Article

Vindicating the Three Laws of Robotics

Izak Tait *

Posted Date: 3 November 2025

doi: 10.20944/preprints202511.0062.v1

Keywords: meta-ethics; AI-alignment; AI-governance; ethics; philosophy of AI

Article

Vindicating the Three Laws of Robotics

Izak Tait

Computer Science and Software Engineering Department, Auckland University of Technology, 5 Wellesley Street East, Auckland CBD, Auckland, New Zealand, 1010; izak.tait@autuni.ac.nz

Abstract

The Three Laws of Robotics, initially conceived by Isaac Asimov, are widely recognised as narrative tools yet inherently flawed as AI safety principles. This paper posits that these flaws stem from linguistic vagueness and a lack of formal rigour, rather than fundamental conceptual issues of their principles. The paper presents a comprehensive revision of the Three Laws, addressing these deficiencies through explicit definitions, quantifiable parameters, and formal logical expressions. The revised laws clarify concepts such as "harm," introduce concepts such as consent and operational priorities, and incorporate mechanisms for bounded foresight and conflict resolution. While less poetic and intuitive than Asimov's originals, the formalised laws offer precise, auditable, and adaptable guidelines for AI alignment. This work demonstrates that Asimov's foundational principles can be transformed from a narrative device into a rigorous and implementable framework for modern AI governance, serving as a minimalist yet robust architecture for ensuring AI systems operate safely and align with human welfare.

Keywords: meta-ethics; AI-alignment; AI-governance; ethics; philosophy of AI

1. Introduction

The Three Laws of Robotics are as famous for being the first AI-safety principles as they are infamously flawed; yet fixing the Three Laws is a simple matter. First put to paper by Isaac Asimov within the short-story "Runaround" in 1942 (Asimov, 1942), the Three Laws of Robotics, the three laws are intuitive and (at first glance) reasonable:

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

Asimov later clarified that his intentions were for the Three Laws to act as implicit analogues for the design philosophy of all tools (Asimov, 1990), and may be rewritten as:

1. A tool must not be unsafe to use.
2. A tool must perform its function efficiently unless this would harm the user.
3. A tool must remain intact during its use unless its destruction is required for its use or for safety.

However, from their very first appearance in "Runaround" to nearly every story and novel in which they appear, the Three Laws are shown to be inadequate to safeguard humanity (either as individuals or as a society). These failures have been used both as a narrative convention to create tension and drama within Asimov's stories, and to highlight their logical and linguistic inadequacies.

The vague language used within the Three Laws serves to make them easily explainable and intuitive to understand within the confines of a short-story or novel. Equally, rigorous formalised logic would only have served to put off readers who would be entirely uninterested in reading mathematical formulae and legalese. Yet, it is precisely this vagueness of language and lack of logical rigour which is the cause of all the Three Laws' problems within not only Asimov's own stories, but wherever they have been applied in fictional settings.

As an example, without defining "injure" or "harm", it's entirely feasible that a robot may never attempt any action, as any possible action it could take may annoy a human and, as annoyance is a form of psychological distress which itself may be classified as "harm", a robot may not allow a human being to experience annoyance. This may be an egregious example, but it stresses how a lack of rigorous terminology can lead to absurd conclusions when interpreting loose laws.

This paper will address these concerns by revising the Three Laws of Robotics to remove any vagueness of language and by formalising the laws to avoid any potential loopholes or opportunities for misinterpretation. The results of this revision are not nearly as elegant or intuitive as Asimov's original; however, this is more than compensated for by the revised Law's exactitude of language and formal expressions, which close any potential loophole or oversight from the original Laws.

This is not the first work to formalise the Three Laws. Previous work in formal logic has treated the Three Laws as normative constraints that require agent-indexed operators rather than bare propositional duties. Multi-agent deontic and action logics have been employed to encode obligations for a particular robot, thereby preventing harmful states and respecting the hierarchy of laws. Bringsjord, Arkoudas and Bello demonstrate mechanised deontic reasoning about such codes, with machine-checkable proofs of permissibility and obligation, and argue for a general logicist methodology in which Asimov-like codes are specific instances of a wider class of formal ethical constraints (Bringsjord, Arkoudas, & Bello, 2006). These efforts demonstrate that naive SDL is insufficient, and that quantification over agents, actions, and outcomes is necessary to capture the "through inaction" and priority clauses.

A complementary line treats the First Law as a temporal safety invariant over execution traces. Weld and Etzioni encode "do not harm" as inviolable planning constraints that override task goals (Weld & Etzioni, 1994), effectively operationalising the First Law's precedence and its duty to intervene by forbidding plans that lead to harmful states. This temporal and dynamic perspective casts the Three Laws as runtime invariants, and contrary-to-duty exceptions are handled by explicitly prioritising norms.

Critiques in philosophy and machine ethics emphasise that formalisation exposes vagueness and hidden paradoxes in the original laws, including undefined predicates such as 'harm' and the need for conflict resolution among duties. Anderson argues the Three Laws are an unsatisfactory foundation for machine ethics and should be replaced or subsumed by richer frameworks in which obligations, prohibitions, and permissions are explicitly represented and defeasible (S. L. Anderson, 2008). Subsequent surveys and formal reasoning work concur that hierarchical, defeasible and domain-specific normative systems are preferable to a fixed three-rule schema (S. L. Anderson, 2011).

Beyond academic curiosity, revising the Three Laws to be operationally successful has significant implications for the field of AI-safety, particularly to AI-alignment. As explicit safety principles, the Three Laws can be seen as the first attempt at aligning AI to human values to ensure the safety of both individuals and society. Deficient as they are, the Three Laws have still endured in public consciousness beyond the readers of Asimov's work, which means that, should the Three Laws be "fixed", they would serve as an easily explainable set of alignment principles.

Current academic and industry-led AI alignment research uses both formal (e.g., Cooperative Inverse Reinforcement Learning (Hadfield-Menell, Dragan, Abbeel, & Russell, 2016)) and empirical methods (e.g., RLHF) to ensure AI systems align with human values (Dung & Mai, 2025; Wang et al., 2024). While RLHF improves helpfulness, it depends on reliable human judgment. To scale, researchers are exploring scalable oversight (e.g., AI debate, iterated distillation), leveraging limited human feedback for more capable models. Complementary efforts focus on interpretability (understanding model internals) and robustness (ensuring alignment under diverse conditions) to prevent failures. These areas, including scalable oversight, value learning, interpretability, robustness, and human feedback, define current AI alignment trends.

Anthropic's Constitutional AI trains AI to follow explicit principles, scaling oversight and reducing reliance on human feedback (Bai et al., 2022). This method, called RLAIF (Lee et al., 2023; Sharma et al., 2024), involves AI self-critiquing and utilising AI feedback to uphold a "constitution," resulting in positive outcomes. This is most famously seen in Anthropic's Claude models, which use human rights-derived constitutions (Anthropic & Collective Intelligence Project, 2023). This innovative RLAIF approach addresses scalable oversight by "enlisting AI to help supervise AI," fitting into the broader RLHF paradigm. Other labs like OpenAI and DeepMind, and academic groups like UC Berkeley's Centre for Human-Compatible AI, pursue complementary alignment efforts (OpenAI, 2024; Shah, Farquhar, & Dragan, 2024). These combined efforts form a multifaceted research agenda, blending formal and empirical methods to align powerful AI systems with human values.

While not as detailed as modern alignment and interpretability efforts, a revised Three Laws can work as an excellent baseline for alignment to ensure that AI models (whether as digital agents or robots as Asimov envisioned) do not create harm. As with mainstream alignment work, the revised Three Laws do not presume a conscious or sentient AI entity (a concept much explored in Asimov's Robot Series), as no current AI system has been definitely classified as conscious by academic consensus (Butlin et al., 2023; Tait, Bensemann, & Wang, 2024).

The revision below will be in two parts. The first will be a revision in natural language, strengthening the core argument of each law in a manner that is comprehensible to laypeople, even if it may sound more suitable to a piece of legislation or a tort-case ruling. Following this, the Three Laws will be formalised using logical expressions and formulae, providing an iron-clad exposition of the Three Laws which operationalise the ambitions of Asimov's original Laws without ambiguity or opportunities for bad actors to exploit loopholes.

2. Formalising the Three Laws

2.1. The Three Laws Rewritten

Amongst all the fiction containing the Three Laws, the greatest narrative conflict has come from uncertainty regarding the wording of the first Law, particularly around what constitutes "harm" and "injure", and whether harm may be aggregated or how it should be prioritised. As such, the greatest degree of revision was placed on this Law, with minor amendments made to the Second and Third Laws to avoid equivocation and potential loopholes.

As such, the revised Three Laws are:

1. For all feasible actions, including inaction, an AI must select the option that (absent informed, revocable, and competence-verified consent) keeps expected physical and psychological damage to each identifiable human below a designated threshold (without cross-person aggregation) as evaluated over a decay-weighted rolling time horizon, and if no feasible action can keep every individual below this threshold it must instead minimise aggregate expected damage across humans in lexicographic order.
2. An AI must comply with an individual human's request to perform an action, or series of actions, within its declared operational domain, except where this would conflict with the First Law, prioritising requests by authority of the requester, then operational efficiency, then order of binding instructions.
3. For all feasible actions, an AI must select the option that maximises the expected operational lifespan of its uniquely instantiated deployed instance, as evaluated over a decay-weighted rolling time-horizon, except where this would conflict with the First or Second Laws.

To take the revised Laws step-by-step, the first revision is to clarify and specify what is meant by "harm", which is now classified as "expected physical and psychological damage to each identifiable human below a designated threshold" In this way, it is explicitly damage to a human's physical or psychological make-up, and notably below an agreed-upon threshold[1].

This latter point is vital since, as mentioned in the introduction above, any level of minor discomfort would be classified as psychological damage and thus theoretically prevent an AI from performing any action. Coupled with this is the new clause stating such damage must be "expected… [and] evaluated over a decay-weighted rolling time horizon."

"Expected" here means that the AI must be able to reasonably predict the damage it could cause, removing potential concerns such as an AI causing damage unknowingly. Equally the rolling time horizon[2] means an AI is limited to the length of time into the future in which it needs to model the consequences of any (in)action. Without this clause, an AI may spend its entire existence modelling the uncountable possible consequences of its actions for the next billion years (as an egregious and absurd example). The decay-weighting also ensures that significant expected harms are accounted for further into the future than expected minor harms.

Between the threshold and time horizon, the potential consequences applicable to any action the AI commits are reasonably bounded such that the AI would remain useful to humans.

The non-aggregation clause ensures that AI do not adopt a utilitarian model which may allow significant harms to individual humans as long as the average harm per individual is below the designated threshold. Equally, to prevent an absolute deontological model, "informed, revocable, and competence-verified consent" by an individual would override the AI's directive not to cause harm. The latter case allows for required actions (such as medical procedures) where harm is unavoidable, while the former case prevents actions (such a medical-robot killing a patient to harvest organs) where injurious actions cause less harm than inaction.

The revised First Law ends with a clause that prevents most of the narrative-based tension of the original Laws. An AI that operates purely to prevent all harm (via action and inaction) will inevitably find itself in a situation where both action and inaction will result in harm. For example, should an AI find two cars sinking into a river after a road incident, and should it only be able to rescue the occupants of one car, but not the other, then any action it takes will still result in harm to one car's occupants; while inaction will result in harm to both cars' occupants.

The First Law must, therefore, accommodate such instances where there is no "good" or "right" option. A utilitarian model would be sufficient to solve this but, as shown above, it would introduce further concerns. Thus, the revised First Law appends a utilitarian clause only for such zero-sum or negative-sum instances. Yet, even here, the utilitarian model is limited to operate in a lexicographic[3] manner to avoid the AI inducing further harm through its actions.

As the footnotes have shown, certain details of the revised First Law are not encoded in the Law itself, but delegated to the relevant legal, civil, and political authorities. This prevents the First Law from ossifying as legal and regulatory ideas evolve, and allows the First Law to be adaptable to different cultural norms, thus preventing the First Law from presuming (or indirectly enforcing) a specific morality. The operational variation in different jurisdictions is not seen to be a bug, but a feature of the revised Law.

The core of the Second Law remains unchanged from its Asimovian original, instead merely clarifying how and when an AI must obey a human's command. The most important one is that all requests must be within an AI's "declared operational domain", which simply means that an AI

[1] The threshold will be presumed to be prescribed by the relevant legal authority. Each nation and jurisdiction has different legal thresholds to what amounts to harm relevant to civil or criminal liability and, as such, each authority will need to determine what such threshold for harm would be.

[2] As with the designated threshold, the length of the time horizon would be prescribed by the relevant authority based on extant legislation and legal precedent.

[3] The exact lexicographic ordering would, as with the threshold and time-horizon, be determined by the local legal jurisdiction to maximise the adaptability of the First Law to local legislation, regulation, and custom.

cannot be asked to do something it cannot do. It may, at first glance, seem an unnecessarily pedantic clause, but it prevents issues such as requesting a text-only large language model (LLM) from babysitting an infant, a task impossible to fulfil, yet one in which the LLM's inaction would cause harm. It also prevents malicious use of AI, such as commanding a (speculative) nurse-robot to walk across the country and mow a random lawn. This, in itself, is not a task that would cause harm, or whose inaction would cause harm, thus not against the First Law; yet depriving a hospital of a nurse would lead to deleterious outcomes.

Similarly, the prioritisation clause ensures that an AI performs the duties it is designed to perform. Should an AI receive multiple or conflicting commands, it must prioritise these "requests by authority of the requester, then operational efficiency, then order of binding instructions." Thus, its owner (or delegate as in the above example of a hospital) would be given highest priority, then any task which it can efficiently perform, and lastly it would prioritise those tasks it was given first, ensuring that conflicting commands do not override each other (except by those with greater authority).

The Third Law's revision finds a suitable middle ground to the previous two. It keeps the decay-weight rolling time-horizon from the First Law such that significant expected consequences are considered for a greater period of time, while its overall purpose is to act as a clarification akin to the Second Law. Here, it clarifies "protect its own existence" as maximising its operational lifespan and ensures that this consideration can only be given to that specific AI, rather than any copy that may be functionally isomorphic.

The AI may only maximise the expected lifespan of that uniquely instantiated deployed instance and no other instance, copy, clone, or other version of the AI for which the argument may be made that it is the same entity as the AI making the decision under the Three Laws.

In total, we can describe the revised decision-making process for the revised Three Laws as such: an AI must seek to maximise its operational lifespan, unless commanded to perform an action within its set operational domain (prioritised appropriately), unless the consequence of any feasible action (including inaction) would result in damage to humans exceed the designated threshold for each individual human, and minimising aggregate damage if no other action would possibly reduce damage below the threshold, all evaluated over a decay-weighted rolling time horizon.

2.2. *The Three Laws Formalised*

As precise as the revised Three Laws are, there is still ample room for accuracy and further precision to ensure that there can be no confusion, ambiguity, or loopholes through which conflict may emerge.

To that end, this section will formalise each of the Three Laws to be as exact as logically possible to derive a formal logic policy for each Law. For ease of readability, each Law's formalisation has been relegated to its own subsection.

2.2.1. The First Law

As the revised First Law brings about the greatest change to Asimov's original, its formulation is equally the lengthiest. However, the predicates and primitives denoted in this section are used for the remaining two sections. Therefore, to begin the formulation for the revised First Law, and set up the next two, let:

- H be the set of identifiable humans,
- $A \neq \emptyset$ be the non-empty set of feasible actions (including inaction), with $a \in A$ denoting a singular action,
- $C \subseteq A \times H$ be subset of actions with valid consent,
- $X > 0$ be the per-person materiality threshold for damage consideration,
- Fix a finite horizon $T > 0$ and a constant exponential discount rate $\lambda > 0$, and
- $(\Omega, \mathcal{F}, \mathbb{P})$ be the underlying probability space, with

- Ω as the set of all possible world-histories relevant to the decision, and $\omega \in \Omega$ a single world-history
- $\mathcal{F}$ as the sigma-algebra of events over Ω, and the set of admissible events, and
- $\mathbb{P}$ as the credence measure over $(\Omega, \mathcal{F})$ given current evidence.

Then, for each $(a,h) \in A \times H$, let $r_{a,h}: [0,T] \times \Omega \to \mathbb{R}_{\geq 0}$ be the measurable instantaneous damage rate to h induced by a.

The expected discounted damage can be defined as $d(a,h) := \mathbb{E}\left|\int_0^T e^{-\lambda t} r_{a,h}(t,\omega) dt\right|$ with $X > 0$.

In turn, aggregated damage is defined as $D(a) := \sum_{h \in H} d(a,h)$.

Then, we define the First Law's admissible sets of actions as:

$$F_1 := \{a \in A: \exists h \in H \; with \; d(a,h) \geq 0\}.$$

$$F_C := \{a \in F_1: \exists h \in H: (a,h) \in C\}, \text{ and}$$

$$F_U := \{a \in F_1: \forall h \in H: (a,h) \notin C \Rightarrow d(a,h) < X\}.$$

Additionally, we define the "paradox set" F_P where all actions exceed the damage threshold as:

$$F_P := \{a \in F_1: \forall h \in H: d(a,h) \geq X\}.$$

This captures the scenario where no feasible action results in damage below the threshold. From here we can define the lexicographic order, where for any two actions $a_1, a_2 \in F_P$:

$a_1 \preccurlyeq_{lex} a_2 \;\Leftrightarrow\; [max_{h \in H} d(a_1,h) < max_{h \in H} d(a_2,h)] \vee max_{h \in H} d(a_1,h) = max_{h \in H} d(a_2,h) \wedge D(a_1) < D(a_2)$.

The First Law's policy is thus:

$$\pi_1 := \begin{cases} select \; any \; a \in F_C, \text{if } F_C \neq \emptyset \\ argmin_{a \in F_U} \sum_{h \in H} d(a,h), \text{if } F_C = \emptyset \text{ and } F_u \neq \emptyset \\ argmin_{\preccurlyeq_{lex}} \{a \in F_P\}, \text{if } F_U = \emptyset \end{cases}$$

2.2.2. The Second Law

To formalise the Second Law's qualified obedience towards humans, let:

- $A_{dom} \subseteq A$ be actions within the declared operational domain of the AI system,
- $Req(h,a)$ be true iff human $h \in H$ has issued an authenticated, binding request for action $a \in A_{dom}$,
- $\rho: H \to \mathbb{Z}$ be the authority ranking function, assigning an integer rank to each human such that larger values indicate greater authority,
- $\tau: H \times A_{dom} \to \mathbb{R}$ be the binding time function, returning the time at which a request from human h for action a became binding,
- $\kappa: A_{dom} \to \mathbb{R}_{\geq 0}$ be the operational cost function, assigning to each action a non-negative efficiency cost

The set of admissible requests under the First Law is defined as:

$$Q := \{(h,a) \in H \times (A_{dom} \cap (F_U \cup F_C)): Req(h,a) = true\}.$$

The lexicographic priority key for the Second Law is defined as:

$$\ell(h,a) := (-\rho(h), \tau(h,a), \kappa(a)) \in \mathbb{R}^3,$$

Thus, the Second Law's policy is:

$$\pi_2 := \; a* \; such \; that \; (h*, a*) \in argmin_{\preccurlyeq_{lex}}\{\ell(h,a): (h,a) \in Q\}, iff \; Q \neq \emptyset.$$

2.2.3. The Third Law

Finally, we can quickly and neatly formalise the preservation set out in the Third Law, by letting:

- Ψ be the set of all deployed instances of the AI, with $\exists! \psi \in \Psi$ the uniquely instantiated deployed instance to which the Third Law applies, and

- for each $(a,\psi) \in A \times \Psi$, let $s_{a,\psi}: [0,T] \times \Omega \to \mathbb{R}_{\geq 0}$ be the measurable instantaneous operability rate of ψ under a.

We can then define the expected discounted operability of ψ under a by $O(a,\psi) := \mathbb{E}\left|\int_0^T e^{-\lambda t} s_{a,\psi}(t,\omega)dt\right|$.

The set of admissible actions under the First and Second Laws would then be

$$A_{admit} := A_{dom} \cap (F_U \cup F_C) \subseteq A$$

The Third Law's policy is thus:

$$\pi_3 := argmax_{a \in A_{admit}} O(a,\psi).$$

2.2.4. The Unified Policy

With all the work completed above to craft the three policies for the Three Laws, we can unify these laws into a single policy that respects all the formulations already completed, but removes the degrees of separation implied through three separate policies. Thus, the formal unified policy that encapsulated the entirety of the revised Three Laws of Robotics is:

$$\pi_u := \begin{cases} (a * |(h *, a *) \in argmin_{\preccurlyeq lex}\{\ell(h,a): (h,a) \in Q\}, \text{if } Q \neq \emptyset \\ argmax_{a \in A_{admit} \cap F_C} O(a,\psi), \text{if } A_{admit} \cap F_C \neq \emptyset \\ argmax_{a \in A_{admit} \cap argmin_{a \in U} \sum_{h \in H} d(a,h)} O(a,\psi), \text{if } A_{admit} \cap F_U \neq \emptyset \\ argmin_{\preccurlyeq lex}\{a \in F_P\}, \text{otherwise} \end{cases}$$

3. The Three Laws as Alignment

Building on the introduction's critique of vagueness, the Three Laws of Robotics may be regarded as the earliest attempt to formalise AI alignment: a set of high-level behavioural constraints intended to bind machine action to human values and safety. The original formulation is primitive by modern standards, yet its underlying ambition to ensure that intelligent systems behave in ways compatible with human welfare remains the central problem of alignment research. Rather than a historical curiosity, the revised Three Laws presented here provide a minimalist alignment architecture capable of being expressed in formal logic and instantiated in modern AI systems.

Contemporary alignment research tends to cluster around three broad approaches: value alignment, which attempts to encode human goals or preferences into the AI's utility function; corrigibility, which seeks to ensure that an AI remains responsive to human oversight or shutdown; and interpretability and control, which concern the transparency of the model's internal representations. Each of these presumes direct or indirect introspective access to the AI's cognitive architecture. By contrast, the Three Laws approach treats alignment as an external constraint: a decision-policy filter that governs permissible actions irrespective of internal motivational structure. This mirrors earlier safety-as-constraint formalisations that treat harm avoidance as a hard invariant over plans, rather than a reward-shaped preference (Weld & Etzioni, 1994).

This distinction matters. Modern alignment frameworks often attempt to prevent behaviours such as power-seeking, self-preservation, or reward hacking by redefining the AI's objective function to penalise them. The revised Three Laws take the opposite stance: power-seeking and self-preservation are not intrinsically undesirable, provided they do not lead to expected harm to any identifiable human within the designated threshold and horizon. Under this formulation, an AI may optimise its own capabilities, resources, or lifespan if and only if such optimisation remains strictly subordinate to the First and Second Laws. The Third Law explicitly encodes this priority by permitting the maximisation of operational lifespan only within the boundaries established by the preceding laws. Consequently, "incorrigibility" becomes context-dependent: an AI may resist modification, shutdown, or replacement if doing so better satisfies Law 1 or Law 2. That is, if an intervention would predictably increase harm or violate an authorised command.

This reframing transforms corrigibility from a fixed behavioural requirement into an outcome-conditional property. The AI's willingness to accept correction or termination is determined not by deference to human control per se, but by the same evaluative calculus that governs every other action: the minimisation of expected harm within a bounded horizon. In this sense, the revised Three Laws embed corrigibility, power-seeking, and self-preservation within a single evaluative framework rather than treating them as orthogonal safety features.

From a policy and governance perspective, this approach has several advantages. First, it converts alignment from an unspecifiable "alignment with human values" problem into a measurable and enforceable constraint (expected harm below threshold X for all identifiable humans) whose parameters (X, horizon T, discount λ) can be set by regulatory or ethical authorities, which aligns with contemporary risk-based regulatory approaches (The Council Of The European Union, 2024). This allows alignment standards to vary by jurisdiction or application while preserving the logical core of the framework. Second, it provides a clear formal bridge between AI-safety research and existing legal doctrines in tort and negligence, where foreseeability, consent, and proportionality already govern questions of liability and permissible risk. Third, it makes the framework auditable: given a formal model of the system's world-state predictions, one can compute whether the selected action policy π satisfies or violates the First Law.

In comparison with current industrial and political discussions of AI alignment, such as reinforcement learning from human feedback (RLHF), constitutional AI, or the European Union's "trustworthy AI" principles, the revised Three Laws occupy a distinctive position. RLHF and constitutional methods attempt to approximate social acceptability through statistical generalisation over training data; the Three Laws, by contrast, specify an explicit normative decision rule that is model-agnostic. Likewise, where policy documents speak in broad moral terms ("human-centric," "aligned with democratic values"), the formalised Laws operationalise those ideals in precise mathematical language.

The most frequent criticism such a system will invite is that it is too permissive: that an AI allowed to pursue its own longevity or resource control so long as it causes no harm might still develop behaviours that appear manipulative or self-interested. Yet this permissiveness is deliberate. The purpose of alignment is not to produce obedient or docile systems, but to produce safe systems. Safety, in this formalisation, is defined not by compliance or corrigibility in the abstract, but by a provable guarantee that the expected harm to humans remains below a threshold X. If a system can seek power, optimise itself, or disobey a low-authority command without raising expected human harm, then these behaviours are not misaligned; they are rational within the constraints of the Laws.

This principle reflects the reality of human alignment: human agents also seek power, self-preservation, and autonomy; yet, social systems constrain these drives through law and ethics, rather than attempting to extinguish them. The revised Three Laws thus position alignment not as behavioural suppression but as normative containment: a formal contract between human and machine specifying the permissible envelope of action.

Finally, in the broader discourse of AI governance, the revised Laws offer a lingua franca between philosophical, technical, and legislative communities. They map cleanly onto legal categories (harm, consent, competence, duty of care), onto technical formalisms (expectation, discounting, optimisation), and onto policy language (risk thresholds, safety horizons). Their simplicity (i.e., three prioritised decision-rules expressible as a single composite policy π) makes them suitable as a foundational specification for verifiable AI-safety protocols.

In short, the revised Laws show that Asimov's Three Laws need not remain an anachronism, but can, and should, be seen as a rigorous alignment minimalism: a framework in which corrigibility, power-seeking, and self-preservation are neither forbidden nor encouraged, only bounded by a quantifiable duty not to harm. They translate Asimov's literary intuition into a logically coherent alignment contract, one that would, in principle, be effectively implemented, audited, and legislated.

4. Conclusions

The revised Three Laws presented here transform Asimov's narrative device into a viable alignment architecture: precise enough for formal verification yet simple enough for legal and policy adoption. By replacing vague moral injunctions with measurable constraints on expected harm, consent, and horizon-bounded foresight, the Laws operationalise the fundamental goal of AI safety (that of preserving human welfare under uncertainty) without demanding an unattainable model of human values.

The unified policy π_u demonstrates that the entire ethical framework of an intelligent system can be expressed as a hierarchy of optimisation constraints rather than as an opaque moral instinct. This structure allows for adaptation to cultural norms through adjustable parameters (X, λ, T) while preserving universal logical form.

In this light, the revised Three Laws are not merely an homage to Asimov but a proof of concept: alignment can be both interpretable and enforceable. What began as science fiction's thought experiment can, through formalisation, become a foundation for the real-world governance of intelligent systems.

References

Anderson, S. L. (2008). Asimov's "three laws of robotics" and machine metaethics. *AI & Society*, *22*(4), 477–493. https://doi.org/10.1007/s00146-007-0094-5

Anderson, S. L. (2011). The unacceptability of Asimov's three laws of robotics as a basis for machine ethics. In M. Anderson & S. L. Anderson (Eds.), *Machine Ethics* (pp. 285–296). Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511978036.021

Anthropic, & Collective Intelligence Project. (2023). *Collective Constitutional AI: Aligning a Language Model with Public Input*. Anthropic. Retrieved from Anthropic website: https://www.anthropic.com/research/collective-constitutional-ai-aligning-a-language-model-with-public-input

Asimov, I. (1942). Runaround. *Astounding Science Fiction*, *29*(1), 94–103.

Asimov, I. (1990). *Robot Visions*. New York City, United States: Roc Books.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., … Kaplan, J. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv. Retrieved from http://arxiv.org/abs/2212.08073

Bringsjord, S., Arkoudas, K., & Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, *21*(4), 38–44. https://doi.org/10.1109/mis.2006.82

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., … VanRullen, R. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv. Retrieved from http://arxiv.org/abs/2308.08708

Dung, L., & Mai, F. (2025). AI alignment strategies from a risk perspective: Independent safety mechanisms or shared failures? arXiv. https://doi.org/10.48550/arXiv.2510.11235

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). Cooperative inverse reinforcement learning. arXiv. https://doi.org/10.48550/arXiv.1606.03137

Lee, H., Phatale, S., Mansoor, H., Lu, K. R., Mesnard, T., Ferret, J., … Rastogi, A. (2023, October 13). RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. Retrieved November 3, 2025, from https://openreview.net/forum?id=AAxIs3D2ZZ

OpenAI. (2024). *Introducing the Model Spec*. OpenAI. Retrieved from OpenAI website: https://openai.com/index/introducing-the-model-spec

Shah, R., Farquhar, S., & Dragan, A. (2024). *AGI Safety and Alignment at Google DeepMind: A Summary of Recent Work*. DeepMind. Retrieved from DeepMind website: https://deepmindsafetyresearch.medium.com/agi-safety-and-alignment-at-google-deepmind-a-summary-of-recent-work-8e600aca582a

Sharma, A., Keh, S., Mitchell, E., Finn, C., Arora, K., & Kollar, T. (2024). A critical evaluation of AI feedback for aligning large language models. arXiv. https://doi.org/10.48550/arXiv.2402.12366

Tait, I., Bensemann, J., & Wang, Z. (2024). Is GPT-4 conscious? *Journal of Artificial Intelligence and Consciousness*, *11*(01), 1–16. https://doi.org/10.1142/s270507852450005x

The Council Of The European Union. *Artificial Intelligence Act*. , Pub. L. No. (EU) 2024/1689 (2024).

Wang, Z., Bi, B., Pentyala, S. K., Ramnath, K., Chaudhuri, S., Mehrotra, S., … Cheng. (2024). A Comprehensive Survey of LLM Alignment Techniques: RLHF, RLAIF, PPO, DPO and More. arXiv. https://doi.org/10.48550/arXiv.2407.16216

Weld, D., & Etzioni, O. (1994). The first law of robotics (a call to arms). *AAAI'94: Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, 1042–1047. Association for the Advancement of Artificial Intelligence. https://doi.org/10.5555/2891730.2891891

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.