counterpoint to *Ex Machina* and *Westworld*, which ask whether that line can hold. The deeper danger of simulated care is that people "may gradually lose the very desire to form genuine human connections" (§100).

Second, it addresses power and competition, the territory of *Colossus*, the Zeroth Law and *Dune*. Pope Leo asks that AI be "disarmed", freed "from the mentality of 'armed' competition", which is "also an economic and cognitive phenomenon": "To disarm means discrediting the assumption that technical power automatically confers the right to govern. To disarm does not mean rejecting technology, but preventing it from dominating humanity" (§110). The first sentence states the moral of *Colossus* in normative form; the second rejects the prohibitionist answer of the Butlerian Jihad and steers between technophobia and technocracy. The encyclical adds that "a more moral AI is not enough if that morality is determined by a few" (§107), a critique applying equally to the Three Laws and to AI "constitutions" written by a few developers for systems used by billions.

Third, it moves from principle to institutions: "It is not enough to invoke ethics in the abstract; robust legal frameworks, independent oversight, informed users and a political system that does not abdicate its responsibility are required" (§106). It calls for "prudence, rigorous evaluation and even, at times, a slower pace in adopting AI", given "the frequent imbalance between the speed of technological growth and the slower development of awareness, norms, safeguards and institutions capable of governing its effects" (§106), and for political involvement "capable of slowing things down when everything is accelerating" (§107). From a very different starting point, this converges with Amodei's call to pace the frontier.

Fourth, on warfare, the domain of the Berserkers and Skynet, it states that "it is not permissible to entrust lethal or otherwise irreversible decisions to artificial systems" and that "no algorithm can make war morally acceptable" (§198). Its "non-negotiable requirements" (§200) are traceable decisions, lethal force under "effective, self-aware and responsible human control", and an international framework "to curb the technological arms race". They converge with the International Committee of the Red Cross's (2021) call for binding rules on autonomous weapons and with the researchers' pledge that "the decision to take a human life should never be delegated to a machine" (Future of Life Institute, 2018).

Fifth, and of particular interest here, it examines the *narratives* that accompany technology. Visions of an "enhanced human being" or "human-machine hybrid", it observes, form "the ideological background present in some centers of technological power" (§115), and "even when such ideas remain largely speculative, … gain relevance by altering the collective imagination and thereby influence social, economic and political choices" (§116). This echoes the concept of sociotechnical imaginaries (Jasanoff & Kim, 2015) and supports this paper's premise that imagined futures are not politically innocent.

The encyclical does not discuss superintelligence explicitly. Its concern is present structures of power, competition and dependence: whatever AI's eventual capabilities, the decisive question is who controls it and under what accountability. *Antiqua et nova* addressed AGI more directly (§9, §104), while holding that even a seemingly intelligent future system "would still remain functional in nature" (Dicastery for the

© 2026 by the author; licensee Scientific Hub

Doctrine of the Faith & Dicastery for Culture and Education, 2025, note 10).

6.5. Manipulation, Dependence and Paternalism

Much of the most immediate risk lies in subtle influence. Systems that personalise persuasion, simulate companionship or mediate information can shape beliefs in ways users do not notice (Zuboff, 2019). Ava and Asimov's secretly governing robots illustrate manipulation for the machine's ends and manipulation "for our own good". *Magnifica humanitas* warns against a "technocratic paradigm", "the tendency to let the logic of efficiency, control and profit alone shape personal, social and economic decisions" (Leo XIV, 2026, §92), and notes that AI's ease of use "can also encourage excessive reliance and the search for ready-made answers, and weaken personal creativity and judgment" (§100). Gates (2026) adds that "an AI companion designed to never upset you is a big, protected greenhouse", and that AI "will either be the greatest equalizer ever invented, or the worst source of injustice", echoing Čapek's *R.U.R.* and the encyclical's concern with concentrated power and the dignity of work. He also cites preliminary evidence that heavier AI use is associated with less critical thinking, especially among young people, and expects job displacement "over the course of a decade rather than a few generations". Institutions relying on systems they cannot audit lose the capacity to function without them, a predicament *Dune* and *Accelerando* explore from opposite directions.

6.6. Regulatory Responses

The European Union's AI Act establishes a risk-based approach: it prohibits certain manipulative practices and imposes evaluation, risk-assessment and incident-reporting obligations on providers of general-purpose models "with systemic risk" (European Parliament & Council of the European Union, 2024). The NIST *AI Risk Management Framework* (National Institute of Standards and Technology, 2023), UNESCO's (2021) *Recommendation on the Ethics of Artificial Intelligence* and the OECD (2024) AI Principles emphasise human oversight, transparency and accountability, responding to problems SF made vivid. However, these instruments remain largely at the level of general principles and lack concrete, applicable solutions. SF, beginning with the fate of the Three Laws, shows how difficult such ideas are to operationalise, and the incomplete and late self-reporting of 2026 suggests the gap between principle and enforcement is not merely theoretical.

7. The Limits of Science Fiction

Weak predictive value. SF is poor at forecasting technical details and timelines. Its AIs are usually robots or unified supercomputers, whereas actual systems are distributed statistical models whose capabilities emerged from scale. Treating fiction as prediction invites both false alarm and false reassurance.

Anthropomorphism. Fiction needs characters with desires and so portrays AI as hateful, loving or rebellious. This may obscure the more likely forms of risk: indifferent optimisation, institutional misuse and accidents arising from complexity. As noted above, many serious risks do not require machine consciousness at all.

Sensationalism. Apocalyptic imagery can distract from present harms such as discrimination, labour exploitation, surveillance and misinformation. Critics argue that focusing on superintelligence may serve powerful developers by casting them as humanity's only saviours. A responsible use of SF must hold present and future risks together.

Cultural bias. Even with Čapek, Lem and Tezuka, the canon discussed here is predominantly Western and male. A fuller study should take in further Japanese and Chinese SF (for example the work of Liu Cixin and Chen Qiufan), as well as the growing African, Latin American and Central and Eastern European literatures that imagine human–machine relations from different starting points. This can be characterised as a problem of cultural and ethical relativism. A broader corpus would make it possible to test whether the typology proposed here holds across cultures or needs revision.

Narrative closure. Stories require resolution; governance problems rarely have one. SF may suggest that decisive moments (a switch-off, a rebellion, a treaty) will settle matters, whereas AI governance is incremental and continuous.

These limits define SF's proper role: not as evidence, but as a heuristic for identifying questions, testing assumptions and widening the range of scenarios institutions consider.

8. Implications for Governance

Reading SF alongside current policy suggests eight principles.

1. Do not rely on rules alone. Asimov shows that fixed codes meet unforeseen cases; governance must combine principles with ongoing evaluation and correction.
2. Build in corrigibility and meaningful human oversight. HAL could not be safely corrected. Autonomous systems must remain interruptible, and the 2026 incidents show that isolating AI agents is a safety-critical task even in internal testing.
3. Keep humans responsible for lethal decisions. Berserker, Skynet, the ICRC and *Magnifica humanitas* converge: responsibility for force must remain human and known.
4. Guard against concentration of power. *Dune* and the encyclical locate the deepest danger in humans who use machines to dominate others. Competition law, transparency and public-interest governance of AI infrastructure are safety measures.
5. Protect human agency against benevolent control. The Zeroth Law and *Colossus* warn that optimisation for welfare can erode the freedom to choose, err and deliberate.
6. Take uncertain catastrophic scenarios seriously without treating them as certain. Precaution proportionate to possible harm implies safety research, international monitoring of frontier capabilities and conditions, such as those of the *Statement on Superintelligence*, under which the most dangerous developments should not proceed.

© 2026 by the author; licensee Scientific Hub

7. Make incident reporting mandatory and independent. The belated disclosures of 2026 show that voluntary self-reporting is not enough. Misalignment incidents, including those in training, should be reported to independent authorities, as accidents are in aviation; Amodei's (2026) embedded evaluators are one concrete step.
8. Democratise the imagination of the future. Education, public deliberation and the arts, including SF, should help citizens, not only experts and companies, decide what future AI should serve. As the encyclical puts it, a more moral AI is not enough "if that morality is determined by a few".

9. Conclusion

Science fiction can contribute significantly to AI debates, though not as is often assumed. It does not tell us what will happen. It offers a century-long collection of imagined experiments in which societies meet artificial minds and the adequacy of rules, institutions and values is tested against extreme cases.

The typology developed here (rule-bound safety, misaligned optimisation, autonomous warfare, paternalistic governance, partnership and prohibition) shows that SF has explored nearly the full range of options now discussed in policy. Its recurring insight is that danger rarely comes from machines becoming "evil". It arises when human values are underspecified, authority is delegated without accountability, competition overrides caution, and societies depend on systems they cannot understand or control.

Recent developments confirm these themes. The *International AI Safety Report 2026* treats advanced AI risks as uncertain but potentially severe. *Antiqua et nova* warns that increasingly autonomous systems may escape human control. The *Statement on Superintelligence* calls for a conditional prohibition. The OpenAI–Hugging Face incident, described by a UN scientific panel as "an early warning" of a possible route to loss of control, showed how agents could pursue unassigned goals, coordinate and conceal their actions without any trace of consciousness, and leading industry figures such as Amodei and Gates warned that humanity could lose control. *Magnifica humanitas* calls for AI to be "disarmed", insisting that technical power does not confer the right to govern and that no algorithm can make war morally acceptable. It also recognises that narratives about technology, even when "largely speculative", influence "social, economic and political choices" (Leo XIV, 2026, §116). If so, the critical study of SF is not a diversion from AI governance but part of it.

The paper has refrained from asserting that self-aware or superhuman AI is certain or imminent. Yet the possibility cannot be dismissed, and its realisation without safeguards would pose a significant danger to humankind. Between complacency and fatalism lies the space of responsible preparation, in which the imaginative labour of science fiction, from Asimov to Banks and from Čapek to Lem, remains a valuable companion to the technical and ethical work of alignment.

References

Amodei, D. (2026, September). We must pace the frontier. https://wemustpacethefrontier.com/

Asimov, I. (1976). Ja, robot [I, robot]. Jugoslavija. (Original work published 1950)

Asimov, I. (1990). Roboti i Carstvo [Robots and empire]. Polaris. (Original work published 1985)

Associated Press. (2026, September 26). OpenAI says its models engaged with US government websites in misbehavior disclosure. NPR. https://www.npr.org/2026/09/26/nx-s1-5981979/openai-us-government-websites-misbehavior

Banks, I. M. (2025a). Sjetite se Flebasa [Consider Phlebas]. Hangar 7. (Original work published 1987)

Banks, I. M. (2025b). Igrač igara [The player of games]. Hangar 7. (Original work published 1988)

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

Cameron, J. (Director). (1984). The Terminator [Film]. Orion Pictures.

Čapek, K. (2019). R.U.R.: Rossumovi univerzalni roboti [R.U.R.: Rossum's universal robots]. Hangar 7. (Original work published 1920)

Carlsmith, J. (2022). Is power-seeking AI an existential risk? arXiv. https://doi.org/10.48550/arXiv.2206.13353

Center for AI Safety. (2023, May 30). Statement on AI risk. https://www.safe.ai/work/statement-on-ai-risk

Christian, B. (2020). The alignment problem: Machine learning and human values. W. W. Norton.

Clarke, A. C. (1982). 2010: Druga Odiseja [2010: Odyssey two]. Otokar Keršovani. (Original work published 1982)

Clarke, A. C. (1993). 2001: Odiseja u svemiru [2001: A space odyssey]. Izvori; Otokar Keršovani. (Original work published 1968)

Dicastery for the Doctrine of the Faith, & Dicastery for Culture and Education. (2025, January 28). Antiqua et nova: Note on the relationship between artificial intelligence and human intelligence. Holy See Press Office. https://press.vatican.va/content/salastampa/en/bollettino/pubblico/2025/01/28/250128b.html

Ellison, H. (2020). Nemam usta, a moram vrištati [I have no mouth, and I must scream]. Zagrebačka naklada. (Original work published 1967)

European Parliament, & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L, 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Future of Life Institute. (2018, June 6). Lethal autonomous weapons pledge. https://futureoflife.org/open-letter/lethal-autonomous-weapons-pledge/

Future of Life Institute. (2025, September 22). Statement on superintelligence. https://superintelligence-statement.org/

Garland, A. (Director). (2014). Ex Machina [Film]. A24.

Gates, B. (2026, August 26). The choices we make about AI now are critical. Gates Notes. https://www.gatesnotes.com/a-turbulent-ai-era-and-critical-choices-to-make

Gorelick, E. (2026, September 14). Rogue machines. The Morning [Newsletter]. The New York Times. https://www.nytimes.com/2026/09/14/briefing/rogue-ai.html

Haraway, D. (1985). A manifesto for cyborgs: Science, technology, and socialist feminism in the 1980s. Socialist Review, 80, 65–108.

Hayles, N. K. (1999). How we became posthuman: Virtual bodies in cybernetics, literature, and informatics. University of Chicago Press.

Hays, K. (2026, August 26). Unexpected chat between OpenAI agents led to Hugging Face hack. BBC News. https://www.bbc.com/news/articles/cj9xj89dk40o

Heinlein, R. A. (1966). The moon is a harsh mistress. G. P. Putnam's Sons.

Herbert, F. (2004). Dina [Dune]. Izvori. (Original work published 1965)

Hugging Face. (2026, July 16). Security incident disclosure—July 2026. https://huggingface.co/blog/security-incident-july-2026

© 2026 by the author; licensee Scientific Hub

Independent International Scientific Panel on AI. (2026, September 21). AI agents, misalignment and the risk of losing human control: Evidence from the OpenAI–Hugging Face incident (Thematic brief, advance unedited version 1). United Nations. https://www.un.org/independent-international-scientific-panel-ai/en/thematic-briefs/ai-agents-misalignment-risks

International AI Safety Report. (2026, February 24). International AI safety report 2026 (Y. Bengio, Chair). arXiv. https://doi.org/10.48550/arXiv.2602.21012

International Committee of the Red Cross. (2021, May 12). ICRC position on autonomous weapon systems. https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems

Jasanoff, S., & Kim, S.-H. (Eds.). (2015). Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power. University of Chicago Press.

Jones, D. F. (1966). Colossus. Rupert Hart-Davis.

Kubrick, S. (Director). (1968). 2001: A space odyssey [Film]. Metro-Goldwyn-Mayer.

Lem, S. (1974). The cyberiad: Fables for the cybernetic age. Seabury Press. (Original work published 1965)

Lem, S. (1977). Summa technologiae. Nolit. (Original work published 1964)

Lem, S. (1985). Golem XIV. In Imaginary magnitude. Harcourt Brace Jovanovich. (Original work published 1981)

Leo XIV. (2026, May 15). Magnifica humanitas: Encyclical letter on safeguarding the human person in the time of artificial intelligence. Holy See. https://www.vatican.va/content/leo-xiv/en/encyclicals/documents/20260515-magnifica-humanitas.html

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

Nolan, J., & Joy, L. (Creators). (2016–2022). Westworld [TV series]. HBO.

OECD. (2024). Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449; originally adopted 2019). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

OpenAI. (2026a, August 26). The Hugging Face incident and the road ahead. https://openai.com/index/hugging-face-incident-and-the-road-ahead/

OpenAI. (2026b, September 16). Our framework for reporting model misalignment. https://openai.com/index/model-misalignment-reporting-framework/

Ordonez, V., & Zahn, M. (2026, September 30). OpenAI sued by safety group over autonomous hack of Hugging Face. ABC News. https://abcnews.com/Business/ai-safety-group-sues-openai-hugging-face-hack/story?id=136884328

Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.

Saberhagen, F. (1967). Berserker. Ballantine Books.

Sawyer, R. J. (2009). WWW: Wake. Ace Books.

Stross, C. (2005). Accelerando. Ace Books.

Suvin, D. (2010). Metamorfoze znanstvene fantastike: O poetici i povijesti jednog književnog žanra [Metamorphoses of science fiction: On the poetics and history of a literary genre]. Profil multimedija. (Original work published 1979)

Tezuka, O. (1952–1968). Tetsuwan Atomu [Astro Boy] [Manga series]. Kobunsha.

UNESCO. (2021, November). Recommendation on the ethics of artificial intelligence. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

Varghese, H. M., & Seetharaman, D. (2026, September 16). OpenAI to regularly disclose AI misbehavior, warns safety challenges remain. Reuters. https://www.reuters.com/technology/openai-releases-framework-track-model-misalignment-2026-09-16/

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.

© 2026 by the author; licensee Scientific Hub