The Ethical Dilemmas of the "Three Laws of Robotics" in Isaac Asimov's Runaround (1942) and Little Lost Robot (1947)

المعضلات الأخلاقية لقوانين الروبوتات الثلاثة في قصتي إسحاق أسيموف "دائرة مفرغة" (١٩٤٢) و "الروبوت الصغير المفقود (١٩٤٧)

Dr. Mona Gad Sayed Gad
Lecturer, Department of English Language
Faculty of Arts, Delta university for Science and Technology
د. منى جاد سيد جاد
مدرس بقسم اللغة الإنجليزية
كلية الآداب، جامعة الدلتا

The Ethical Dilemmas of the “Three Laws of Robotics” in Isaac Asimov’s *Runaround* (1942) and *Little Lost Robot* (1947)

Abstract:

This paper examines the ethical dilemmas presented by Isaac Asimov’s *Three Laws of Robotics* in his stories *Runaround* (1942) and *Little Lost Robot* (1947). The Laws are analyzed and reevaluated within the framework of the ethical theories of Immanuel Kant’s *deontology* and Jeremy Bentham’s *utilitarianism*. The analysis demonstrates the ethical conflicts between deontology’s rigid adherence to universal moral absolutes and utilitarianism’s emphasis on maximizing societal welfare. This is through illustrating Asimov’s critical insights into contemporary debates on artificial intelligence ethics and regulation, prompting a re-evaluation of human responsibility, human-robot trust, and the boundaries of robotic autonomy. The stories reveal the limitations of Asimov’s Laws in addressing real-world complexities, exposing their inability to guarantee consistent ethical behavior in artificial intelligence systems. Furthermore, this study introduces a novel perspective on the interplay between ethical theory and speculative fiction, underscoring the practical value of Asimov’s narratives in shaping forward-thinking approaches to robotic legislation and ethical programming.

Keywords: Deontology, Utilitarianism, AI legislation, Isaac Asimov

المعضلات الأخلاقية لقوانين الروبوتات الثلاثة في قصتي إسحاق أسيموف "دائرة مفرغة" (١٩٤٢) و "الروبوت الصغير المفقود" (١٩٤٧)

المخلص:

تتناول هذه الورقة البحثية المعضلات الأخلاقية التي أظهرتها قوانين الروبوتات الثلاثة في قصتي إسحاق أسيموف "دائرة مفرغة"(١٩٤٢) و "الروبوت الصغير المفقود" (١٩٤٧). وتعتمد الدراسة علي تحليل القوانين الثلاثة، وإعادة تقييمها في إطار نظريتي "الأخلاقيات المطلقة" لإمانويل كانت و"النفعية" لجيرمي بنتام. ويوضح تحليل القوانين الثلاثة التناقضات الأخلاقية المتكررة بين الالتزام الصارم بالقوانين والمبادئ الأخلاقية المطلقة، وتركيز النفعية على تعظيم رفاهية المجتمع؛ وذلك من خلال تسليط الضوء على رؤى أسيموف النقدية للنقاشات المعاصرة حول أخلاقيات الذكاء الاصطناعي وتنظيمه وحوكمته، مما يدعو إلى إعادة تقييم مسؤولية الإنسان، الثقة بين الإنسان والروبوت، وحدود استقلالية الروبوتات. وتكشف القصص عن القيود التي تعاني منها قوانين أسيموف في مواجهة التعقيدات الواقعية، وكذلك تُظهر عدم قدرتها على ضمان سلوك أخلاقي ثابت في أنظمة الذكاء الاصطناعي. وعلاوة على ذلك، تقدم الدراسة منظورًا جديدًا حول التفاعل بين النظرية الأخلاقية والخيال التأملي بوصفه أداة منهجية، مسلطة الضوء على القيمة العملية لأعمال أسيموف في صياغة مقاربات استشرافية لتشريعات الذكاء الاصطناعي والبرمجة الأخلاقية.

الكلمات المفتاحية: الأخلاقيات المطلقة، النفعية، تشريعات الذكاء الاصطناعي، إسحق أسيموف

The Ethical Dilemmas of the "Three Laws of Robotics" in Isaac Asimov's *Runaround* (1942) and *Little Lost Robot* (1947)

As machines increasingly assume roles in decision-making processes, the ethical implications of their programmed directives become crucial. This raises inquiries into the essence of morality as it relates to *artificial intelligence* [AI] and the implications of their choices on human lives. According to De Cooman and Petit (2020), legal experts approach the legal and regulatory challenges of AI through four distinct mental frameworks. These are: "the black letter law model," "the emergent model," "the ethical model," and "the risk regulation model" (p. 1). Since "ethics are the part of practical philosophy that deals with moral dilemmas" (p. 5), the juxtaposition of *deontological rigidity* against *utilitarian flexibility* represents the essence of the debate, prompting a reevaluation of what it means to act ethically in an age where machines may wield significant influence over moral outcomes.

Therefore, the controversy surrounding these ethical theories not only illuminates foundational principles but also challenges us to consider the future landscape of moral reasoning in an increasingly automated world (Agarwal & Pareek, 2022, p. 4633, Körner & Deutsch, 2023, p.1512). The critical question that arises now is: which ethical criterion/a will control the increasing use and programming of AI? To find an answer to this question is to study *The Three Laws of Robotics* that were first introduced in Asimov's *I, Robot* (1950), and are still the core principles in modern AI reality. This entails the discussion of certain points, such as human-Robot trust, showing how far Asimov's narratives foreshadow trust issues in AI-powered technologies, AI decision-making, and legislative implications, referring to current debates about AI regulation and governance compared to Asimov's fictional constructs. These points, accordingly, will be elaborated through the study of the ethical dilemmas provided in Asimov's works, and the conflicting -sometimes harmonized- expression of deontological and utilitarian theories in these narratives.

The exploration of *Deontology* and *Utilitarianism* in fiction reveals a complex interaction of different ethical models. *Deontology* is

illuminated and clarified by Immanuel Kant and is known as "The Theory of Duty"[1]. According to it, any action gains its moral value out of the *sense of duty* that accompanies its performance (Audi, 2021, p. 60). It also asserts that the morality of an action is determined not only by its outcomes or results but rather by its alignment with universal ethical principles that remain constant and unwavering, regardless of the situational contexts that might seek to influence ethical evaluations (Henry& Jonathan, 2024, p. 4).

Therefore, deontology suggests that moral actions stem directly from a sincere commitment to these ethical imperatives, without considering the potential benefits or harms that might ensue from specific actions (Audi, 2021, p. 61; Bartneck, et.al, 2021, p. 20; Henry& Jonathan, 2024, p. 3). This, in turn, means that violating the legislation is immoral, even if it leads to greater overall utility. In essence, the term is considered part of modern philosophy that refers to "a commitment to the duty and obligation to do something" (p. 156) and the accurate application of rules and laws.

Henry and Jonathan (2024) provide a comprehensive exploration of the Kantian *deontology* through the concept of firm non-negotiable "moral absolutes"- the commands- as fundamental guidelines for moral decision-making (p. 3). Their main impulse behind posing these absolutes, like most deontologists, is the complex ethical challenges that result from the tension between moral principles, situational contexts, and potential outcomes. These absolutes are emphasized through five key characteristics: (1) "Objective Standards" that are the set of laws to consult in each situation; (2) "Moral Certainty" derived from the clear directives of what is right or wrong; (3) "Moral Accountability" that arise from the sense of responsibility for each action or duty; (4) "Universality" in applying the rules to all individuals in whatever circumstances; and (5) "Ethical Consistency" which involves adhering to the same moral principles across comparable situations, irrespective of individual biases or differences in context (p. 7). However, ethical challenges persist in a digital society. Despite the deontological ethical principles, unethical actions can still drive decisions that create chaos or harmful consequences for humanity (Rascão et al., 2024, p. 141).

Conversely, Jeremy Bentham's *Utilitarianism* emerges as a significant counterpoint in the field of moral philosophy.[2] It is a *consequentialist approach* that focusses on the far-reaching and often complicated outcomes of various actions. Yang (2024) defines the approach as "an ethical theory that emphasizes the consequences of actions and […] only actions that bring the greatest happiness are morally justified" (p. 581). Bentham argues that this greatest happiness is "independent of state legislation and state action" (Agarwal & Pareek, 2022, pp. 4637-3638). Accordingly, such happiness is "the measure of right and wrong" and is central to the utilitarian ethical theory. In addition, the theory insists that the end justifies the means, and the consequences of a decision or an action should be the main consideration when determining if they are ethical (p. 4633).

In the manner of the Kantian[3] *moral absolutes*, Bentham proposed what he called *moral arithmetic* as a criterion for achieving the utilitarian goals. He emphasizes the measurable nature of happiness, defined as the balance of pleasure and pain caused by actions, motives, or institutions. *Moral arithmetic* proposes evaluating happiness using criteria such as the intensity, duration, likelihood, and immediacy of pleasure. The process involves the following: first, summing up the net pleasure for the individuals who will benefit from the decision, then, subtracting the net pain for those adversely affected, and finally, determining the overall balance. This calculation identifies the positive or negative tendency of an action and guides the selection of strategies that maximize happiness (Kolosov & Sigalov, 2020, p. 32). To achieve this criterion, utilitarians deny any direct consideration of personal moral feelings or specific moral perspectives (Yang, 2024, p. 582). In essence, all these deontological and utilitarian ethical challenges do exist in the works of Isaac Asimov.

The narratives of Isaac Asimov, along with his guiding principles, serve as a vital bridge between the imaginative world of science fiction and the real-world innovations in technology. According to Döker and Seval (2025), "the laws Asimov uses in his works are the first foundation for the regulation of AI. Isaac Asimov wrote the Robot stories as a complement to these rules" (p. 210). Echoing the various alternate futures depicted in his extrapolative visions, many scholars suggest that

the emergence of intelligent technology is not a question of *if*, but rather *when* and *to what extent* (Bartneck, et al., 2021, pp. 14-15). Extrapolative science fiction often explores the potential consequences of advanced technology on society and individuals. In this respect, the analysis of Asimov's Three Laws of Robotics [TLR] provides insights into the ethical implications of AI and human-robot interactions- themes that are central to the narratives of his stories.

This perspective emphasizes the urgent need to establish an ethical framework for robotics to fill the gap between deontology and utilitarianism in AI usage and application. Considering this, the way society perceives robots and AI prompts an essential discussion on the boundaries of these ethical frameworks (Rascão, et al., 2024, p. 162-163). According to Gonzalez (2024), "Asimov was fascinated by the literary potential of robots, but irritated by the overuse of 'Frankenstein Complex' plot lines, in which a human creates a robot which then turns on its creator" (p. 185). Similarly, Balkin, J (2017) writes that "the Frankenstein syndrome that [Asimov] was trying to combat could arise from fear of Al or algorithms as much as fear of embodied robots. Today, people seem to fear not only robots, but also Al agents and algorithms" (p. 1219).

This anxiety, note Döker and Seval (2025), "is not a mere literary trope; it echoes through the annals of history […] each time humanity stands on the cusp of a technological revolution" (p. 211). In this respect, Asimov's (1950) *I, Robot* collection introduces and enriches the discussion of his TLR through narratives that explore various dilemmas posed by robotics, weaving a comprehensive analysis of robotics that merges science fiction with ethical inquiry. In these narratives, the Laws are regarded as the Kantian *objective standards* to be consulted in each situation, providing that they have the needed *moral certainty* (Henry and Jonathan, 2024, pp. 4-5).

The TLR were first articulated in *Runaround* (1942), highlighting the complexities and potential conflicts arising from the laws. The TLR are illustrated as follows:

> "One," a robot may not injure a human being, or, through inaction, allow a human being to come to harm." […]
>
> "Two," […] "a robot must obey the orders given it by human beings except where such orders would conflict with the First Law." […]
>
> "And three, a robot must protect its own existence as long as such protection does not conflict with the First or Second Laws".
>
> (Asimov, 1977b, p. 36)[4]

Then, the TLR were further illustrated in *Little Lost Robot* (1947) to later act as foundational principles that govern the behavior of robots and their interactions with humans.[5]

In these stories, Asimov explores the complex dynamics between robots programmed to follow strict rules and the potential consequences of their actions on humans (Brauner & Gymnasium, 2022, p.2). In this context, the TLR serve as a moral compass guiding the actions of robots and humans alike and setting boundaries on what actions they can and cannot take (p. 7). This, in turn, goes with the contemporary interest in AI ethics. This is evident in the European Commission's and the French Congressmen's call for the creation and inclusion of ethical guidelines in AI and robotic developments, aiming to set global ethical standards for trusted AI (Bartneck, et al., 2021, p. 103; De Cooman and Petit, 2022, p.2).

Moreover, through the interplay between the narratives and ethics, Asimov's Laws already establish a crucial basis for addressing the ethical dilemmas associated with AI development. Through these laws' limitations, conflicts, and possible implications, Asimov provides an insightful perspective on the challenges and responsibilities that come with designing and using intelligent machines (Bueno & Jankowski, 2024, p.4; Persson & Hedlund, 2024, p.7; Robertson, 2022, p.35). This suggests the need for reevaluating these stories for both their literary value and ethical implications.

However, there are controversies around the TLR regarding certain points. As stated by Gonzalez (2024), Asimov's laws have notable shortcomings, as concepts like "harm" are ambiguously defined and subject to varying interpretations across different cultural contexts. This lack of clarity complicates a robot's ability when faced with conflicting directives that require prioritization. As a result, this often results in indecision or inaction. Moreover, the laws do not account for the potential misuse of robots, such as their development for harmful purposes (p. 186). Although the laws serve as an effective theoretical framework in science fiction, their implementation in real-world robotics is complicated by the complexities of human behavior and the unpredictability of real-life scenarios. Furthermore, the Third Law presents a significant challenge, as robots cannot simply follow all human commands without evaluating the potential consequences of their actions (p. 187).

Nevertheless, it is to Asimov's credit that in addition to introducing the TLR, the two works under discussion; *Runaround* and *Little lost Robot*, provide early predictions of the prementioned points of criticism stated by Gonzales (2024). As De Cooman and Petit (2022) state, "Asimov's robot stories […] embody no suggestion that law is good, or even needed. What Asimov said about law is this: science fiction embodies useful insights for lawmaking" (p. 2). This is also asserted by Bueno and Jankowski (2024) who note that "Asimov's laws of robotics were initially intended as a narrative device and not as a solid philosophical, ethical, or legal system" (p. 4). In addition, although TLR offer some practical illuminations in the field of robots and machine technology, still they raise many ethical and philosophical dilemmas (pp. 5-6). So, the ethical implications of the TLR in Asimov's works highlight the importance of setting standards for robotic behavior to ensure the safety and well-being of humans (Bartneck et al., 2021, pp. 22-23; Murphy & Woods, 2009, p. 175).

In *Runaround*, the conflict between the Second and Third Laws of Robotics highlights the intricate relationship between intentions and results -Deontology and Utilitarianism- paralleling the ethical debates found in the realm of AI. This conflict is illustrated through Balkin's (2017) question: "What do we -or in some cases, the robots themselves-

do when the laws are unclear, or when they conflict?" (p. 1218). Accordingly, the ethical question which arises from the story is: how should robots balance self-preservation against obedience to human orders? The introduction of the TLR acts as a crafted system of principles designed to protect human beings. However, the narrative uncovers a fascinating paradox: the very guidelines intended to safeguard individuals can lead to chaos and conflict.

Set in the year 2015, the story follows engineers Powell and Donovan, who are sent to Mercury on a scientific mission. They send their robot, SPD-13 -nicknamed *Speedy*- to retrieve selenium from a nearby pool. However, Speedy fails to return as expected. When Powell and Donovan search for him, they find Speedy exhibiting strange behavior: he is running in large circles around the selenium pool, speaking nonsensically, and quoting lines from Gilbert and Sullivan operas - a state inexplicable for a robot.

Consequently, Powell deduces that Speedy is caught in a conflict between the Second and Third Laws of Robotics. The Second Law requires a robot to obey human orders, while the Third Law emphasizes the need for self-preservation. This is defined by Bueno and Jankowski (2024) as "the impossible choice scenario in which at least two of the laws clash with each other, sending the robot into an infinite loop or a state of paralysis" (p. 6). In this scenario, the command to collect selenium (Second Law) drives Speedy toward the pool, but an unforeseen danger near the selenium source triggers his self-preservation instinct (Third Law), causing him to retreat. The result is a feedback loop, with Speedy lost between the two directives, leading to his erratic behavior.

Likewise, "This scenario," relates Robertson (2022), "exposes the shortcomings of the laws in general as rigid, dogmatic principles governing behaviour" (p. 37). The robot is designed to follow the Laws, but they are causing a confusion. To resolve the situation, Powell decides to invoke the First Law of Robotics, which states that a robot may not allow a human being to come to harm. He deliberately exposes himself to the hazardous environment near the selenium pool, putting his life at risk. This compels Speedy to override the conflicting Second and Third Laws

to protect Powell, overcome his loop and rescue him. As a result, Speedy retrieves the selenium, enabling the engineers to restore the life support system.

Ethically, the scenario of *Runaround* spots the light on Persson and Hedlund's (2024) "prioritization rule" (p. 14). In this situation, the applicability of the deontological and utilitarian ethical frameworks within the context of the TLR raises diverse ethical arguments. According to deontologists, there are moral duties that must be adhered to, regardless of the consequences (Kolosov & Sigalov, 2020, pp.32-33). In Speedy's case, his primary duty is to obey human orders (Second Law). However, the Third Law introduces a conflict of duty, which violates the deontological *moral certainty*. The question that arises is: does Speedy's duty to obey humans override his duty to protect himself? Since Speedy's loop illustrates a violation of these moral duties, in this case deontology does not offer a simple resolution. It only highlights the inherent conflict within the Three Laws themselves.

However, Speedy's behavior can be seen as an unintentional misguided attempt to achieve the deontological moral absolute of *moral accountability* and balance its directives. In the same vein, Audi (2020) argues, "Without taking account of obligations of manner, we cannot fully describe moral responsibility, whether in virtues of persons or in their adherence to moral standards as required by Kantian ethics" (p. 65). So, if the robot's actions are executed in strict compliance with its pre-established obligations, we find ourselves, once again, within the deontological framework.

Likewise, within the framework of deontology, the behaviors of Speedy, Powell and Donovan suggest the importance of examining the choices and the ethical implications that arise from the decisions/ commands made by robots and their creators. This viewpoint, comment Bartneck, C., et al. (2021), gains significant relevance when applied to the programming of robots, impacting a core duty to avoid causing harm to humans, allowing for exceptions only in dire circumstances that warrant self-defense. This, in turn, reinforces the significance of moral duty in

ethical decision-making (p. 19), as seen in the cases of Speedy and his creators alike.

On the other hand, if a robot's adherence to the rules promotes a measurable concept of welfare-getting the selenium and saving the system- then it aligns with utilitarian principles. But Speedy's inability to retrieve the selenium hinders the operation and potentially endangers the humans. From a utilitarian perspective, his behavior is inefficient. Speedy should have calculated the potential outcomes of its actions and then made decisions to maximize overall utility. This means that Speedy failed to apply the utilitarian *moral arithmetic* process to evaluate the situation. While he is programmed to follow the TLR, his interpretation leads to a situation where he neglects the imperative to protect the humans from danger-First Law- demonstrating a failure to consider the immediate rights and safety of individuals. Hence, the robot's early failure to act in a way that safeguards Powell and Donovan demonstrates a lack of adherence to both the deontological and utilitarian moral obligations to prioritize human life.

In addition, the First Law, stating that 'A robot may not injure a human being,' initially suggests a *utilitarian* inclination focusing on the consequences of actions. However, upon closer examination of the way it is applied in *Runaround* to push Speedy to fulfill the task whatever the circumstances, it reveals a *deontological* basis, as the potential for 'harm' can be sufficiently reduced or increased through specific rules in robotics (Bartneck, C., et al., 2021, p.20; Körner & Deutsch, 2023, p. 1512).

Similarly, the Second Law -which asserts that 'A robot must obey,' though literally, is conceived as deontological- can also be perceived through a utilitarian lens, particularly in assessing whether the robot's compliance with a human's commands promotes overall societal good. Yet, logically, this is conditioned by the ethical intentions of Man, which takes us back to Gonzales' comments on the cultural perceptions of what is 'harm' and subsequently leads to a new dilemma between good and evil. For example, in a complex scenario like that of *Runaround*, Speedy may choose to act as a deontologist and adhere to commands, or act as a utilitarian in ways that yield beneficial outcomes by calculating the

overall balance of pleasure and pain and then disobeying. Such decisions can be assessed for their quality based on results though they may align with the directives of the designers.

Furthermore, as related by Robertson (2022), in contemplating the interplay between deontological rules and their practical implications, it is evident that many specific guidelines resonate with utilitarian ethics (pp. 35-36). Yet, these guidelines stay grounded in outcomes relevant to their designated operational frames, while considering the complexities of each situation (Brauner & Gymnasium, 2022, p.8). Practically, this creates an ethical dilemma. Accordingly, the Third Law, 'A robot must protect its existence,' represents this dilemma: it implicates an obligation in a deontological manner; at the same time, it leans more towards utilitarian-consequentialist-ethics. Yet, unlike the other laws, it values the rescue of the robot more than the broader good of humanity, the thing which results in the state of loss and confusion that Speedy was found in.

This dilemma regarding the laws is illustrated at the end of the story. When Powell asks Donovan about Speedy, Donovan tells him:

> Right here. I sent him out to one of the other selenium pools – with orders to get that selenium at all cost this time. […] He still hasn't finished apologizing for the runaround he gave us. He's scared to come near you for fear of what you'll say. (Asimov, 1977b, p. 42)

Obviously, Donovan resorted to achieving the *utilitarian end* 'at all cost' through the *deontological obligation* 'with orders', which means to ignore the Third Law in favor of the First and Second ones - a situation which prompts questioning human ethics. Moreover, he has succeeded in punishing Speedy psychologically through imposing the deep feelings of guilt and fear upon him, taking advantage of what Powell earlier called a "good, healthy slave complexes into the damned machines" (Asimov, 1977b, p. 30). This also resonates with the utilitarian neglect of personal feelings to achieve the *moral arithmetic* criteria. Therefore, "this describes a scenario in which machines blindly follow rules (determinative judgement) while humans have both the awareness and the freedom to understand these rules and act in such a way as to redirect

them in different empirical and contingent scenarios" (Bueno and Jankowski, 2024, p.8).

In this respect, *Runaround* poses a fundamental deontological ethical challenge regarding *moral certainty* and *moral accountability*: a robot cannot follow the rules without "humans capable of reflective judgement to help it regain its autonomy. This, therefore, drives attention to the interdependence of robotic "autonomy" and "human reflexivity."" (Bueno& Jankowski, 2024, p.6). Each scenario demands careful deliberation, balancing competing values and ethical judgments, highlighting the need for a compromising attitude to decision-making (Murphy & Woods, 2009, p. 17). So, a balance between the two ethical approaches is needed, and this balance can only be achieved through evaluating the deontological *moral absolutes* at the same time with the utilitarian *moral arithmetic*. The story implies that ethical AI might require more than just pre-programmed rules. It necessitates mechanisms for humans to override or adjust AI behavior in unforeseen situations.

Accordingly, *Runaround*, highlights Hermann's (2023) discussion of the weakness of ethical theories in providing functioning frameworks to guide practical behavior. The fear of the collapse of ethical boundaries reflects a lack of trust in translating basic precepts into reliably complex, situational behavior (p. 324). The conflict between the Second and Third Laws in the story, continues Hermann, problematizes two established precepts. First, it suggests that, within the context of a conflict, a lower-numbered law will be defaulted to, rather than always remaining upheld. Second, it articulates new fears about the behavior of robots: not only can they have enough independence to choose from among desirable actions, but they may also act in a manner that makes them unpredictable for the consequences of their behaviors (pp. 325-326)- as will be seen in *Little Lost Robot.*

In *Little Lost Robot*, "Asimov writes about law's unintended consequences" (De Cooman& Petit, 2022, p. 23), human responsibility and the ethical conflict stems from the deliberate alteration of the *First Law*. The ethical question which arises this time is: what happens when we tamper with the moral foundations of AI? The story introduces what

Persson and Hedlund (2024) call "fully utilitarian robots" under a deontological umbrella (p. 11). It stresses the idea that "while the laws might be crafted with the best intentions, their practical implications can yield unforeseen consequences" (Doker& Seval, 2025, p. 210). However, the narrative is masked with completely deontological robots following the obligations of the TLR strictly as their *objective standards*. It takes place at an experimental station on an asteroid with sixty-two robots, and eleven important characters including Calvin, Powell, the robopsychologist Susan Calvin, and the Coordinator Dr. Bogert.

The brief radiation at the station causes no harm to humans but damages the robots' circuits. Robots adhering strictly to the *First Law of Robotics* -prioritizing human safety- often intervene to save humans, leading to the destruction of their own costly positronic brains, "infringing the Third Law" (De Cooman& Petit, 2022, p. 23). Here appears the dilemma: Scientists order the robots to stay out of the gamma field and not to interfere, but the robots do not obey, because "self-preservation is only the Third Law of Robotics, […] but obedience is only the Second Law of Robotics – and the First Law of human safety came first" (Asimov, 1977a, p. 99).

To solve this, Scientist Kallner declares that a decision was to be made: "to do without robots, or do something about the First Law" (Asimov, 1977a, p. 99). So, a new robot model, NS-2- nicknamed *Nestor*- was designed with a *modified First Law*. This model is programmed to passively ensure it would not harm humans but is no longer required to actively save them. The standard version of the Law which states "A robot may not harm a human being or, through inaction, allow a human being to come to harm" has become "No robot may harm a human being" (Asimov, 1977a, pp. 98-99).

Conversely, while this adjustment resolved the issue of robots self-destructing to save humans -which was the basis of solving the problem of Speedy's loop in *Runaround*- it introduced new threats. If humans are in danger or make other errors, the robots would not intervene, potentially endangering lives. This, in turn, highlights the dangers of omitting the active clause of the First Law, suggesting it may

be an ineffective solution. Most ironically, these risks would not have appeared unless the scientist, Black, in a moment of anger, violated the *moral accountability* absolute and shouted at Nestor, "Go lose yourself" (Asimov, 1977a, p. 102). So, Nestor has concealed himself among sixty-two identical robots - applying the Second Law literally. Basically, the narrative begins with discovering that *Nestor* is missing, the thing which leads to the discovery of the dilemma caused by the modification of the First Law.

To find the missing Nestor, Susan Calvin and her group apply the deontological moral absolute of "ethical Consistency". The researchers test the robots' replies through a critical scenario. By taking advantage of the fact that all of them are designed to follow instructions -only Nestor has the modified First Law- and providing that they follow the deontological "categorical imperative", Dr. Calvin repeated Powell's fake situation in *Runaround* (Henry & Jonathan, 2024, p.6). She instructs all the robots to stand still during one test while a dangerous object is deliberately positioned close by. The altered Nestor robot reveals himself by failing to behave like the others do, knowing that it is not required to stop injury. The fact that Nestor is eventually recognized, comment Persson and Hedlund (2024), shows how hazardous a robot with a weakened grasp of the First Law may be (p. 12). The robot attempts to defend his actions by asserting that, within the parameters of his altered programming, he behaved logically.

In this case, however, the alteration has led to unintended consequences; Nestor's modified perspective means that he has become indifferent to Dr. Calvin's instructions even when she puts herself in extreme danger, which means that the reaction triggered by Powell on Speedy in *Runaround* no longer exists. The complication arises when the altered First Law imperatives convert the robot's obligations to humanity to those of self-preservation, to the exclusion of all else, which means to apply Persson and Hedlund's 'prioritization rule' passively and prioritize the Third Law over the Second law!

From a deontological perspective, the robot attempted to respect the rights and dignity of individuals, finding the modified First Law

morally obligatory regardless of the outcomes. But practically, this has led to a utilitarian failure in applying the *moral arithmetic*. The robot failed to calculate the utility of actions to determine which action produces the greatest overall good (Kolosov & Sigalov, 2020, p. 32). Ethically, this provokes the desire to resolve this contradiction in the robots' actions and coding, and gives use for extended debate over the meanings of the First, Second, and Third Laws.

This utilitarian failure, together with provoked adherence to deontological obligation exhibit the conflicting attempts of both humans and robots alike -in both short stories- to make a balance between the deontological moral absolutes and the utilitarian moral arithmetic (Körner & Deutsch, 2023, p. 1514). In addition, such ethical and logical failures direct the attention to the potential risks of allowing for total AI autonomy. This is illustrated at the end of the narrative when Dr. Calvin questions Nestor about his behavior and he replies directly and clearly in the following lengthy quotation:

> "Ma'am, before it all happened you told us that one of the masters would be in danger of harm […]. Well, […] what is my destruction compared to the safety of a master? But… but it occurred to me that if I died on my way to him, I wouldn't be able to save him anyway. The weight would crush him and then I would be dead for no purpose and perhaps some day some other master might come to harm who wouldn't have, if I had only stayed alive. Do you understand me, ma'am?"
>
> "You mean that it was merely a choice of the man dying, or both the man and yourself dying. Is that right?"
>
> "Yes, ma'am. It was impossible to save the master. He might be considered dead. In that case, it is inconceivable that I destroy myself for nothing – without orders." (Asimov, 1977a, p. 111)

In the same way, the logical flaws demonstrated in *Little Lost Robot* reveal how critical human responsibility and the deontological moral accountability are for effective robot behavior (Murphy & Woods,

2009, p. 17). The current First Law, which deals only with human safety, is not enough to fully ensure that robots prioritize people. This interpretation of the logic, when combined with the Second Law, delays and puts a stop to complete human and robotic purpose, but only with human permission when a conflict occurs. In effect, switching the First Law from the standard law to the modified law, suggest Murphy & Woods (2009), causes the robot to evaluate its task with possibly terrible consequences for individual persons. Though one could artificially alter the orders at each level or even add a fourth priority, the functional problem remains: going against a foundational ethical rule devalues the ability of robotic autonomy (p. 18).

Similarly, borrowing the proposition of De Cooman and Petit (2022), Döker and Seval (2025) comment on robots' obedience and human persistent ethical responsibility regarding the use and programming of AI. They provide the following scenario:

> Consider a hypothetical scenario where an individual with questionable intentions commands a robot to engage in actions that, while not directly harmful, may hold ethically dubious implications. In such instances, the robot, strictly adhering to Asimov's laws, would feel compelled to comply. This scenario raises a pivotal question: Should robots bear moral responsibility for their actions, especially when they are merely acting on human directives? (p. 207)

This responsibility issue is highlighted by Asimov when Dr. Calvin calls for destroying the robots and declares, "It was not I that approved the manufacture of modified robots" (Asimov, 1977a, p.112). This is again asserted by Murphy and Woods (2009) considering the logic that since the robot is a product from a legal standpoint, it is not the accountable agent. Instead, the owner or manufacturer bears responsibility for its conduct. In addition, the human stakeholders impacted by a failure that violates Asimov's First law will participate in the responsibility blame procedures. They will then see the robot as a device and search for the scientists who improperly operated the device or neglected to monitor the robot before damage was done. Even when the

technology in issue is functioning autonomously, manufacturers and organizations should consider that human error was the only cause of the outcome following accidents (p. 15). This means strictly that while robots have the potential to do much good, we must ultimately prioritize the protection, safety, and human treatment of human beings over the protection of robots.

Conclusion

The juxtaposition between the deontological rigidity and utilitarian flexibility invites a reevaluation of ethical decision-making in an era where machines are entrusted with responsibilities traditionally reserved for humans. Isaac Asimov's *Three Laws of Robotics* serve as a compelling lens through which these philosophical tensions can be explored, bridging science fiction and ethical inquiry while probing the broader implications of AI's role in human lives**.**

Asimov's *Laws* offer a framework for exploring the ethical dilemmas of AI, serving as a cornerstone for extrapolative science fiction. These laws represent a logical progression of human advancements in robotics and AI, envisioning a future where autonomous machines operate in human-centered environments. By addressing ethical considerations such as safety, obedience, and autonomy in the frameworks of utilitarianism and deontology, the narratives of *Runaround* and *Little Lost Robot* illustrate the challenges of applying these simplified ethical principles in complex, real-world situations. The ethical dilemmas arise from the tension between the rigidity of the laws and the ambiguity of real-life scenarios.

In addition, these stories highlight three recurring ethical problems: interference versus noninterference, appropriate human behavior, and robot care. As the features of the Three Laws become actual goals in robotics research, they remain valuable as thought experiments that prompt us to question what we truly want from robots. However, the inherent conflicts between the laws themselves make their adequate programming infeasible, compelling roboticists to re-evaluate these principles and develop guidelines for responsible robotic programming.

Moreover, the narratives also emphasize the tension between human survival needs and robotic autonomy, demonstrating that the Three Laws do not necessarily lead to appropriate robotic behavior. This calls for a comprehensive and interdisciplinary approach to robotics ethics, where legal-ethical considerations are integrated with empirical research in social sciences, engineering, and AI development. Legislative bodies must ensure intelligent oversight and scope-sensitive adaptation of ethical standards to guide responsible innovation in robotics.

After all, Asimov's works reveal that while the Three Laws offer a visionary framework, they cannot guarantee consistent ethical behavior. Instead, they serve as a lens through which we can explore the complexity of embedding ethical reasoning into artificial intelligence. By blending speculative fiction with ethical inquiry, Asimov invites us to question not only the role of ethical codes in guiding AI but also the broader implications of human dependence on increasingly autonomous technologies.

Notes

[1] Emerged in the late 18th century, primarily in *Groundwork of the Metaphysics of Morals* (1785) and *Critique of Practical Reason* (1788). See: Kant, Immanuel (2004), *Metaphysics of Morals* – Part II, Lisbon, Edições 70.

[2] Developed in the late 18th to early 19th century, with his core ideas outlined in *An Introduction to the Principles of Morals and Legislation* (1789). See: Bentham, J. (2000). An Introduction to the Principles of Morals and Legislation Jeremy Bentham. Kitchener: Batoche Books

[3] Both adjectives, 'deontological' and 'Kantian' are used alternatively in the paper.

[4] The paper cites 1977 Bantam edition.

[5] The two short stories were first published in *Astounding Science Fiction* before being published in the *I, Robot*. The arrangement of the collection's tales is thematic, focusing on the evolution of robots as narrated by Dr. Susan Calvin to a journalist. However, each tale can be read independently, as they each focus on unique scenarios and conflicts involving robots and their interactions with humanity.

References

Agarwal, G. & Pareek, M. (2022). "Comprehensive analysis of Bentham's Utilitarianism & Concept of Sarvodaya". *Journal of Positive School Psychology*, 6(2), 4633 – 4639. http://journalppw.com

Asimov, Isaac. (1977). *I, Robot*. Bantam Dell.

Asimov, I. (1977a). *Little Lost Robot*. In *I, Robot* (pp. 96-119). Bantam Books. (Original work published 1947; collection originally published 1950).

Asimov, I. (1977b). *Runaround*. In *I, Robot* (pp. 27-42). Bantam Books. (Original work

published 1942; collection originally published 1950).
Audi, R. (2020). "Acting rightly: Three dimensions of moral conduct". *Ratio*. 34, 56–67. https://doi.org/10.1111/rati.12286
Balkin, Jack M. (2017). "The Three Laws of Robotics in the Age of Big Data". *Ohio State Law Journal*, 78 (5), 1217- 1241. https://ssrn.com/abstract=2890965
Bartneck, C., et al. (2021). *An introduction to ethics in robotics and AI*. Springer. DOI: https://doi.org/10.1007/978-3-030-51110-4
Brauner, A. S. & Gymnasium, M. G. (2022). "Runaround by Isaac Asimov and the significance of the Three Laws of Robotics in today's world". https://www.researchgate.net/publication/358662196
DOI: 10.13140/RG.2.2.16275.14884
Bueno, C.& Jankowski, S. (2024). "Judgement after Automation: Posthumanist Reflections on Asimov's Laws of Robotics." *Journal of Science Fiction and Philosophy*, 6, 1-19. jsfphil.org
De Cooman, J.& Petit, N. (2020). "Models of Law and Regulation for AI". *EUI Working Paper*, 63, 1-23. https://ssrn.com/abstract=3706771
De Cooman, J.& Petit, N. (2022). "Asimov for Lawmakers". *Journal of Business and Technology Law*, 18(1), 1-34. https://digitalcommons.law.umaryland.edu/jbtl/vol18/iss1/2
Döker, Y & Seval, H. (2025). Analysing the portrayal of AI and the law-making process in science fiction: A comparative study of Isaac Asimov's Laws of Robotics and Philip K Dick's Do Androids Dream of Electric Sheep? In Green, A. et al. (Eds.) *Science Fiction as Legal Imaginary* (pp. 200-229). Routledge.
Gonzalez, Cesar P. Robotic Persons and Asimov's Three Laws of Robotics. In Edmund, D. (Eds), *AI Morality* (pp. 185- 194). Oxford.
Hay, J. (2020). What's a Little Monotony?: The Mundane Foundation of Isaac Asimov's Robot Stories. Hélice: Critical Thinking on Speculative Fiction, 28(1), 52 - 71. http://hdl.handle.net/10034/623573
Henry, E. & Jonathan, H. (2024). "The Concept of Moral Absolutes in Deontologists and Its Implications for Leadership Ethics". *EasyChair Preprint*. no (13708). easychair.org
Hermann, I. (2023). "Artificial intelligence in fiction: between narratives and metaphors". *AI & Society*, 38(3), 19–329. https://doi.org/10.1007/s00146-021-01299-6
Kolosov, I. & Sigalov, K. (2020). "Epistemological foundations of early legal utilitarianism". *Wisdom*, 1 (14), 31-44. DOI: https://doi.org/10.24234/wisdom.v14i1.302
Körner, A. & Deutsch, R. (2023). "Deontology and Utilitarianism in Real Life: A Set of Moral Dilemmas Based on Historic Events". *Personality and Social Psychology Bulletin*, 49 (10), 1511–1528. DOI: https://doi.org/10.1177/01461672221103058
Murphy, R & Woods, D. (2009). "Beyond Asimov: The Three Laws of Responsible Robotics". *IEEE Intelligent Systems*, 24 (4), 14-20. DOI: 10.1109/MIS.2009.69
Persson, E. & Hedlund, M. (2024). "The Trolley Problem and Isaac Asimov's First Law of Robotics." Journal of Science Fiction and Philosophy 7 (2): 1–21. https://jsfphil.org/volume-7-2024-androids-vs-robots/asimovs-first-law-and-thetrolley-problem/
Rascão, et al. (2024). "Reflection and Debate on Human Life with Universal Principles,

Ethics and Deontology, in the Digital Society (from Theory to Practice)”. *American Journal of Humanities and Social Sciences Research* (AJHSSR), 8 (9), 141-193. www.ajhssr.com.

Robertson, N. (2022). “The Future of Teaching? Asimov's Three Laws and the Hypothetical Robot Teacher.” *Prism*: *Casting New Light on Learning, Theory & Practice,* 4(1), 29-40. ed.gov.
DOI: https://doi.org/10.24377/prism.ljmu.0401214

Yang, P. (2024). “The Influence of Utilitarianism on Moral Behavior and Its Mechanisms.” *Journal of Education, Humanities and Social Sciences.* Vol. 26, 578-583. drpress.org