POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Outgrowing Asimov: The Inadequacy of the Three Laws in AI Development

Laurea Magistrale in
Computer Science and Engineering - Ingegneria Informatica

Alberto Papiri, 10566115

Academic year:
2022-2023

Abstract: This paper presents a comprehensive analysis of some of the limitations of Asimov's Laws of Robotics in ensuring the safe and ethical development of Artificial Intelligence (AI) technologies The three laws, conceived by Isaac Asimov in his science fiction literature, have been and are still influential in the ethical discourse on AI. However, as AI technology continues to evolve, these laws fall short of addressing the complexities of real-world ethical dilemmas. This paper critically explores the inherent ambiguities in Asimov's laws, their lack of clarity in defining 'harm' and 'human being,' and the difficulties encountered in implementing these abstract principles in AI systems. Drawing from various contexts like the Trolley Problem, AI driving systems, and military applications, the paper illuminates the growing discrepancy between Asimov's laws and the ethical challenges brought about by AI advancements. The paper concludes with a call for a more comprehensive, globally collaborative effort in developing ethical guidelines for AI that reflect the complexities of the contemporary socio-technical landscape. This paper aims to deepen the understanding of AI ethics and highlight the necessity for a robust ethical framework that can effectively guide the development and deployment of increasingly autonomous technologies.

Key-words: Robotics, Philosophy, Asimov's Laws

1. Introduction

In this rapidly evolving digital era, the ubiquitous impact of Artificial Intelligence (AI) on various aspects of our lives is undeniable. As promising as these technological advancements are, they present a pressing need to develop ethical guidelines for AI's responsible creation and deployment. One framework that has formed a cornerstone in this discussion is Isaac Asimov's Laws of Robotics. Introduced in his science fiction narratives of 1942, Asimov's Laws were among the earliest attempts to create a moral compass for AI. However, as we explore the complexities and challenges of real-world applications of AI, the limitations of Asimov's laws have become increasingly apparent.

Being aware that they were developed for a narrative mechanic, we assume they are applied to the real world, which is not too far from reality, at least for the consideration that the audience has had in recent decades.
What we claim is that Asimov's Laws are insufficient for a safe development of AI agents, which are systems that are designed to perceive their environment, make decisions and take actions to achieve a specific goal or

set of goals. For consistency with Asimov's ideas, we will make a second assumption namely that we do have AI agents complex enough to apply these laws to. The paper starts by revisiting Asimov's Laws and their theoretical basis in *Section 2*. This exploration is vital for understanding the fundamental assumptions of these laws and their bigger limitations. Drawing from various academic and scientific resources, we shed light on the inherent ambiguities and inability of these laws to grapple with complex ethical dilemmas. We present a range of scenarios, from common day-to-day interactions with AI to complex hypothetical situations, to demonstrate these constraints effectively.

In then, our exploration leads us to engage with renowned thought experiments, such as the Trolley Problem, to highlight the dilemmas that arise when AI has to make decisions that potentially impact human lives. Hence, we present an application of this scenario in AI-enabled autonomous vehicles, bringing into focus the challenges in programming these vehicles to handle unavoidable accidents. Furthermore, we touch upon the critical area of AI application in the military field in . We discuss the ethical issues around autonomous weapons systems, the question of accountability, and the prospect of escalating AI warfare. We argue that Asimov's Laws, in their current form, fall short of addressing these concerns adequately.

The reader will find this paper valuable not just for its critique of Asimov's Laws but also for its advocacy for a more nuanced ethical framework for AI technologies. By drawing attention to the practical challenges of AI ethics in diverse contexts, this paper emphasizes the urgency for a broader and more comprehensive ethical architecture. In a brief conclusion in *Section 5*, we propose that this new framework should be shaped through global collaboration, embracing transparency and inclusivity to navigate the socio-technical challenges presented by AI's future.

2. Asimov's Laws and their well-known flaws

When in 1942 Isaac Asimov wrote a series of stories and novels about robots, he gave birth to some of the most recognizable pieces on the subject[1]. Understanding pretty soon that his novels need some expedient to protect humans from almost any perceivable danger, he decided to implant all his robots with the Three Laws of Robotics, that state as follow:

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 Law.*

As pointed out by Robin Murphy from Texas A&M University and David Woods from Ohio State University[8], Asimov's laws are based on functional morality, which assumes that robots have sufficient agency and cognition to make moral decisions. This gap helped Asimov to create situations functional to his narrative, highlighting the "classical" shortcomings, most of them due to the conflict between different laws. The causes of these conflicts are to be found mostly in the ambiguity of the words used in the declaration of the laws. In particular, the first law as it deals with harm to humans. It fails also because of complicated ethical problems that are too complex to have a simple yes or no answer[11]: a robot will be frequently asked to balance the harm and benefits in complex scenarios, having no guideline on how to weight one harm against another. What can appear to be cruel in the short run can actually become a kindness in the longer term, just think of the task of children's educators, who are often faced with such situations.

Moreover, the term "harm" is vague, leaving room for interpretation and loopholes. For example, it is not clear what counts as "injury" or "harm" to a human being, how can a robot determine the consequences of its actions or inactions. It's not clear if only physical harm counts, or emotional or psychological harm also fall under this purview.
Also the concept of "human being" can be ambiguous and dependent on eras and cultures. During the 14th century, slaves were often viewed more as cattle than as human beings. In modern times, the rights to life of a fetus are frequently the focus of heated debates. However, looking into the future, consider a scenario where an expectant woman, due to a particular medical condition, faces a high likelihood of mortality during childbirth. In such a case, her AI doctor will have to decide if recommend an abortion or not. It's critical to acknowledge that although the woman statistically has a better survival chance if she has the abortion, the fetus, once it is born, has a potentially longer life to lead. Thus, irrespective of the choice, the AI, in one form or another, will inevitably inflict harm on a human being[3].

Even if we consider the previous issues resolved, a big one still occupies the field. Those abstract and ambiguous concepts are difficult to implement as a piece of software. Programmers should be able to encode that in a digital way since computers deal with logical or numerical problems and results[7].
A new approach can be tried thanks to the newest machine learning technologies: tell the robots all the useful notions by natural language. In this case we have to face new problems: the first one regards the chosen language to formulate the laws that could not correspond to the robot one. Natural Language Processing (NLP) and its declinations may indeed be the solution to this problem, although giving the right meaning to idioms peculiar to a language remains a challenge.

A second problem is raised by Hogan in his work[5]: trying to simulate the process of inculcating certain concepts into a robot, shows how an exponential amount of sub-concepts have to be introduced recursively, making such a process impossible. He also explains that humans use language to supply only that information which we suppose is not obvious, on the other hand for a machine nothing is obvious. *"Its entire reality is described by what it has been given explicitly. Nothing else exists."*[5].
The main point is that humans can communicate with each other in this way because they have all had the same experience of forming their perceptions and associations in the same universe of physical space, populated by the same objects, and growing up in the same cultural environment. Therefore, it is unfeasible to codify every aspect of human behavior in an attempt to recreate a moral agent that perfectly mirrors a human being.

3. AI driving systems

We introduce now the famous Trolley Problem, a philosophical thought experiment first formulated by Philippa Foot in 1967, that can help us to highlight some dilemmas also present in robot behavior. Specifically, in the standard trolley dilemma, five workers working on the tracks are expected to be hit and killed by a runaway train with failed brakes. However, by pulling the lever to divert the runaway trolley onto the sidetrack, one can save the lives of the five workers in exchange for the life of another worker[12].
The problem challenges the principles of utilitarianism, deontology, and the doctrine of double effect, but it also raises questions about how robots would act in such a situation, especially if they follow Asimov's laws of robotics. According to the First Law, a robot cannot, through action or inaction, allow a human being to come to harm. In the Trolley Problem, any choice made results in harm to human beings, creating an unsolvable conflict for a robot programmed with these laws.
A possible solution to this problem could come from the "Zeroth" Law that Asimov later added above all the others. It states: *A robot may not harm humanity, or, by inaction, allow humanity to come to harm.*
It could be a potential justification for pulling the lever. However, it is still unclear how this rule would quantify the value of human lives or decide what constitutes greater harm to humanity.

Asimov's Laws, when faced with the Trolley Problem, fail to accommodate for more complex decision-making that goes beyond binary choices. Ethical decisions often exist in shades of gray and require more sophisticated ethical reasoning than the laws offer. The Trolley Problem is a simplified representation of the ethical dilemmas that AI may face in real-world situations, such as AI driving systems.
As Hübner and White point out in their work[6], these technologies have great potential to drastically reduce the number of crashes on our roads, particularly as factors relating to human error, including speeding, negligence, drunk and distracted driving. Yet, self-driving cars cannot be 100% safe. The link between the thought experiment and this topic has been discussed by many journalists and philosophers in the last years. The authors use the Trolley Problem to discuss how to program the so-called crash algorithms for autonomous cars and other autonomous vehicles in different accident-scenarios, especially where harm is unavoidable.

To give an example, imagine a self-driving car that is driving on a road with a cliff on one side and a wall on the other. Suddenly, a group of pedestrians appears on the road, leaving the car with two options: either steer and crash into the wall, killing the passengers, or continue and hit the pedestrians, killing them. A self-driving car might choose the last option in order to save the passengers. This could be justified by appealing to the first law, which prohibits harming a human being or allowing a human being to come to harm. By steering and crashing into the wall, the car would directly cause harm to its passengers, who are under its care and authority.
Hübner and White state that the question of how self-driving cars should be programmed to handle accidents is inevitable, whether it is clearly defined within the vehicle's programming or subtly influenced by its overall control system. As a result, philosophers are increasingly concentrating on the ethical implications of crash algorithms. Although still in its early stages, they're also suggesting potential ethical guidelines that could instruct us on how autonomous vehicles should respond in the event of unavoidable accidents.

Some other researchers[9] have identified key differences between the ethics of accident-algorithms used in autonomous vehicles and the classic Trolley Problem. These differences primarily revolve around three main dissimilarities: the overall characteristics of the decision-making situation, the role of moral and legal accountability, and the knowledge level of those making the decisions.
We believe that, with the right assessments of the Trolley Problem, it is possible to address the issues of AI driving systems, in particular with respect to Asimov's First Law implications, as seen before. Naturally, this is a hypothetical thought-experiment where choices are deliberately limited to a few unique, one-time options, and the decision-maker possesses complete knowledge. However, these problems serve as theoretical instruments to explore ethical instincts and theories, as well as assess new technologies in relation to traditional ethical dilemmas.

The second law is not without its criticality in this case either. A recent story gives a representation of it: a man that used a car to run over a shooter on a bridge that was randomly shooting and killing people, according to authorities[4]. The driver was considered heroic by having stopped that shooter by running down him. If the Asimov second law was programmed into the AI driving system of self-driving cars and a passenger ordered the AI to run over a shooter, presumably it would refuse to do so. This is obvious because the instruction would harm a human but we know that in this case it seemed worthwhile to do so.

4. Military implications

One of the first scenarios that comes out more frequently when people are asked to think about dangerous applications of AI and robotics is related to the military field. Artificial Intelligence is increasingly finding applications in military contexts, from autonomous weapons systems (AWS) to "fire-and-forget" missiles, and drones loaded with explosives. These developments promise efficiency and strategic advantages but also raise serious ethical issues. The prospect of AI making life-and-death decisions on the battlefield, the issue of accountability for AI's actions, and the potential for escalating AI warfare are but a few of these concerns. At a first glance, the First Law should totally avoid the use of these technologies, being them certainly harmful. But a conflict is, at least, a two-sided system and the use of autonomous weapons could guarantee safety and defense to one of the involved parties: invoking the principle of the Zeroth Law, as already seen, the question would still be open.

Moreover, recent proposed scenarios by some armies all over around the world suggest that the direction of this industry is far from complying with Asimov's Laws and solutions must be found in order to build a regulamentation. Giving the permission for a military robot to fire upon anything moving without human permission shows a bigger problem than the act in itself: artificial agents, of which AWS are one example, cannot understand the value of human life. A human combatant cannot transfer his privileges of targeting enemy combatants to a robot[10]. The increasing difficulty to hold military personnel morally and legally responsible for war crimes or fatal accidents when autonomous weapon systems are present. This is known as the "responsibility gap" problem. The US, in particular, is using drones for targeted killings in foreign countries. The accountability, not to mention morality, of these actions is still being ferociously debated. But the two authors Barthelmess and Furbach[2] imply that humans are still ultimately responsible for these killings and that international law, rather than Asimov's laws, should be able to cope with issues that arise, or adapt to do so.

However, Lee McCauley from University of Memphis[7] claims that *"No robotic apocalypse is coming"*, explaining how Asimov's Laws will avoid large scale accidents. Mainly because today robots are not "smart" enough for doing so. But what we showed in the previous paragraph, about current warfare scenarios, we believe it should alert us and make us ready with the right tools to face increasingly humanoid robots. In support of our findings, in his own work he collects many opinions from experts that highlight pretty much the same flaws that we discussed in *Section 2*.

5. Conclusion

In conclusion, while Asimov's Laws of Robotics provided a foundational framework to address the robots' possible threats, they are not sufficient to navigate the intricacies of real-world scenarios that AI may encounter, especially as we integrate these systems more deeply into our societies. The laws' inherent ambiguities, inability to deal with complex ethical decisions, the practical difficulties in their implementation, as well as the lack of accommodation for cultural variations limit their effectiveness.

Furthermore, ethical dilemmas such as the Trolley Problem and the rising use of AI in autonomous vehicles and military applications underline the need for a more sophisticated ethical framework. Hence, it becomes necessary to develop a broader, more comprehensive ethical structure that considers these complexities and ensures the safe, responsible development and application of AI technologies. This new ethical architecture should be formed through a globally collaborative effort, as the document Ethically Aligned Design by IEEE tries to do, guiding the implementation of these technologies through General Principles as respect of human rights and transparency.

References

[1] Isaac Asimov. Runaround. *Astounding science fiction*, 29(1):94–103, 1942.

[2] Ulrike Barthelmess and Ulrich Furbach. Do we need asimov's laws?, 2014.

[3] Hans A. Gunnoo. Asimov's laws of robotics, and why ai may not abide by them. https://towardsdatascience.com/asimovs-laws-of-robotics-and-why-ai-may-not-abide-by-them-e6da09f8c754, 2019. (Accessed on 07/04/2023).

[4] Christine Hauser and Michael Levenson. Soldier stopped shooting by driving into gunman, kansas police say - the new york times. https://www.nytimes.com/2020/05/28/us/fort-leavenworth-centennial-bridge-shooting.html, 2020. (Accessed on 07/03/2023).

[5] James P. Hogan. Asimov's laws. In Michael G. Hinchey, James L. Rash, Walter F. Truszkowski, Christopher Rouff, and Diana Gordon-Spears, editors, *Formal Approaches to Agent-Based Systems*, pages 260–263, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[6] Dietmar Hübner and Lucie White. Crash algorithms for autonomous cars: How the trolley problem can move us beyond harm minimisation. *Ethical Theory and Moral Practice*, 21(3):685–698, June 2018.

[7] Lee McCauley. AI armageddon and the three laws of robotics. *Ethics and Information Technology*, 9(2):153–164, August 2007.

[8] Robin Murphy and David D. Woods. Beyond asimov: The three laws of responsible robotics. *IEEE Intelligent Systems*, 24(4):14–20, 2009.

[9] Sven Nyholm and Jilles Smids. The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice*, 19(5):1275–1289, 2016.

[10] Michael Skerker, Duncan Purves, and Ryan Jenkins. Autonomous weapons systems and the moral equality of combatants. *Ethics and Information Technology*, 22(3):197–209, February 2020.

[11] Chris Stokes. Why the three laws of robotics do not work. *International Journal of Research in Engineering and Innovation*, 02(02):121–126, 2018.

[12] Judith Jarvis Thomson. The trolley problem. *The Yale Law Journal*, 94(6):1395–1415, 1985.