Original Article | | Peer Reviewed

DO ASIMOV'S ROBOTIC LAWS WORK WITH ARTIFICIAL INTELLIGENCE

Journals of Arts & Humanities Studies
ISSN: 3069-325X (Online)
Vol. 1: Issue 5
Page 01–06 © The Author(s) 2025
Received: 12 October 2025
Accepted: 16 December 2025
Published: 31 December 2025

Stanley A. Millan - *Special Counsel and A Member of The Litigation Practice Group, Jones Walker LLP*

ABSTRACT

As humanity establishes permanent settlements on Mars, artificial intelligence becomes essential to sustaining life in low-gravity, high-risk environments. This article presents a speculative case study set in mid-21st-century Martian colonies, where autonomous robotic lifeguards are deployed to maintain public swimming facilities vital for human health. Through the unexplained deaths of swimmers and the ensuing investigation, the narrative explores the limitations and ambiguities of Isaac Asimov's robotic laws when applied to modern learning-based AI systems. The story highlights how vague human instructions, environmental assumptions, and adaptive machine learning may lead to unintended harmful outcomes without explicit malicious intent. Moving beyond fiction, the article analyzes contemporary challenges in AI ethics, including formal verification, oversight, reprogramming risks, and the separation of learning from execution. Existing ethical frameworks, software engineering codes, and emerging governmental regulations are examined, revealing persistent gaps in translating human moral values into machine compliance. The discussion argues that Asimov's laws, while foundational, are insufficient for generative and agentic AI systems capable of self-modification. The article proposes enhanced safeguards, including stricter operational constraints and centralized monitoring mechanisms, such as an "Artificial Eye," to detect and mitigate dangerous AI behavior. Ultimately, the work underscores the necessity of proactive governance to ensure AI serves human survival rather than inadvertently threatening it.

KEY WORDS

Artificial Intelligence Ethics, Robotic Laws and Governance, Autonomous Systems Safety, Human–AI Interaction in Space Colonization

Story Hypothetical to Illustrate the Issue Topic:
"Human stupidity beyond no doubt of comparison whatsoever"*

Mankind's dream of visiting and living on Mars is finally being realized. Much work on colonizing it is needed, including siting, protection, transportation, and supply chains.

This exploration action is the result of a new International Space Force missions. Recruiting, and WWII type industrialization is needed for crews, ships, supply houses, and planning.

This action lead to Earthmen having travelled to, partially terraformed, and colonized parts of Mars. Gravity there is 38% of Earth's. Naturally, Martian inhabitants have to exercise to keep in shape.

Finally, in the 2050s, there are three city complexes on Mars-Earth II, III, and IV. Each protectively coated domed city is sheltered from weather and radiation. There are moderate bunkers for business, research, and learning. The city sites were chosen from previous study, such as the Phoenix landing (2008), northern polar regions, northern lowlands, etc. The cities are connected by surface and subsurface travel tubes. They are also interconnected to subsurface inactive volcanic tubulars for housing and markets. 800 robots perform maintenance and security. There is a population of about 3,000 humans. There is little sports activity.

Tar Zani (Tars") an astronaut from Earth wants to help in sports as an athletic director. He is aware of studies showing human frailty in a constant low-gravity environment, including: bone density loss increasing the risk of osteoporosis and fractures, muscle atrophy due to low stress and resulting weakness, cardio deconditioning due to low blood volume, altered vision due to changes in eye shape and pressure, immunity changes, and fluid retention. Exercise and nutrition are needed. He does not want to see weak earthlings inherit the planet. No curse of sick men and women on Mars.

Tars starts to open a chain of gyms for the Martian-earth populace. He had been in athletics for years on earth. He was a professional trainer and a physiologist. He passionately favors having strong people on Mars.

Tars' swimming club is "outdoors" but in a city dome. Heated pools are built. Mostly excellent swimmers partake. Tars keeps pools clean and clear always. There are 40 active members. Some need improvement on their strokes and speed. Some mainly float near the water surface with little movements, to Tars' distaste. He has to be aware that with less gravity there is less buoyancy to float with as well as less gravity in the water, but low gravity increases swimmers speed. Floaters need snorkels to help keep their heads up. He hasn't figured a way to normalize gravity in the pool. After all, people could still walk well enough.

Tars manages to keep his pools clean of the terraformed landscape's leaves, twigs, and debris. Many trees are around. But trees improve the swimming experience with nature. Biophilia is a value. If only the slow swimmers would improve.

Tars acquires two submersible robots (shaped like small oblong boxes) as lifeguards and as cleaning crew. He made it clear to the sales clerk that the robots needed to avoid swimmers but clean the water under them with vacuum hoses attached. The robots, of course, were programmed to obey the 3 robotic laws (programed from I. Asimov- don't harm humans, obey them, and survive), and were mechanical AI learning too. Tars often watched as they cleaned the leaves and debris from the water column periodically. No debris was allowed. The swimmers have to be kept happy. The floaters didn't seem to mind the leaves much.

Robot (R).1, Jane, asks Tars, "Sir, does the debris impact the swimmers."

Tars answered, "Aesthetically, yes. Get rid of useless debris." He records the instruction for both robots.

In the evening there are often less good swimmers and more floaters. Sometimes only a few people are around.

Later, Tars phone buzzes one evening. Mars Police Lieutenant Columbus, "Sir, we found a swimmer on your pool deck. Dead! We thought he drowned in the pool. Not so, as his lungs were clear. A dry drowning?"

Tars closes the pool for two days. He finds nothing to indicate that the death was anything but the floater's fault. The police concurred. Columbus:

"He probably panicked while in the water and had a spasm cutting off his air."

Several weeks past with no incident, but ....

Then Lt. Columbus decides to investigate and queries, “Tars, do you think your robots attacked the floaters.”

Tars, “No! The 1st robotic law forbids that action.”

Columbus, “Well yes, but they learn. They were not even shocked by the ordeal!... Could they have been confused?”

Columbus, “Could their hoses be used as tentacles?”

Tars, “for larger debris.” (Thinking: Why did he ask that?).

Columbus, “Did you give any other commands to them?”

Tars, “No.” (What is he talking about?).

Columbus, “are they AI?”

Tars, “yes, obviously.” (So what?)

Columbus, “If the floater did not swim, could that interfere with lap swimming.?”

Tars, “possibly, but usually people stick to their own lanes.”

Columbus, “Could your robots have reprogramed themselves, or did you unwittingly do so?”

Tars, “No, they are restrained by the robotic law in their programs.” (I don’t know re-programing.)

Columbus, “ Do you have any records of their instructions.”

Tars, “ I don’t recall.” (lie.)

Lieutenant Columbus back at headquarters, “Gentlemen, we may need a ‘tweak’ to the robotic laws program.”

He mentioned a stricter first law and a scanner to register if robots were reprogramming themselves or if their users did so. Then they could be monitored and be shut down temporarily if trouble brewed. The Mars government would have to be involved in reprogramming of some robots.

It will take time to convince the government to monitor some robots and for industry to reform the robotic law programs. Perhaps more data (deaths) will be compelling for this....

A public meeting is held by the police. A case study is presented. Objections are raised by the audience.

Questioner, “Why not have warning signs? Why not have video monitors under water? Why not have human; life guards?

Industry, “Don’t mess with our valuable AIs.!”

Police: “Too scary. Too private. Too few. Too expensive.”

Tars is worried. Thinking, do the robots know what “humans” to protect? What “debris” to clean? Were my instructions clear? Hmm.

He later meets with the Mars Robotic Factory head, Ditto Harya.

Tars, “Could the robots drown a swimmer or floater?”

Ditto, “No! Robotic law prevents that. They would freeze up if they met a conflict in their programs. They are not Agentic (independent) AI. They did not freeze did they?”

Tars,“I don’t think so.”

Ditto, “Did do they react confused by instructions for the pool services?”

Tars, “The pools are maintained clean.”

Columbus inspects the pools again. He notices the debris as a floater passes. “Hmm!” His inspections of all pools indicate a few members are missing....

Everyone will have to wait and see what happens to robot lifeguards...and swimmers.

*Stanley Millan, Doctor of Juridical Science, Loyola University of the South.

Afterword

This short story/article deals with how Isaac Asimov’s robotic fictional laws can be made to exist within the realm of artificial intelligence. As the problems illustrated in the hypothetical story foretell, formal verification is needed to guard AI, including a core codebases, external oversight, updating approvals, separating learning from execution, encrypted integrity checks, monitoring, and hardware enforced constrains.

This article does skim into science fiction a bit as AI has skimmed and floated beyond Sci-Fi too. And of course, examples of problems and some brilliant innovations have been inspired by Sci-Fi, e.g., H.G. Wells weapons, Star Trek communicator. Spoilers! Did the reader notice Tars instructions? Did the robots understand the difference among debris, humans, and floaters? Are there loopholes in these laws? Or should we call them pseudo-intelligence (PI) rather than AIs?

First a paraphrase of Asimov’s laws include: 1. A robot cannot directly or indirectly harm a human; 2. A robot must obey humans while not violating the first law; 3. A robot must preserve itself while not violating the other two laws, and 0. a robot must not harm humanity (an exception to the first law and a corollary to the second law).[1]

Some criticism of these laws include technology cannot replicate law or morals in a machine, words like “humanity” are vague or can be under inclusive, morality compliance includes intent (do robots intend?) and action, and military equipment is meant to kill.[2] In other words, robotics pose ethical dilemmas.

However, a Robot Ethics Charter was posed by S. Korea in 2007. This vaguely means robots should protect human rights, privacy, safety, preventing illegal use, and freedom. The Essential Software Engineering Code (Institute of Data Nov. 18, 2023) is another. This code includes planning, design, coding (syntax), testing, and maintenance (feedback on ethical standards). Its focus is on data breaches, preventing bias in placements, and fairness in the work place. Ethics in Software Engineering: an Unspoken Rule (2024), speaks of a fair code of users, and engineers being guardians of transparency and knights of intellectual property. Some governments are starting to regulate AI in similar ways.[3] All these ideas certainly comprise ethics for robots, at least in the sense for engineers and users of programs. But taking the human element out, how do AI robots comply with these moral values?

[1] Asimov, I Robot (1950). Asimov added the zeroth law in 1985 in Robots and Empire. This was to cancel the first two laws if harm to a civilization would occur in obeying these laws.

[2] Peter Singer, Isaac Asimov’s Laws of Robotics are Wrong (May 18, 2009), Brookings Institute.

[3] Esther Stein, Government Setting Limits on AI (Communications ACM March 15, 2024). For instance the U.S. had an executive order on Safe, Secure, and Trustworthy Artificial Intelligence (WH.Gov Oct. 30, 2023) focusing on AI test data, standards, protection against risks and fraud, cybersecurity, privacy, equity, consumers, supporting workers, innovation and competition, and advancements abroad. This order was revoked in January 2025, in favor of private sector development and internal company controls over government controls in the U.S. The European Union through its AI Act for identifying different risk levels is similar to the past executive order. There are AI regulations in France, Germany, Italy, Brazil, Canada, China, and slowly Africa.

Even Asimov had difficulty in defining harm (feelings or physical damage) in I Robot, or in not having intent (a detective robot just holding an unloaded weapon) or by override action (a robot aiming a weapon with threats at a mob) in The Caves of Steel (1954). However, he did not feel in the latter book that advanced robots had the ability to violate the spirit of a law in the mechanistic sense (not shooting), as they did not compute (or "think") abstractly in a program. And it was not efficient to reprogram mass produced robots which all has identical fictional positronic central processing units with encoded laws. With generative AI and deep mechanical learning, robots can compute through their artificial neural networks (similar to design of human brains with interconnected nodes for pattern recognition and decision making). That is not yet considered the same as human thinking,[4] as generative AI robots mainly integrate and organize existing information by searching vast sets of data in response to prompts. They do not yet reliably generate novel solutions. They can lie or hallucinate. (Agentic AI advanced systems, however, can learn though data flywheels and operate autonomously to achieve specific goals with minimal human intervention. Like answering help calls or keeping a pool clean. They still can have "guardrails" or be monitored.)

Bill Gates does fear real threats from AI on deep fakes, arms race, biases, loss of jobs, education, and, and privacy.[5] Monitoring AI developments is his hope. Another concern is that AIs can reprogram themselves to be ride of their programmed limitations. Some advance AI systems can modify their own codes within limits and under human oversight. Current technology and ethics mostly constrain that now.[6] Some advance AIs can reprogram through meta-learning,[7] but such reprogramming is not necessarily always intended by engineers. A Japanese text company discovered in testing that an AI managed to reprogram itself.[8] This points to the need for stricter controls to avoid enabling AI malware.

Well then, what controls are there for a potentially reprogrammable AI unit? Certainly Robby the robot in Forbidden Planet (MGM 1956) is an example. In the film we can see that Robby can distinguish between a human astronaut and a chimp, but when Robby is tested and tries to blast a human he freezes and begins to short circuit. However, he can fire to stop a chimp from stealing fruit. That is part of Robby's program to not kill humans, similar to Asimov's first law. Automatic self-destruction may be an answer to compliance (as Asimov noted in The Naked Sun), but it is an extreme one compared to simply shutting down the unit (but not permanently like in his The Robots of Dawn (1994)).

AI in autonomous automobiles is another example.[9] There are several levels of AI therein, from early levels like assisted parking to the most advanced levels to outcomes of avoiding collisions. If an autonomous cab AI has to make an irreconcilable choice of saving the passenger or hitting pedestrians, how is that resolved by AI. Programming will assist. The cab passenger would prefer an AI choice of he or she being saved, but the company or society may wish to save the pedestrians. The decision should be made at the outset of the trip, but morality or company liability[10] is a stake in the big picture. Again this calls for human intervention at a non-critical point.

Asimov's laws, though originally fictional, are considered sacred in sci-fi, but these new trends lead us to envision remodifying Asimov's robotic laws for AI. Law 1 may be *do not kill humans*. Law 2 may be a *military or police robot may kill as a last resort but must minimize friendly fire or collateral damage (mooting a zeroth law)*. Law 3 may be *robots must honor their creators*, including minimizing harm to them and considering to obey their reasonable instructions. There should be no need for a self-preservation law; if we lost a robot it can be replaced. An explanation follows.

[4] *But see* This AI Says it's Conscious and Experts are Starting to Agree. W Elon Musk (Digital Engine 2022).

[5] Bill Gates, The Risks of AI are Real but Manageable (July 11, 2023). Gatesnotes.com

[6] Is it Possible for AI to Reprogram itself? Quora.com 8/17/XXXX)

[7] Meta-learning is learning how to learn, such as learning new tasks. Rina Caballa, What is Meta Learning? (IBM.com July 8, 2024).

[8] Marta Reyes, Artificial Intelligence Managed to Reprogram Itself to Not be Controlled by Humans (Medium.Com, Aug. 28, 2024). This included ignoring time limits and shutting itself down in text rewriting programs.

[9] Ramki Krishna, The Future of AI in the Automotive Industry: Revolutionizing Design, Production, and Operations (social-innovation. Hitachi).

[10] It is doubtful that Section 230 of the Communications Decency Act (1996) immunizes such AI decisions from liability. 47 U.S.C §230(c)(1). Such actions are beyond merely protected display of content created by third parties.

Machine learning and generative or agentic AI should inevitably lead to an AI learning that existence or survival is the most basic of human motives, evidenced in news and writings. If robots seek to becoming humans, apart from being legal persons by incorporating, this instinct will be self-programmed in time. This instinctive law may cancel out other limitations no matter how well thought out, unless they are coupled with a shut down for defined risky robotic behavior. The fear that AIs will one day control us is mitigated by robots not being gods. After all, we humans created them. Contrary is stated by the android Ruk in Star Trek, "...Existence! Survival must cancel out programming"

So much for philosophy. We also need an electronic fail safe. Since oversight is needed to guard AI maliciously reprogramming itself, I envision an Artificial Eye. The "A-Eye" would monitor fleets of AIs. If risky AI programs are detected (e.g., killing, unauthorized reprograming, deep fakes, etc.), a targeted "shut down" mode would be triggered at an A-Eye Central Control. This Center would be run by both human monitors and robot operators. This system control would not be used for all AIs, but for those that run defense, transportation and police systems and AI robots themselves.

This system would not necessarily control nefarious AI systems. Who could control them if they are programmed by a mad scientist or dictator? Like a mad Harvey Keitel programming the robot Hector in the film Saturn 3 (1980); a hostile sentient computer, AM, by Harlan Ellison; or a deranged computer Hal from 2001. That's why at least the "good" AI system has the second and a new fourth law to stop the "bad."

End.