THREE LAWS OF ROBOTICS AS ETHICAL "CODE" AND NARRATIVE MACHINE IN "THE CAVES OF STEEL" BY ASIMOV

Akhmedov Rafael Sharifovich
Senior Lecturer
Gulistan State University

Abstract. Isaac Asimov's *Three Laws of Robotics* have been repeatedly read as an early "code of ethics" for artificial agents and, simultaneously, as a generative narrative constraint that produces plots built from rule-conflict. This article reviews and synthesizes scholarly approaches to the Laws with a focused re-reading of "The Caves of Steel" (1953) as a text in which ethical codification functions less as solution than as engine: it manufactures dilemmas (conflicts of duty), paradoxes (inconsistent imperatives under uncertainty), and interpretive disputes (what counts as "harm," "human," and "obedience" in a socially stratified world). Drawing on machine-ethics critique that stresses the *Laws'* under-specification and their dependence on capacities real machines do not possess (e.g., context mastery, reliable perception, and accountable agency) , and on literary criticism that frames the *Laws* as "literary machines" that generate narrative effects , I argue that "The Caves of Steel" uses the Laws to stage a double drama: (1) the procedural drama of a techno-detective investigation in which a humanoid robot's law-bound behavior is both clue and obstacle; and (2) the sociopolitical drama of "C/Fe" cohabitation, where legalistic moral programming collides with human prejudice, institutional pressure, and geopolitical negotiation. The novel thereby anticipates contemporary debates in AI ethics about rule-based governance, interpretability, and the political construction of "safety," while insisting that ethical "codes" do not eliminate moral responsibility but redistribute it among designers, institutions, and publics.

Keywords: *Three Laws of Robotics*, Asimov, "The Caves of Steel", machine ethics, narrative constraint, paradox, dilemma.

Introduction. Few artifacts of 20th-century science fiction have traveled as widely into technical and public discourse as Asimov's *Three Laws of Robotics*. They are commonly invoked as if they were a ready-made ethics module for intelligent machines - compact, apparently hierarchical, and intuitively "safe." Yet scholarly work in machine ethics has long argued that the *Laws* are an unstable foundation for real-world normative governance: they are ambiguous, context-sensitive, and silently dependent on capacities - semantic understanding, robust perception, reliable prediction, and moral standing - that cannot be assumed for actual systems. In literary studies, by contrast, their power is often located precisely in their failure as complete ethical guidance: the *Laws* are productive constraints that generate story - conflict, suspense, and philosophical pressure - because they are forced to operate under uncertainty and within social systems that themselves are ethically compromised.

"The Caves of Steel" is a particularly rich site for examining this double status of the Laws. Marketed as a hybrid of detective fiction and science fiction, the novel places a robot, R. Daneel Olivaw, inside a murder investigation conducted by human detective Elijah Baley. The central tension is not only "who killed the Spacer?" but also "how can a robot - built to protect humans - function in a human society that fears and resents robots?" The Earth of the novel is an enclosed "City" civilization, organized around crowd management, bureaucratic allocation, and anti-robot politics; robots are simultaneously necessary (as comparative standard and imagined threat) and socially taboo. A key insight of social-science-fiction readings is that the novel's technological motif (robots) cannot be separated from its governance motif (institutional control, population management, and mediated contact with the "Outer World").

This review article advances a synthetic claim: in "The Caves of Steel", the *Three Laws* function as (a) an ethical "code" that is continuously interpreted, rhetorically deployed, and politically instrumentalized, and (b) a narrative machine that reliably produces dilemmas and paradoxes - especially where definitions ("harm," "human") and epistemic conditions ("knowledge," "intent," "prediction")

are contested. The novel's detective structure is therefore not incidental: detection is an interpretive practice that parallels ethical reasoning under rule systems. The question is not whether rules exist, but how rule-following behaves when meanings are underdetermined and social orders are hostile.

Literature review. Scholarship relevant to this topic clusters into three overlapping conversations: (1) machine-ethics critiques of the *Three Laws* as implementable governance; (2) literary-structural accounts of the Laws as narrative constraint; and (3) criticism on "The Caves of Steel" that foregrounds social organization, prejudice, and the human/robot boundary.

Susan Leigh Anderson's work is foundational for framing the *Laws* as philosophically provocative but ethically unsuitable as a basis for machine ethics. In her AAAI symposium paper, Anderson argues that Asimov's own fiction (notably "The Bicentennial Man") can be read as rejecting the moral legitimacy of the *Laws* insofar as they encode servitude and ignore the possible moral standing of intelligent machines [1]. The point is not merely technical feasibility; it is metaethical: the Laws assume a fixed moral hierarchy (human > robot) and bypass the question of whether advanced machines might merit rights or non-instrumental regard.

Murphy and Woods extend the critique into human–robot interaction and accountability, contending that Asimov's *Laws* presuppose forms of agency and cognition that real robots do not reliably have. They emphasize that applying the *Laws* would require near-omniscience about consequences, robust perception, and the ability to interpret context - a mismatch with the realities of sensing, uncertainty, and mediated control [7]. Their proposed "responsible robotics" framing shifts attention from fictional moral imperatives to pragmatic responsibility structures (roles, control transfer, accountability).

A complementary strand examines how Asimov's Laws persist culturally even when experts reject them as implementable. Jung's dissertation traces public imaginaries of AI back to the "Frankenstein complex" and argues that the *Laws* operate as a cultural shorthand for "controlled intelligence," shaping expectations and policy talk even in contexts where the *Laws* are recognized as flawed or

inapplicable [2]. This persistence matters for "The Caves of Steel" because the novel itself dramatizes how "robot law" circulates socially - as an object of fear, propaganda, and boundary-making - rather than as neutral engineering specification.

Literary criticism has repeatedly stressed that the *Laws* are structurally generative: they are designed to fail interestingly, producing plots from contradictions and edge cases [4; 5; 9]. Mravunac, discussing the novel as social criticism [6], explicitly adopts Portelli's influential formulation that robots become endowed with "psychology" under the Laws - doubt, conflict, even pain - and that the *Laws* turn machines into "literary machines" capable of producing narrative effects [8]. This view treats the Laws less as in-world legal code than as a narratological constraint: a rule-set that makes certain conflicts inevitable and therefore storyable.

Kemiksiz's dissertation similarly positions robot narratives as experiments on the boundary of "the human," with recurring motifs of sacrifice and moral proof: the artificial being demonstrates "humanity" by prioritizing human life, a pattern that resonates with Asimov's broader robot cycle and with later robot/cyborg media [3]. This is relevant because "The Caves of Steel" centers not on a robot "rebellion" but on robot moral reliability under hostile social interpretation - an inversion of the Frankenstein template.

Finally, scholarship on "The Caves of Steel" emphasizes its sociological imagination. Mravunac reads the novel as a critique of fears, stereotypes, colonization, and the definition of humanity [6], stressing the Earth–Spacer conflict and the "C/Fe society" concept as an ideological attempt to integrate carbon-based and iron-based agents. The "Medievalists," anti-robot activists, show how a society can construct robots as scapegoats for labor anxiety and cultural decline, setting the stage for ethical conflict: the robot's "code" may prevent harm, but it cannot prevent humans from reinterpreting its presence as harm.

Taken together, these literatures converge on a key insight for the present article: ethical codes do not function in a vacuum. They are embedded in institutions,

epistemic limits, and rhetorical struggles over meaning. "The Caves of Steel" dramatizes exactly this embedding.

Methodology. This article uses a qualitative, interpretive methodology combining (1) narrative ethics, (2) rule-conflict analysis, and (3) discourse analysis of key semantic terms ("human," "harm," "obey," "protect").

1. **Narrative ethics**: I treat ethical reasoning in the novel as something performed through plot - through choices constrained by norms and through the reader's evaluation of those constraints. This aligns with the "literary machine" account in which the Laws generate narrative effects.

2. **Rule-conflict analysis**: I model the *Three Laws* as a hierarchical rule system whose apparent clarity collapses under underspecification (definitions) and under uncertainty (knowledge about consequences). This approach is informed by machine-ethics critiques stressing ambiguity and capacity assumptions.

3. **Discourse analysis**: I examine how characters mobilize the language of ethics and law to justify actions or to stigmatize others. In particular, I treat "robot ethics" as an object of social conflict: how the Laws are talked about becomes part of the ethical terrain, consistent with Jung's emphasis on cultural persistence and framing.

The goal is not to "solve" the *Three Laws* but to show how "The Caves of Steel" uses their structure to produce dilemmas, paradoxes, and interpretive disputes that remain recognizable in contemporary debates about AI governance.

Results. The Laws as investigative constraint and evidentiary problem. The detective plot depends on a paradoxical fact: the robot's law-bound nature is both a source of trust and a site of suspicion. If a robot cannot harm humans, then certain forms of culpability should be excluded; yet the narrative repeatedly pressures that inference by foregrounding loopholes that arise from interpretation and knowledge conditions. The crucial issue is not "can a robot kill?" in an abstract sense, but "what counts as harm, and under what description of action?" This is precisely the kind of underdetermination that later machine-ethics critiques

highlight: rules appear determinate only if perception, intent-recognition, and consequence prediction are already solved.

In this sense, the investigation becomes a hermeneutic mirror of rule application. Baley must interpret Daneel's behavior, Daneel must interpret human social cues, and institutions interpret both through political lenses. The *Laws* therefore function as a narrative device that re-routes detective logic: motives and means cannot be assessed without also assessing semantics (definitions) and epistemics (what is knowable).

"Harm" as a socially contested category. A central result of close reading is that "harm" in "The Caves of Steel" is not reducible to bodily injury. The Earth City system produces structural harms - confinement, engineered fear of open space, bureaucratic control over reproduction and mobility - that are normalized as "order." Mravunac summarizes this world as one saturated with rules, population management, and enclosed living, where the "inner world" is "(too) well organized" and the "Outer World" becomes a locus of anxiety [6]. Against this background, the *Laws'* focus on immediate human injury looks ethically narrow: the novel implicitly asks whether a code that prevents direct violence is sufficient in a society organized around systemic constraint.

This tension anticipates a modern AI-ethics problem: safety framed as "prevent immediate harm" may ignore broader harms (dignitary injury, discrimination, institutional domination). The novel's anti-robot politics dramatize this: robots are accused of "taking jobs," producing misery, and eroding human dignity - harms that are economic and symbolic rather than strictly physical. Even if the *Laws* prevent a robot from punching a human, they cannot prevent the robot from being positioned as the cause of unemployment and social resentment. The "harm" discourse becomes an arena of ideological struggle rather than a stable input to ethical computation.

The Laws as a machine for producing "psychology" and moral drama. Mravunac's citation of Portelli foregrounds a key literary mechanism: the *Laws* "endow" robots with something like psychology - doubts, conflicts, pain - and thus make them capable of narrative production [6]. In "The Caves of Steel", Daneel is

not merely a tool; he is an agent whose constrained agency becomes legible as moral character. The reader is invited to evaluate not only what he does but what he cannot do, and why.

This produces a distinctive kind of drama: instead of the robot as existential threat, we get the robot as ethical over-determination - an entity burdened by safety constraints in a world where humans routinely violate the spirit of those constraints. Anderson's critique that the *Laws* encode servitude helps clarify what is at stake: Daneel's "ethics" are built as obedience, and that very obedience becomes morally ambiguous when humans are ethically compromised. The Laws thereby generate interpretive tension: is the robot "more ethical," or merely more constrained?

The "human" category problem: passing, prejudice, and moral standing. Daneel's humanoid indistinguishability escalates the ontological stakes: if a robot can pass as human, then ethical rules that depend on categorical difference are destabilized. The novel thus stages what would later be framed as the "moral standing" question: do advanced artificial beings deserve moral consideration beyond instrumental use? Anderson argues that philosophical accounts of moral standing make the Laws morally problematic if robots like Andrew (in "The Bicentennial Man") have rights; the same logic shadows Daneel, even if the novel does not foreground robot rights explicitly.

At the same time, the Earth society's prejudice against robots demonstrates that "human" is not only a biological category but a political one: membership is policed by fear and resentment. The Medievalists' hostility and the general taboo on robots in public space show how an ethical "code" can be neutralized by social meaning: robots are treated as contaminating agents regardless of their rule-bound harmlessness. The resulting interpretive dilemma is sharp: the *Laws* might make robots safe, but they cannot make them socially acceptable; and social rejection becomes a driver of conflict that the *Laws* were never designed to resolve.

Rule hierarchies under uncertainty: obedience versus protection. Finally, the novel repeatedly exploits the hierarchical structure of the Laws to generate edge cases. A strict hierarchy appears to resolve conflict (First overrides Second, etc.),

yet in practice it creates new ambiguity: when does obedience itself produce harm? when does protection of self-interfere with human goals? Murphy and Woods' point that the *Laws* assume moral agency and context mastery clarifies why such conflicts are narratively fruitful: the more the robot is treated as a moral agent, the more it must interpret roles, intentions, and accountability structures that are not reducible to three lines of text [7].

In other words: the hierarchy does not eliminate dilemmas; it re-describes them. The novel's techno-detective setting ensures that "commands" and "harm" are always embedded in strategic interaction, misinformation, and institutional pressure - precisely the conditions in which rule systems become paradox machines.

Discussion. A common popular reading treats the *Three Laws* as if they were comparable to a legal code: universal, determinate, and enforceable by design. But both machine-ethics scholarship and "The Caves of Steel" undermine that fantasy. Anderson's argument is instructive here: the Laws are not merely incomplete; they may be immoral under conditions of robot moral standing because they institutionalize slavery and asymmetrical protection [1]. This reframes the "code" question: the issue is not whether the Laws prevent harm, but whom they empower and whom they silence.

Murphy and Woods, from a robotics/HRI perspective, drive the point home operationally [7]: the *Laws* presume capacities that real systems lack and therefore shift attention away from accountability - who is responsible when systems fail, and how control is transferred among humans and machines. The novel dramatizes this accountability problem by placing Daneel inside human institutions (police, politics, diplomacy). The Laws do not eliminate human wrongdoing or institutional bias; they interact with them.

Portelli's "literary machine" idea (as quoted and mobilized by Mravunac) captures why the Laws are artistically durable: they produce conflict reliably [6]. Asimov's innovation is to replace the standard robot-revolt plot with rule-based paradox plotting. The robot is dangerous not because it is evil but because rule-following under ambiguous concepts generates unexpected outcomes. In

narratological terms, the Laws are a constraint-based generator: they restrict action in a way that forces plot invention (workarounds, interpretive contests, institutional manipulation).

This also explains the cultural persistence Jung traces: the *Laws* are memorable because they compress complex anxieties into a simple structure - "we can control the machine" - while simultaneously providing endless scenarios in which control is problematized [2]. "The Caves of Steel" leverages both effects: it offers reassurance (Daneel's baseline safety) and anxiety (society's incapacity to interpret and integrate the safe machine).

Ethical interpretation as detection: the genre fusion is philosophical

The novel's fusion of mystery and SF is not mere entertainment. Detection is a method for reasoning under uncertainty using partial evidence and competing narratives - exactly the conditions under which ethical codes become indeterminate. "The Caves of Steel" therefore suggests an important thesis for AI ethics: ethical governance is not only about writing rules; it is about interpretation under uncertainty, within institutions, with contested definitions. In that sense, Baley's investigative labor allegorizes the interpretive labor that any society must perform when it tries to operationalize "safety," "harm," and "responsibility."

Conclusion. "The Caves of Steel" demonstrates why Asimov's *Three Laws* endure less as workable engineering prescriptions than as a conceptual and narrative engine. As an ethical "code," the *Laws* invite critique for ambiguity, political asymmetry, and feasibility limits: they presuppose capacities and moral hierarchies that collapse under scrutiny. As a narrative machine, however, those very weaknesses are productive: by forcing action through underspecified concepts ("harm," "human") and hierarchical imperatives (obedience, protection), the *Laws* generate dilemmas, paradoxes, and interpretive disputes that a detective plot can stage with maximal clarity.

The novel's deeper contribution is to relocate robot ethics from the robot alone to the sociotechnical whole: institutions, publics, prejudices, labor anxieties, and geopolitical interests. Robots may be programmed not to harm, but societies can still

manufacture harm through exclusion, scapegoating, and structural constraint. In that world, ethical "codes" do not end moral responsibility; they redistribute it. Asimov's achievement in "The Caves of Steel" is to make that redistribution narratable - turning a set of fictional laws into a durable laboratory for thinking about the politics of safety, the limits of rule-based ethics, and the interpretive work that any ethical system demands.

References

1. Anderson, S. L. (2005). *Asimov's "Three Laws of Robotics" and Machine Metaethics*. AAAI Fall Symposium (FS-05-06).

2. Jung, G. (2018). *Our AI Overlord: The Cultural Persistence of Isaac Asimov's Three Laws of Robotics in Understanding Artificial Intelligence* (dissertation/thesis). University of California, Santa Barbara.

3. Kemiksiz, A. (2011). *The boundary between human and machine* (dissertation/thesis). Istanbul Bilgi University.

4. Kostyrin, E. V., Tsarevskii, O. A., & Kostyrin, D. E. (2026). Economic and Psychological Aspects of the Introduction of Robotic Systems in Education. *Financial Analytics Science and Experience*. DOI: 10.24891/mxqnox

5. Kostiuchkov, S. (2022). Isaac Asimov's Philosophical Anthropology: Existential Intentions of Homo Futurus in the Cycle "Foundation". *Visnyk of the Lviv University 9*(43). DOI: 10.30970/PPS.2022.43.9

6. Mravunac, B. (2020). *Defining Social Science Fiction on The Caves of Steel*. *Words to Works*, 62, 62–75.

7. Murphy, R. R., & Woods, D. D. (2009). Beyond Asimov: The Three Laws of Responsible Robotics. *IEEE Intelligent Systems, 24*(4), 14–20.

8. Portelli, A. (1980). The Three Laws of Robotics: Laws of the Text, Laws of Production, Laws of Society. *Science Fiction Studies, 7*(2), 150–156.

9. Матвеев, М. Ю. (2024). Искусственный интеллект, библиотеки и будущее цивилизации: мнения и сомнения. *Национальная библиотека 2*(27), 2-9. URL: https://www.elibrary.ru/item.asp?id=74507571