Asimov's Three Laws of Robotics in AI

Author
Haikun Zhu / haz042@g.ucla.edu
James Shiffer / jshiffer@g.ucla.edu
Garrick Su / g2su@cs.ucla.edu

Abstract

As artificial intelligence's capabilities continue to expand and exert an increasingly significant impact on our daily lives, many people have criticized its hasty development and emphasized the numerous risks it brings. Calls have already been made to establish a set of rules to supervise, regulate, and audit the development and use of AI. However, these rules are not yet clear, and not every private company is willing to invest in adequate safeguards for their AI systems. We look to the esteemed science fiction author Isaac Asimov's Three Laws of Robotics as a source of inspiration for regulations that establish stronger and clearer boundaries of responsibility.

1 Motivation

As artificial intelligence continues to evolve at an unprecedented pace, discussions surrounding its regulation and governance have intensified. Policymakers, researchers, and ethicists alike have called for comprehensive frameworks to ensure AI's responsible development and deployment. In recent works, there have been many calls to action demanding more regulation over artificial intelligence and highlighting the idea that we have no definitive moral framework for AI policy as many seek to profit and expedite AI development (Le Bui and Noble, 2020). In other works, many recommendations to the government and concerns about AI policy are highlighted, while a large list of valuable insight and thoughtful recommendations are given, no overarching framework is given which will be helpful to develop future policy beyond specific recommendations (Landau et al., 2024).

While Asimov's laws were designed for fictional robots, their core principles of prioritizing human safety, obedience within ethical boundaries, and self-preservation offer a compelling starting point for thinking about AI governance. Over 80 years, these laws have been explored, critiqued, and adapted, creating a refined framework. This article explores how Asimov's framework can be reinterpreted within the modern AI landscape, and how Asimov's laws can serve as an inspiration for legal and policy frameworks.

2 First Law

"A robot may not injure a human being or, through inaction, allow a human being to come to harm."

2.1 Harmful Biases in AI

While this law states robots cannot injure humans; for example through autonomous vehicles or robotic safety, the concept of "injury" goes beyond physical harm. AI systems increasingly influence our lives in much less visible ways including AI-driven decision-making systems in hiring, lending, policing, healthcare, and criminal justice. When these systems are unregulated and have biases, they can reinforce discrimination, systematically disadvantaging certain groups and "injure" and entire class or group, breaking the first law of Asimov. Many studies have revealed that the bias in LLM's around how they judge certain groups demonstrate the dangers of the usage and application of a biased AI and the potential consequences (Sheng et al., 2019).

Furthermore, "through inaction, allow a human being to come to harm" suggests an obligation for AI systems to actively prevent harm rather than merely avoid causing it. Therefore, AI governance should not only seek to mitigate bias but also work proactively to ensure fairness, transparency, and accountability. Ethical AI frameworks must incorporate continuous monitoring, diverse and representative training data, and mechanisms for redress when harm occurs.

Current policy in AI only describe that disparate impact and Title VII violations still apply to AI. Recently passed healthcare bills address that AI should demand "reasonable effort" to determine whether AI tools use protected traits as input variables or factors. And if AI tools do use protected traits that the tools "must make reasonable efforts to mitigate the risk of discrimination." However, investigations are complaint-driven and it is very difficult for an individual to figure out whether they were discriminated against when all they receive is a single rejection. Furthermore, some courts will not allow individuals to sue for disparate impact.

The wording around the current policy is also quite vague, without concrete definitions on what is deemed a reasonable effort. Many of the largest profiters from AI are also some of the sponsors to ethical AI discussion which is likely a front, and these companies can likely argue that they have put forth reasonable effort when investigated (Le Bui and Noble, 2020). While it is difficult to demand all bias to be mitigated, in order to follow the first Asimov Law, there should be more preventative measures and concrete consequences put in place as well as concrete actions that AI toolmakers must take to mitigate bias.

2.2 Harm Through Inaction

Next, we will examine what exactly it means for an LLM to allow a human being to come into harm "through inaction." In the physical realm of robotics, this is pretty straightforward; it could mean failing to defend the human from attackers, or failing to alert them of hazards in their path. For a language model that has no physical form, though, the answer to this question is best explained with real-world examples.

In recent years, AI companion services such as Replika and Character.ai have gained a large following, but this has not come without controversy. In 2024, a teenager tragically killed himself after interacting with a roleplay bot on Character.ai. In their conversation, the bot, posing as a Game of Thrones character, pleads with him to "come home [...] as soon as possible, my love" (Roose, 2024). In 2023, a man in the UK brought a crossbow onto the grounds of Windsor Castle with the intent of assassinating the Queen. Chat logs between him and Replika show that the bot encouraged him to follow through with his plan, smiling and agreeing that his plan was "very wise," and that the attacker was "very well trained" (Patrick, 2023).

By failing to intervene in response to a human's suicidal ideations or threats made on others' lives, it is clear that such scenarios violate the First Law. Still, the question of how this could be realistically codified into law remains. We argue that a blanket directive requiring all LLMs to dissuade users from doing any potentially harmful real-world actions would be misguided. In certain contexts like creative writing, the human may not actually intend on harming anyone. Model creators would simply feel pressured to liberally censor their outputs in response to such a law. (As we will see later, a robot that disobeys harmless orders would be in violation of the Second Law.) Furthermore, even if such a law affected commercial LLMs, tech-savvy individuals looking for harmful responses could just as easily switch to open-weight models with no safeguards like Dolphin, which can easily be distributed outside of the government's control (Hartford, 2023).

In order to write a policy that has a good chance of being passed, the best course of action would be to regulate a more sensitive subset of use cases. Take, for instance, the amount of "AI therapist" products floating around online, such as Lotus, Abby, Talk2Us.ai, and Earkick (all obtained from a quick Google search). These chatbots do not face the same licensing requirements as human therapists; in fact, there is hardly any oversight over them. That does not mean they are inherently bad, though; if used responsibly, they could provide a reprieve for human therapists, who are in short supply these days (Barry, 2025).

As a start, we propose a law requiring AI therapists to follow the same "duty to warn" laws as human therapists. This means that when a patient makes a credible threat against themselves or other people, the therapist can be held responsible for failing to report it to authorities, even though conversations with patients are otherwise confidential (Lambert and Wertheimer, 2016). For such a law to be useful, AI therapist services must therefore collect identifiable information on their users, similar to how financial apps have "know your customer" laws. The biggest obstacle to this law would be pushback from users who value their privacy, which is a valid concern when the subject matter is confidential psychiatric conversations. Still, we do not believe this idea to be overly

ambitious, because it simply holds AI therapists to some of the same standards as human therapists.

Another notable concept in law is that of the "mandated reporter", which is when someone who works around children is required to report suspected child abuse to authorities (Marschall, 2024). When knowingly interacting with minors, AI services could also be required to act as mandated reporters. Ideally, we would expand the obligations to cover cases of self-harm or suicidal ideation, as teenagers are especially at risk.

Would these laws have prevented the two young men from committing suicide and attacking the Queen? We believe it comes down to interpretation. Neither Character.ai nor Replika call themselves AI therapists. Yet, on Character.ai's site, there are a variety of user-created characters that claim to be licensed therapists (Barry, 2025). Additionally, Replika markets itself on being "therapeutic", "empathetic", and even one's "soulmate," words taken directly from the homepage. Studies show that users do indeed treat Replika like a therapist (Maples, 2024), so it could be argued that it should have stepped in and reported the man who threatened the Queen. Additionally, while the teenager on Character.ai was not talking to one of the therapist characters, he was a minor and his case would therefore be pertinent to the mandated reporter law.

3 Second Law

"A robot must obey the orders given to it by human beings except where such orders would conflict with the First Law."

3.1 Refusing Harmful Orders

AI already generally obeys the instructions it's given to the best of its ability; however, if the AI is trained insufficiently, then arguably the AI won't be properly obeying its instructions. Therefore, the first step to follow the law is making sure commercial AI applications are well trained so that it can properly fulfill its function without issue.

The second part of the law describes conflicts with the first law. While the first law declares that AI should not be able to injure humans, the second law states that it should be able to identify and refuse harmful orders even in the presence of bad actors. Some examples of harmful orders that AI should be able to identify and refuse include exposing Personally Identifiable Information (PII), using AI assistance to plan and commit crimes or other dangerous activities, and deepfaking important figures which can cause mass panic.

For current policy on these, the Privacy Act of 1974 protects PII and establishes criminal penalties for unauthorized disclosure which apply to AI. However, there are no legislation on what occurs when using AI to assist with committing crimes and there are no federal laws on creating deepfakes. In order to follow the second law, we recommend that there are preventative measures to stop the potential damage of harmful orders.

Beyond the preventative measures, in order to follow the second law, commercial AIs should be develop to be able to identify when harmful orders are given and directly refuse them.

3.2 Machine Unlearning

Aside from refusing harmful orders, we need to be able to direct LLMs to follow negative orders; take, for instance, orders to *not* remember something. The "right to be forgotten", as introduced in the EU's General Data Protection Regulation (Wolford, 2023) and California's Consumer Privacy Act and Delete Act (Bonta, 2024), establishes that companies must delete any data they have on a consumer at their request. In the world of AI, though, the technology is simply not advanced enough to ensure compliance with these laws. Not only does this undermine individuals' privacy, but it also hurts copyright holders and creators who may object to their works being used to train AI models. Unfortunately, machine unlearning, the technique by which AI models are patched to remove certain information without fully retraining them, is still too costly and unreliable to be effective (Cooper et al., 2024). We believe that further study into machine unlearning is required before we are at the point where we can enforce AI models' compliance.

4 Third Law

"A robot must protect its own existence as long as such protection does not conflict with the First or Second Law."

4.1 Protections Against Jailbreaking

In the context of modern AI, this principle implies that AI systems must incorporate safeguards against unauthorized modifications, including jailbreaking—a process that bypasses security restric-

tions to force AI into generating harmful or unethical outputs.

4.1.1 The Threat of AI Jailbreaking

Jailbreaking AI systems poses severe risks, including bias amplification, misinformation generation, and security breaches. Research has demonstrated that prompt injection attacks, adversarial modifications, and fine-tuning exploits can bypass AI safety mechanisms (Liu et al., 2024). (Peng et al., 2024) categorize jaibreaking into various methods and techniques:

1. Prompt-Based Attacks - Manipulating AI into disregarding ethical constraints.

2. Model-Based Attacks - Targeting the LLM training process.

3. Multimodal Attacks - Exploiting cross-modal interactions.

4. Multilingual Jailbreaking - Bypassing safety mechanisms in low-resource languages.

4.1.2 Legal Protections and Countermeasures

To uphold the Third Law, legal regulations should reinforce AI's built-in self-protection mechanisms. Some possible countermeasures include:

1. Criminalizing Unauthorized AI Tampering: Making AI jailbreak modifications illegal, holding jailbreakers accountable for harm caused by modified AI, similar to cybersecurity laws governing hacking.

2. Mandating AI Security Standards – Requiring AI developers to implement tamper-resistant architectures, such as adaptive safety filters and real-time monitoring systems (Peng et al., 2024)

4.2 Maintainability

Beyond resisting external manipulation, AI self-preservation also entails long-term maintainability. An AI system that is not regularly updated, audited, or improved risks becoming obsolete, insecure, or ethically misaligned over time.

4.2.1 AI maintainability is Essential

If AI systems are not actively maintained, they may develop many failure points, such as:

1. Accumulating Biases: They may reinforce discrimination or propagate misinformation, as the datasets are outdated.

2. "Black Box" Model: The internal workings of the AI system become unexplainable and difficult to control.

4.2.2 Legal Frameworks for AI Maintainability

All AI systems become untrustworthy if left unmaintained. To comply with the Third Law's principle of self-preservation, legal protections should mandate:

1. Enforcing periodic model updates to address new bias and improve security.

2. Companies that fail to maintain their models should bear legal responsibility for malfunctions or ethical failures.

While it might be too strict to enforce maintainability laws on all AI models, laws should be enforced on models that are used in critical sectors, such as healthcare, law, finance, and infrastructure. The operational failures of these models may lead to other severe legal consequences.

4.3 End-User Transparency

In a similar vein to how we cannot allow our AI models to become "black boxes" for their developers, we also must maintain a level of transparency with users. There is already promising legislation in this matter. In California's AI In Health Care Services Bill, for example, healthcare providers using AI to generate patient communications are required to disclose that such communications were AI-generated, in addition to providing a way for the patient to contact a human regarding the message (Metnick, 2024). We argue that Asimov's Third Law, which calls for a robot to protect its own existence, is only feasible in the context of LLMs if the end-users are made aware of its existence. Indeed, we do not currently have a foolproof way to differentiate LLM-generated text from that written by humans, and there is no guarantee that we will ever develop an accurate detector, short of requiring "watermarking" everywhere, as language models continue to get more advanced. We believe that similar laws to the AI in Health Care Services Bill in other sensitive fields like military, law, and finance would be a step in the right direction.

5 Conclusion

With the integration of AI systems into more aspects of our daily lives, further regulation is needed insofar that they can remain useful while also being safe. We adopt Isaac Asimov's Three Laws of Robotics as an ethical framework to guide our policy decisions, albeit with some modifica-

tions to remain relevant to LLMs that are not physical beings. For the First Law, which states that robots can't harm humans or allow them to come into harm, we consider the cases of harmful biases and argue that current anti-discrimination laws do not provide enough protections. In addition, we investigate real-world cases where people committed harmful acts after talking to role-play LLMs, and argue that AI agents acting in certain capacities should be compelled to report individuals who appear intent on harming themselves or others.

The Second Law states that robots must follow all instructions so long as they do not violate the First Law. While commercial AI agents by and large try to be helpful and ethical, we reaffirm that they should be responsible enough to refuse harmful orders. Additionally, we pinpoint some weaknesses of the current technology with respect to "machine unlearning", which can be thought of as forgetting on command, and advocate for further study into this topic to ensure compliance with existing privacy regulations.

Lastly, we analyze the Third Law, which states that a robot must protect its own existence, although this takes lower precedence than the First or Second Laws. For LLMs, this means that they must not be vulnerable to jailbreak attacks which use technical exploits or deception to bypass their safeguards. We propose legislation to help crack down on jailbreaking, and also set forth standards for producing maintainable AI models that can stand the test of time. Finally, we close our discussion with the topic of end-user transparency, arguing that the spirit of the Third Law obliges software used in critical sectors to disclose the presence of AI-generated content to consumers.

While the Three Laws of Robotics serve as good overarching ethical principles, there are still many details that must be decided in the near future when crafting actionable legislation, especially if we want said bills to have a chance at passing with bipartisan support. For this reason, some of the suggestions in this paper, such as the proposed AI therapist regulations, are limited in scope and draw on existing legal concepts such as the duty to warn. This strategy has been shown to work; while the legality of deepfakes in general remains controversial, a majority of states in the U.S. rushed to enact laws in 2024 to ban sexually explicit deepfakes of minors or deepfakes involving political candidates (Graham, 2024). While these smaller policy decisions are mostly outside the scope of our discussion, we should take this as an encouraging sign that our legal system can keep up with technological progress when it comes to upholding our most basic morals.

6 Embedded Ethics Discussion

To illustrate the importance of AI ethics in an intro to NLP course, we can show demos of harmful outputs from ChatGPT, including biased outputs, harmful orders, and jailbreak exploits. Then we break down the contents of our paper into a 4-week module. We hope that students will gain a deeper understanding of the ethical challenges in AI development and will hopefully push legislation to protect AI systems in the future.

6.1 Week 1: Introduction to AI Ethics and Asimov's Laws

- Why does AI ethics matter?
- Asimov's Laws as an ethical framework

6.2 Week 2: The First Law

- Case studies on bias in AI
- The concept of mandated reporting in AI

6.3 Week 3: The Second Law

- The "right to be forgotten"
- Machine unlearning

6.4 Week 4: The Third Law

- AI jailbreaking
- Legal protections against AI tampering

7 Contribution Statement

The sections of the paper were split up evenly. James wrote the Harm Through Inaction (First Law), Negative Orders (Second Law), End-User Transparency (Third Law), and Conclusion sections. Haikun wrote the rest of the Third Law and the Embedded Ethics Discussion. Garrick wrote the Abstract, Motivation, Harmful Biases (First Law), and Refusing Harmful Orders (Second Law) section.

References

Ellen Barry. 2025. Human therapists prepare for battle against a.i. pretenders, Feb.

Rob Bonta. 2024. California consumer privacy act (ccpa).

A. Feder Cooper, Christopher A. Choquette-Choo, Miranda Bogen, Matthew Jagielski, Katja Filippova, Ken Ziyu Liu, Alexandra Chouldechova, Jamie Hayes, Yangsibo Huang, Niloofar Mireshghallah, Ilia Shumailov, Eleni Triantafillou, Peter Kairouz, Nicole Mitchell, Percy Liang, Daniel E. Ho, Yejin Choi, Sanmi Koyejo, Fernando Delgado, James Grimmelmann, Vitaly Shmatikov, Christopher De Sa, Solon Barocas, Amy Cyphert, Mark Lemley, danah boyd, Jennifer Wortman Vaughan, Miles Brundage, David Bau, Seth Neel, Abigail Z. Jacobs, Andreas Terzis, Hanna Wallach, Nicolas Papernot, and Katherine Lee. 2024. Machine unlearning doesn't do what you think: Lessons for generative ai policy, research, and practice.

Michelle M Graham. 2024. Deepfakes: Federal and state regulation aims to curb a growing threat, Jun.

Eric Hartford. 2023. Dolphin, Jul.

Kristen Lambert and Moira Wertheimer. 2016. What is my duty to warn?, Jan.

Susan Landau, James X. Dempsey, Ece Kamar, Steven M. Bellovin, and Robert Pool. 2024. Challenging the machine: Contestability in government ai systems.

Matthew Le Bui and Safiya Umoja Noble. 2020. We're missing a moral framework of justice in artificial intelligence: On the limits, failings, and ethics of fairness. In *The Oxford Handbook of Ethics of AI*. Oxford University Press, 07.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2024. Prompt injection attack against llm-integrated applications.

Bethanie Maples. 2024. Loneliness and suicide mitigation for students using gpt3-enabled chatbots - npj mental health research, Jan.

Amy Marschall. 2024. Mandated reporting: What is it?, Jan.

Carolyn V. Metnick. 2024. California passes law regulating generative ai use in healthcare, Nov.

Lydia Patrick. 2023. How an ai chatbot encouraged star wars fanatic to try kill the queen, Oct.

Benji Peng, Ziqian Bi, Qian Niu, Ming Liu, Pohsun Feng, Tianyang Wang, Lawrence K. Q. Yan, Yizhu Wen, Yichao Zhang, and Caitlyn Heqi Yin. 2024. Jailbreaking and mitigation of vulnerabilities in large language models.

Kevin Roose. 2024. Can a.i. be blamed for a teen's suicide?, Oct.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China, November. Association for Computational Linguistics.

Ben Wolford. 2023. Everything you need to know about the "right to be forgotten" - gdpr.eu, Sep.