Judgement after Automation

Posthumanist Reflections on Asimov's Laws of Robotics

Celis Bueno, C.; Jankowski, S.

Publication date
2024

Document Version
Final published version

Published in
Journal of Science Fiction and Philosophy

Judgement after Automation:
Posthumanist Reflections on Asimov's Laws of Robotics

Claudio Celis Bueno and Steve Jankowski

University of Amsterdam

Abstract

This article argues that contemporary concerns about artificial intelligence often turn on the philosophical question: What is human about judgement? To understand the premise of this popular imaginary, we turn our attention to Isaac Asimov's Laws of Robotics, a set of laws deployed within his science fiction stories to create narratives about the relationships between humans and machines. Not only well-known for entertaining SF audiences, Asimov's laws reflect a shared imaginary regarding the relationship between humans and technology that permeates beyond the realm of science fiction, shaping some of the basic assumptions behind our definitions of politics, humanity, and freedom. Our argument begins by interpreting the stories *Runaround* (1942), *Risk* (1955), *The Bicentennial Man* (1976), and *Foundation and Earth* (1986) through Immanuel Kant and Hannah Arendt's philosophy of judgement. In doing so we pinpoint that the philosophical arcs of these stories are written in the tension concerning who (or what) is capable of both determinative and reflective judgement. Then, following theorists who have interpreted Asimov's "Zeroth Law" through the lens of posthumanism, we argue that Asimov's laws are constrained by the conception of reflective judgement as being inherently anthropocentric and limited to closed systems. In contrast, we advance the argument that reflective judgement emerges only through distributed and contingent systems — systems that include humans and non-humans. What should keep our attention then, is not the existential anxiety over an autonomous artificial intelligence that challenges human superiority, but the politics of creating and maintaining technical systems capable of sustaining distributed forms of reflective judgement.

Keywords:

Judgement, Asimov, Kant, Arendt, Posthumanism, Algorithms

Introduction

Delegating ethical or political decisions to machine-learning algorithms is one of the most vital areas of contention facing contemporary society. In recent years, these technologies have achieved impressive results when performing tasks such as classifying, recommending, or even generating information. Some argue that machine learning algorithms, when properly trained, perform fairer and more objective judgements than human agents (Kaplan 2024). At the same time, the issue of whether these technologies can (or should) perform judgements is tinted by the evidence that despite good intentions,

discriminatory biases are often encoded into these models and present the risk of amplifying social inequalities if deployed at scale (Davis et al. 2021; Eubanks 2017). Finally, some scholars emphasise the existence of a radical gap between human judgement and algorithmic reasoning, arguing for the irreplaceability of human agents in political, ethical, and social decision-making (Cantwell Smith 2019).

Issac Asimov's Laws of Robotics are both a hindrance and a guide to examining these concerns about the relationship between algorithms and judgement. Coming from his mid- to late-twentieth century *Robot* stories, these laws were used as a narrative device to exploit the paradoxical condition of modern technology: on the one hand, we are increasingly relying on machines capable of autonomous behaviour; while on the other hand, we are forced to regulate and control this autonomy in order to bring these technologies into alignment with existing legal, ethical, and political frameworks. Even though these laws were initially intended for science fiction plots, the current development of autonomous lethal weapons systems, predictive-policing algorithms, or self-driving cars is turning the question of specific rules for autonomous machines into a concrete legal, ethical, and political issue (Pasquale 2020; Coeckelbergh 2022). Furthermore, Asimov's laws reflect a shared imaginary regarding the relationship between humans and technology that permeates beyond the realm of science fiction, shaping some of the basic assumptions behind our definitions of politics, humanity, and freedom.

To unpack this shared imaginary, this article focuses on the issue of judgement. In particular, it uses Immanuel Kant's distinction between determinative and reflective judgment in coordination with Hannah Arendt's claim that the latter defines the core of our "human condition"—and the place of struggle of politics, ethics, and ultimately, human freedom—in order to examine the basic assumptions that allow us to differentiate between human judgement and machine rationality. By analysing three different situations that emerge from the deployment of the Laws of Robotics within Asimov's stories we identify a shared ground that informs an anthropocentric definition of machines: the lack of reflective judgement. The second part of the article then examines how Asimov's introduction of one of these laws—the Zeroth Law—opens the door for a novel interpretation of the issue of machine judgement grounded on a posthumanist and non-anthropocentric framework that complements and extends Angela Balzano's rewriting of Asimov's laws.

Human judgement, politics, and algorithms

In the Introduction to the *Critique of Judgment*, Immanuel Kant distinguished between two types of judgement: determinative and reflective (1987, pp. 18–19). Kant defines judgement as the "mediating link between understanding and reason" (1987, p. 5). This means that in the Kantian philosophical system, judgement allows for the laws of nature to harmonize with human freedom (1987, p. 15). Judgment, Kant adds,

> is the ability to think the particular as contained under the universal. If the universal (the rule, principle, law) is given, then judgment, which subsumes the particular under it, is determinative. […] But if only the particular is given and judgment has to find the universal for it, then this power is […] reflective. (1987, pp. 18–19)

On the one hand, our faculty of understanding allows for determinative judgments about natural phenomena: it is the ability to connect a particular case to a universal law. At the same time, our Reason is able to make reflective judgements about ethical and aesthetical phenomena: it can reflectively act as if our actions were guided by universal laws while still preserving the freedom to decide how and when to act on the world.

In her reading of Kant, Hannah Arendt (1992; 2003; 2005) identifies reflective judgement as the central component of human ethics and politics. Judgment, Arendt argues, outlines the basis of a political philosophy (1992, 9–10; 2003, 188; see also Zerilli 2005). Following Kant's distinction between determinative and reflective judgment, Arendt (2003, 189) argues that "thinking" (determinative judgment) is not the same as "judging" (reflective judgement). Thinking is linked to abstractions, rationality, calculability, and universal laws. Judging is linked to particulars, that is, to situated decisions that act in the world and, in doing so, provide their own legitimation (2003, 189). For Arendt, following Kant, reflective judgment is what ultimately differentiates humans from machines, granting the former with freedom, politics, and ethics (2003, 188; see also Zerilli 2005).

In relation to these concerns, the question of judgment has emerged as a central pivot within current debates about the ethics and politics of artificial intelligence. To a large extent, machine learning algorithms can be understood as pattern-recognition technologies aimed at the automation of judgement. Tasks such as classifying, recommending, or even generating information can be interpreted through the lens of judgment: the ability to connect a universal and a particular. At the same time however, it could be argued that computer algorithms can only perform determinative judgments as they can only follow rules but they lack the reflective ability to decide when and how these rules need to be applied. Brian Cantwell Smith (2019) argues that AI systems may be very good at performing calculations ("reckoning"), but that they lack the human ability to perform judgment. In Arendt's words, we could say that algorithms "think" but cannot "judge"; and in Kantian terms, we could say that they perform determinative judgments, but lack the ability to perform reflective ones. This radical (ontological) gap between machines and humans regarding the question of judgment remains one of the key arguments in current debates regarding the dangers of delegating ethical and political decisions to algorithmic systems.

These contrasting positions suggest that while it is necessary to consider the relationship between judgment and algorithms, the character of that relationship is not self-evident. This article mobilises Isaac Asimov's (1942) Laws of Robotics to work through the implications of what reflective judgement means in the context of algorithmic decision-making. Even if Asimov's laws may have been intended as narrative devices and not as a philosophical system or as a guide for practical application, they symptomatically express key presumptions about human-machine relations. Hence, an analysis of the laws' limitations may serve as a mirror on which an anthropocentric imaginary of the relation between culture and technics is reflected.

Asimov's robotic judgement

Asimov's famous Laws of Robotics were initially presented as a set of three laws in his 1942 short story *Runaround*. From this moment onward, they served as a key narrative device for many of Asimov's stories and novels, exploiting possible ambiguities and paradoxes that stemmed from subjecting these laws to complex scenarios. In each plot, the laws are implanted in the robots' "positronic brains" with the main objective of protecting humans from any harm that a robot could do to them, either through direct action or by failing to act in due time. The original three laws were the following (Asimov 1942):

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

Two initial observations about the laws are useful. First, the term "robot" was introduced by Karel Čapek in the 1920s and comes from the Czech word for "forced labourer." As such, a robot refers to a machine that is subjected to human orders to carry out heavy or dangerous tasks. At the same time, however, as Roger Clarke points out, the development of information technologies together "with the merging of computers, telecommunications networks, robotics, and distributed system software" has reinforced a notion of robot as a specific type of machine capable of performing complex actions "independent of human involvement" (1993, 55). This creates a paradoxical situation in which robots acquire a degree of autonomy that needs to be restrained in order to limit the potential harm they could cause to humans. As Persson and Hedlund put it, "the stronger and smarter robots become, the more useful they can be—and the more dangerous. This, in combination with the rapid progress in both AI and robotics, raises the question of how to make robots safe for humans" (2024, 1). The laws of robotics are hence required in light of the growing autonomy of robots (as opposed to non-autonomous machines, which are seen as mere instruments that would require no separate laws to those applying to their human users and/or producers).

The second observation regards the common belief that Asimov's laws of robotics were initially intended as a narrative device and not as a solid philosophical, ethical, or legal system. According to Clarke, for example, the inconsistencies and ambiguities of the three laws respond to their primary role as a "literary device intended to support a series of stories about robot behaviour" (1994, 57). Over time, Clarke adds, Asimov "found that the three laws included enough apparent inconsistencies, ambiguity, and uncertainty to

provide the conflicts required for a great many stories" (57). At the same time, Clarke argues, this narrative device can also help us reveal "problems that might later confront real roboticists and information technologists attempting to establish rules for the behaviour of intelligent machines" (57). Similarly, Wallach and Allen (2008) argue that despite the fact that they "offer little practical guidance as moral philosophy, and their value as specifications for algorithms is questionable" (p. 91), Asimov's stories show that the combination of even these three simple laws can "give rise to many ethical dilemmas" (8), as many scholars and researchers have expertly described (Clarke 1993 and 1994; Miller 2004; Wallach and Allen 2008; Murphy and Woods 2009; Coeckelbergh 2010; Leigh Anderson 2008 and 2011; Gunkel 2012; Balkin 2017; Stokes 2018; Pasquale 2020).

In this paper, however, we would like to approach Asimov's laws from an angle that differs from this common interpretation. As Persson and Hedlund (2024, 3) have argued in a recent paper, Asimov's "main purpose" for introducing these laws was to prevent his stories from falling into a "Frankenstein complex"—a scenario in which human creations would be bound to "turn against their creator." According to Persson and Hedlund, in fact, "the usefulness of the three laws for creating entertaining problems was seen by Asimov himself as a bonus. It was not the main purpose of the laws" (2024, 3). Moreover, Persson and Hedlund show that for Asimov the laws were "obvious from the start" and that everyone is more or less "aware of them subliminally" (Asimov in Persson and Hedlund 2024, 3). Similarly, Salge and Polani (2017, 2) write that despite the different "technicalities and implications of the three laws […] we believe most people would agree with the general sentiment of the rules; Asimov himself argued that these rules are not particularly novel, but govern the design of any kind of tool humanity produces." Following this interpretation, what really stands out for us is how the laws express a shared imaginary—a common sense we could say—of the relationship between humans and technology. More specifically, we argue that this shared imaginary is grounded within a particular notion of human judgement.

For this purpose, the next two sections will demonstrate how some of the laws' shortcomings are in fact caused by a lack of "reflective judgement"—as Sage Leslie-McCarthy (2007) argues, robots have the capacity to *follow the letter of the law*, but not its *spirit*. This premise, we claim, reinforces a popular imaginary in which reflective judgment appears as the decisive element that distinguishes humans from machines, culture from technics.

Three laws and three shortcomings

It has been mentioned that the three laws are usually understood as a narrative device which, precisely because of their generic and abstract character, allow for great storytelling. But philosophically speaking, they are systematically incomplete, full of inconsistencies and ambiguities. As argued above, we believe that these "weaknesses" are symptomatic of a shared imaginary of the modern division between culture and technics according to which reflective judgement lies at the core of human freedom, ethics, and

politics. In what follows, we briefly refer to three particular scenarios in which reflective judgment appears as the key distinctive element separating humans from robots.

The first situation can be characterized 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 (Persson and Hedlund 2024, 7). In Asimov's *Liar*, for example, a robot breaks down "because it is presented […] with an impossible choice: to either reveal the solution to how it got its special power (telepathy), which would hurt the ego of the humans who wants to find the solution by themselves, or not reveal the solution, which would also hurt the humans because they really need the solution" (Persson and Hedlund 2024, 7). This type of situation was later expanded in *Runaround*, where the plot includes a robot who is commanded to carry a load of a given mineral from one space station to another. Following the second law, the robot is forced to obey this human-given order. At the same time, the third law (which had been strengthened in this particular model) prevents the robot from complying because of the dangers that such forced labour would infringe on its own safety. Because of its programmed nature, the robot endlessly processes these two laws and is incapable of deciding which one to obey. The story unfolds in a way that, once the human operator discovers the cause of the paralysis, he decides to place himself in a situation that threatens his own life. This forces the robot to follow the first law, thus exiting the loop between the second and the third law in order to rescue him. This means that in order to follow the spirit of the laws, the robot paradoxically requires humans capable of reflective judgement to help it regain its autonomy, which therefore calls attention to the interdependence of robotic "autonomy" and "human reflexivity."

The second situation emerges from the potential ambiguity of an order issued to a robot using human language. Clarke refers to this as "the ambiguity and cultural dependence of terms" (1993, 57) that undermine the possibility of a non-mediated relation between language and reality. Put differently, to correctly follow the three laws, a robot would need to be able to process the semantic meaning of human language. This includes not just translating the precise meaning of the laws into repeatable and consistent actions, but also interpreting the potential ambiguity that may stem from human orders.

In Asimov's short story *Risk*,

> the vagueness of the word 'firmly' in the order 'pull the bar towards you firmly' jeopardizes a vital hyperspace experiment. Because robot strength is much greater than humans, it pulls the bar more powerfully than the human had intended, bends it, and thereby ruins the control mechanism. (1993, 57)

This opens a more general question about the translatability of the ambiguous and embodied character of human language into the formal character of computer calculability. As many have argued (Winograd 1972; Weizenbaum 1976; Searle 1980; Dreyfus 1999; Levesque 2017; and more recently Bender and Koller 2020), there seems to be a radical gap between the semantic, referential, and context-based nature of human language and the formal (syntactic) logic of computer rationality. This gap would render it impossible to offer a programmable rule capable of subsuming the situated character of the former under

the formalism of the latter. In relation to Asimov's laws, this opens questions such as: how does one offer a formal, computable definition of "harm," or a computable definition of "human?" According to these authors, because human language is always dependent on meaning (Searle 1980), semantic ambiguity (Levesque 2017), or referentiality and intentionality (Bender and Koller 2020), language becomes the central site for challenging uncritical reflections of human-machine interactions. Coming from legal studies, Sage Leslie-McCarthy reframes this distinction in terms of the two interpretations of law. Within the Robot series, there are "various situations in which the literal or formalistic interpretation of law is put in conflict with what can loosely be referred to as the intention or 'spirit' of these laws" (Leslie-McCarthy 2007, 400). As she rightly points out, Asimov's narrative "shows his distinct preference for interpretations that favour the 'spirit' of law" (p. 400), a preference that tends to be reserved for his human characters.

The third scenario created by Asimov's laws that we wish to address here can be referred to as heteronomy: the condition under which an intelligent being remains bound to external laws. As Susan Leigh Anderson (2008; 2011) has argued, Asimov's Three Laws of Robotics are unacceptable "as a basis for machine ethics" since a truly ethical agent does not simply follow externally imposed (heteronomous) rules but enacts a free (autonomous) decision regarding which rules to follow within the context of specific conditions. To develop her argument on the "unacceptability" of the three laws, Leigh Anderson focuses on Asimov's (1976) short novel *The Bicentennial Man*.

The story centres on Andrew, a servant robot programmed with the three laws who slowly develops a desire to be "free." After a long trial and debate, he is finally granted this freedom by a judge; and yet—in spite of this granted freedom—Andrew continues to follow the Three Laws. This creates a conflict between the right to act freely and the need to follow external rules. This conflict becomes crucial in a specific scene where Andrew is bullied by a group of humans but is incapable of defending himself because of the imposition of the First Law. Despite his granted "freedom," Andrew is not a true autonomous agent because he is, above all, ruled by the external laws programmed in his "positronic brain." In light of this situation, Leigh Anderson argues, Andrew is not recognised as a free agent and hence cannot be held "morally responsible for his actions" (2011, 291). In this sense, the three laws highlight the paradoxical situation of "autonomous agents" that must follow rules that keep them in the service of humans. For this reason, Leigh Anderson concludes, Asimov's laws are an unacceptable mechanism for establishing a true ethical autonomous system.

All of Asimov's robots are psychopaths

Upon reflecting on these stories we argue that these three scenarios draw our attention to a shared imaginary regarding the radical difference between humans and machines: the absence of reflective judgement. Each plotline results from the mechanical application of a determinative judgement that attempts to subsume all particular cases to predefined rules. Put differently, for Asimov's three laws to actually work, robots would need to have "sufficient capabilities for judgement" that would cause them to, for example, "frustrate the

intentions of their masters when, in a robot's judgement, a higher order law applies" (Clarke 1994, 57). In this sense, many of Asimov's stories seem to be constructed following an opposition between the capacity of humans to perform reflective judgements and the inability of machines (despite how "smart" they become) to do so. As such, these stories tend to reinforce an anthropocentric distinction between the spontaneity of human reason and the calculability of machine thinking.

In the case of the "impossible choice" scenario, the robot is paralysed by a conflict between two or more laws while the human's judgement is the means to exit paralysis. In the story *Runaround*, by understanding that the robot was caught between two conflicting rules, the human operator was able to introduce a new creative element (risking his own life) to bypass the loop. 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 (reflective judgement). From this first situation then, reflective judgement appears as the capacity to evaluate when a general law should be applied to a particular case (Arendt 2003, 188–189). As such, reflective judgement would entail not the blind obedience of rules, but a form of meta-reflection. This entails a conception of human judgement that is neither deterministic nor mechanistic, but rather "free and spontaneous" (Arendt 2003, 189). While Asimov's robots can only enact determinative judgements (i.e., to apply universal rules to local cases), their human counterparts have the power of reflective judgement which allows them to act based on empirical and contingent data.

From the point of view of the second scenario addressed above, reflective judgement appears as the capacity to evaluate the meaning of an intrinsically ambiguous statement. Because human language is context-based and, as such, entails a structural ambiguity, human meaning is sometimes posed as "untranslatable" into a formal and programmable computer language. For a number of scholars, this untranslatability constitutes the key impediment for an artificial general intelligence (Winograd 1972; Weizenbaum 1976; Searle 1980; Dreyfus 1999; Levesque 2017; Bender and Koller 2020). In Clarke's words, faced with the intrinsic ambiguities of language, robots would have to "exercise judgement to interpret the meaning of words and hence of orders and of new data" and therefore, only a reflective judgement could allow a robot to properly "determine whether and to what extent the Laws apply to a particular situation" (1993, 57). Again, reflective judgement appears as the capacity that distinguishes humans from machines.

The third scenario mentioned referred to the opposition between a true ethical agent (autonomy) and an agent that merely follows external rules (heteronomy). This was exemplified with Asimov's story *The Bicentennial Man*. In this case, too, reflective judgement appears as the capacity to mediate between human freedom (spontaneity) and blind obedience to a rules system (determinative judgement). Mark Coeckelbergh (2010, 236) argues that any artificial agent that only follows rules and does not consider the particulars provided by context (such as emotion and imagination), will be nothing but a "psychopathic robot." Furthermore, Coeckelbergh argues that "it would not only be wrong to call such rule-following robots 'moral', but it might also be dangerous to build them"

(2010, 236). Put differently, to design a robot that can only enact determinative judgements is to design a "psychopath." Following Lacan on psychosis, a psychopath is not someone who does not follow (social) rules, but an agent who forgets that there is a structural gap (an intrinsic ambiguity) between any rule system (belonging to the symbolic domain) and the Real (Lacan 1997, 9). This means that psychotic behaviour is not due to acting outside the rules, but being unable to break them when a given situation demands so. From this perspective, reflective judgement appears as the mediating mechanism between rules and particular cases. Reflective judgement is hence what ultimately provides humans with the freedom and spontaneity to decide when and how to follow a specific rule. To put it in Coeckelbergh's terms, reflective judgement would be essential for building non-psychotic robots.

Based on these three situations, it is possible to define Asimov's robots as machines capable of determinative judgement and Asimov's human characters as agents that possess the ability of both determinative *and* reflective judgement. This establishes a sharp schism between humans and machines. Furthermore, this reading of Asimov's laws through the notion of judgement reinforces to a large extent Arendt's interpretation of Kant, defining reflective judgement as the key to understanding "the human condition," its essential political and ethical character, and the differentiating element between humans and machines. Furthermore, in the current context of algorithmic technologies, this opposition creates a disjunction between two theoretical stances. On the one hand, the belief in the exceptionalism of reflective judgement for ethical and political life (and the corresponding belief in the impossibility of translating reflective judgement into algorithmic rules). From this perspective, trying to automate judgement will result in totalitarian (Arendt), psychopathic (Coeckelbergh), inauthentic (Cantwell Smith), or simply unethical results (Leigh Anderson).

On the other hand, a line of thought that began with Wiener's (1948) cybernetic theory, acquired a cyberfeminist reading through Sadie Plant (2000), and has bifurcated into transhumanist dreams of truly intelligent machines capable of reflective judgement (George et. Al. 2020; Pinka 2020; Kurzweil 2000, 2005, 2013). Despite their radical differences, however, both positions grant a privileged position to the issue of reflective judgment. In the first case, as a way of highlighting the unsurpassable gap between human spontaneous thinking and machinic calculation (and the dangers of not fully differentiating them). In the second case, reflective judgment is used to advocate for the possibility of a true thinking artificial being (an AGI or "singularity"). Nevertheless, as Yuk Hui warns us, it is important to overcome this polarisation that simply reduces the debate to either a "stark rejection of technology" or, by contrast, "a naïve defence of machines" (2021, 228). Instead of simply opposing culture and technics, then, the theoretical challenge today is to understand their mutual co-constitution. This is what Hui defines as the major task of philosophy "after automation" (p. 218).

The reflective precondition of the Zeroth Law

The three Laws of Robotics are not sacrosanct. In 1985, Asimov amended them with the Zeroth Law as a narrative response to many of their shortcomings and aporias (Clarke 1994, 58; Henry 2018, 48). The law stated that "a robot may not injure humanity, or, through inaction, allow humanity to come to harm" (Asimov 1986, 353). This was presented as the highest-order law, which meant that it would always overwrite the other three laws. Hence, from the perspective of the Zeroth Law, humanity as a whole is rendered more important than any single individual—the universal more significant than the particular.

Nonetheless, the abstract character of the Zeroth Law entails a practical impossibility that marks the relation between judgement and algorithms even more prominently.[1] As Barbara Henry (2018, 48) argues, reflective judgement is precisely that which cannot be transferred or translated to a rule but can only be developed through constant exercise. For this reason, the Zeroth Law would entail revolutionary and disruptive consequences for human-robot relations; they demand a "situated judgement" (*"giudizio in situazione"*) that cannot be reduced to a set of formal rules or transferred as a piece of software, but that rather needs to be developed through "practical experience" (Henry 2018, 48). The Zeroth Law hence appears as a meta-law that, in order to work, would require the exercise of reflective judgement from the side of robots. The Zeroth Law could then be considered as a sort of "categorical imperative" for robots, that is, not simply a rule, but a form of meta-reflection that requires reflective judgement to decide how and when a subsequent rule should be applied in epistemological, ethical, or politically ambiguous situations.

Furthermore, it could be argued that the Zeroth Law accentuates the ambiguous character of human language already examined above. Here, the question of how to define a human being is replaced by the issue of how to define humanity. As an abstraction, humanity "may refer to the set of individual human beings (a collective), or it may be a distinct concept (a generality, as in the notion of 'the State')" (Clarke 1994, 58). In the novel *Foundation and Earth*, for example, Asimov (1986) presents an "Aristotelian" scenario in which a group of people ("Solarians") produce robots that only classify as "fellow humans" those who speak the same language ("Solarian"). This means that the Zeroth Law would only apply to those "formally defined" as humans, excluding all others. From this example it could be argued that the question of meaning and understanding is not simply a technical problem, but rather represents the core issue of politics (see, for example, Weil and Bender 2023). Put differently, the question of the ambiguity of language, accentuated in the case of the Zeroth Law, reinforces the profound relation between reflective judgement and politics. As Arendt argues, politics is "based on the fact of human plurality" (2005, 93). Hence, the problem of politics is that of the "coexistence and association" of this plurality, pushing

[1] For a more detailed critique of the shortcomings of the Zeroth Law, see Persson and Hedlund (2024, 13-14).

human animals living together to permanently reflect upon what unifies them and what distinguishes them, constantly redefining the limits of "who" is included (or excluded) as human. A law that reduces this definition to a formal and calculable programme would be, from Arendt's perspective, the end of politics and the beginning of totalitarianism (2003, 31).

This brings us to another dimension of the Zeroth Law that highlights the need for reflective judgement: the relation between information and decision-making. Again in *Foundation and Earth*, one robot discusses the challenges posed by the Zeroth Law:

> In theory, the Zeroth Law was the answer to our problems. In practice, we could never decide. A human being is a concrete object. Injury to a person can be estimated and judged. Humanity is an abstraction. How do we deal with it? (1986, 353)

In relation to this passage, Clarke explains that the Zeroth Law does not actually deal with concrete individuals but with "groups and probabilities" (1994, 58). Accordingly, Miller (2004, 196) contends that the Zeroth Law faces the same problem of any utilitarian philosophy, that is, the "lack of sufficient information to determine what action will be maximally beneficial to humanity." Put differently, when dealing with an abstraction like humanity, decisions will always be based on limited information (Clarke 1994, 59). Because of this, decisions must be based on an estimate of probability and, as such, will require "human (or robot) judgement" (Clarke 1993, 58).

From this perspective, it could be argued that one difference between determinative and reflective judgement is the relation between information and decision-making. Determinative judgement assumes that the universal law contains all of the required information to decide whether to subsume or not a particular case under it. Reflective judgement, instead, acts within the gap between information and decision-making: it presupposes that decisions have to be made based on contingent, partial, and empirical data. The possibility of a decision based on absolute knowledge (determinative judgement) is an abstraction that exists only within closed systems. Information in open systems, instead, is structurally limited. Faced with the impossibility of absolute knowledge, then, reflective judgement functions as an operation that, in each case, bridges the gap between limited information and decision-making in open-system scenarios.

New laws for posthuman relations

According to Barbara Henry (2018, 48), there is a "latent posthumanism" in Asimov's Zeroth Law.[2] This latent posthumanism is significant because it can disentangle the debate of human-machine relations from the anthropocentric frameworks that have dominated it. Asimov (1986) uses the term "Galaxia" to describe a conception of the universe as a single organism composed of planets, suns, human and non-human animals, organic and non-

[2] For an alternative interpretation of the Zeroth Law not grounded in a posthumanist framework, see Persson and Hedlund (2024, 13-14).

organic matter, technical objects, machines, etc. From this perspective, even though individual human beings play an important role within this larger organism, they are nonetheless elements of this "collective being" (Miller 2004, 200). In Miller's view (p. 201), conceiving the universe as a single entity that includes humanity as part of it is one way of solving the applicability problem of the Zeroth Law. We have mentioned that the abstract nature of the Zeroth Law made it practically impossible to be applied using only determinative judgement. This is so if we still consider humanity as the sum of individual human entities. If, on the contrary, we conceive humanity as part of a single organism named "Galaxia," it becomes a concrete individual that can be "cared for" (p. 201). While this shift towards a posthumanist reading of robotic laws is a step in the right direction, Miller still frames "Galaxia" as a "closed system." In this sense, this collective organism can be said to be bound to a linear causality and therefore to determinative judgements (Hui 2021). What is required, then, would be a shift from the linear causality of closed systems towards the recursive causality of open (complex and contingent) systems. In a 2017 conference, Jack Balkin elaborated on this approach:

> When people think about robots in science fiction, they often think of self-contained entities. But today we know that many robots and Al agents are connected to the cloud. That is certainly true of the Internet of things and home robots. It is likely to be true of self-driving cars as well. So the laws of robotics, whatever they are, are also likely to be the laws of cloud intelligences that are connected to the Internet. (Balkin 2017, 1220)

Likewise, Roger Clarke had previously argued for an "open-system approach" to the laws of robotics. According to him, artificial intelligence "no longer comprise independent machines each serving a single location," but systems "designed to support all elements of a widely dispersed organisation" (1994, 62). In this new context of digital telecommunications, "a set of laws for robotics cannot be independent but must be conceived as part of a system" (60). This means that it is not enough to implant the laws on each robot's "positronic brain." Instead, any robot designed to function in our world must be conceived as part of larger (open and complex) systems that include "data collection, decision-analytical, and action processes" through which each individual robot can "apply the laws" (60). Furthermore, Clarke adds, in a complex scenario where human and non-human intelligent agents constantly interact, even human activity must begin to be "perceived as part of a system" (62).

What Balkin's and Clarke's arguments share is an intuition that a more systemic approach to the laws of robotics is needed. This would demand a shift from a notion of judgement as an individual faculty towards a notion of judgement as the distributed capacity of an open and complex system to regulate itself through its informational exchange with the environment. Furthermore, this would entail a non-anthropocentric approach to both human and machine judgement. Living beings, human animals, and advanced technical objects together create a cybernetic garden, a common capability of regulating their activity in relation to a porous territory of systems. Furthermore, creative and inventive judgement is necessary in each entity in order to adapt to changing

environmental conditions. From this perspective, judgement no longer appears as that exclusive human faculty that mediates between the universal and the particular, but as a transversal (non-anthropocentric) regulatory function that constantly negotiates between conservation and innovation, between the organism and its environment. To a certain extent, this position may sound similar to that of Cybernetics. But it differs in one key aspect. Within Cybernetics, the self-governing of a cybernetic organism is presented merely as the corrective action for dealing with asymmetries of information flow and processing. In other words, it is a theory of technics and culture where machines reproduce the functions of living organisms and/or human-like intelligence through a reflective judgment that is understood only as a biological behaviour. But following Arendt, this cannot be the case. Reflective judgment is inherently political. Therefore, what we are suggesting is a twofold shift.

The first key objective is to move away from a notion of reflective judgement as the capacity of one individual (human, animal, machinic) towards a notion of reflective judgement as the informational exchange between distributed systems. Secondly, this exchange is not merely a biological response to the environment. Instead, it is a political process, where the degrees of openness and closedness of these systems are contingent on the situated and partial knowledge of their interdependence with the environment. In other words, the political question cannot be whether or not the system is distributed. That is a given. The political questions are how is this distribution designed from the outset, who has the legitimacy to reconfigure this design, and according to which justifications?

Feminist technology theorists have long filled out the details of these political concerns of cybernetic systems. Sadie Plant's influential texts from the 1990s conceive this shift from closed to open systems as a revolutionary stage towards the overcoming of patriarchal societies. She distinguishes between a notion of intelligence grounded on the linear processes of centralised computers and the new "connectionist" machines capable of parallel processing and possessing learning abilities (Plant 2000, 329). This new distributed processing, she argues, "defies all attempts to pin it down, and can only ever be contingently defined. It also turns the computer into a complex thinking machine which converges with the operations of the human brain" (329). In the 1990s, research in both Artificial Intelligence and neuroscience were converging, each understanding its object as a "complex, connective, distributive machine" (Plant 2000, 329). As Yuk Hui puts it, this was part of a larger turn from a linear to a recursive causality (2021b, 341). According to Plant, complexity theory and the notion of open systems were key aspects of this novel paradigm:

> [T]he parallels proliferate. The complexity the computer becomes also emerges in economies, weather-systems, cities and cultures, all of which begin to function as complex systems with their own parallel processes, connectivities and immense tangles of mutual interlinkings. (2000, 329)

Nevertheless, Plant argues, not all these revolutionary systems were suddenly let free to self-organise. Power apparatuses, sciences, institutions, and academic knowledge attempted to reproduce a centralised, top-down structure in order to control these newly

unleashed self-organizing machines. These institutions and corporations "intended to guarantee the centralised and hierarchical control of market processes, cultural developments and, indeed, any variety of activity which may disturb the smooth regulation of the patriarchal economy" (329). For Plant, Asimov's laws of robotics belong to these attempts to restrict the revolutionary power of open systems by reducing robots to closed entities:

> When Isaac Asimov wrote his three laws of robotics, they were lifted straight from the marriage vows: love, honour and obey. Like women, any thinking machines are admitted on the understanding that they are duty-bound to honour and obey the members of the species to which they were enslaved […] But self-organising processes proliferate, connections are continually made, and complexity becomes increasingly complex. In spite of its best intentions, patriarchy is subsumed by the processes which served it so well. (329)

In a spirit similar to that of Plant, Angela Balzano (2020) calls for a rewriting of Asimov's laws of robotics from the perspective of open systems. In her introduction to the Italian translation of Donna Haraway's *The promises of monsters*, Balzano (2020) argues that Asimov's three laws are of no use to Haraway (nor to any post-anthropocentric understanding of technology) since they presuppose that each robot constitutes a "self-referential identity." Asimov is the type of science fiction "that Haraway invites us to rewrite," she argues (Balzano 2020). Hence, we should follow Haraway rather than Asimov and rewrite the laws of robotics from a post-anthropocentric perspective, that is, from a point of view in which human, nature, and machines do not interact as closed entities, but instead establish an open relation that precedes the individual. The new laws should not merely apply to robots as that "machinic other," but for us as cyborgs, as hybrid networks of human and non-human elements (Balzano 2020; see also Haraway, 2004). Using related language, Leslie-McCarthy suggests that in the context of the Zeroth Law, Asimov stresses "'interpretative' readings of Law over literal readings" which leads to constructing posthuman laws that are "underpinned by a notion of the fundamental kinship of 'intelligent' beings" (2007, 414).

Following Haraway's (2004, 69–70) notion of "inappropriate/d others,"[3] Balzano (2020) defines the posthumanist version of the First Law in the following terms:

[3] Haraway's (2004, 69–70) concept of "inappropriate/d others" refers to "the historical positioning of those who cannot adopt the mask of either 'self" or 'other' offered by previously dominant, modern Western narratives of identity and politics. To be 'inappropriate/d' does not mean 'not to be in relation with'—i.e., to be in a special reservation, with the status of the authentic […]. Rather to be an 'inappropriate/d other' means to be in critical, deconstructive relationality, in a diffracting rather than reflecting (ratio)nality […] To be inappropriate/d is […] to be dislocated from the available maps specifying kinds of actors and kinds of narratives, not to be originally fixed by difference […] The term 'inappropriate/d others' can provoke rethinking social relationality within artifactual nature."

> Humans are not self-sufficient beings. They are part of a complex and articulated collectivity. As such, humans may not injure the 'inappropriate/d others', all forms of organic life, artificial life, more, less, or non-human life. Humans may neither allow that these 'inappropriate/d others' become injured because of their inaction.

Balzano's Second Law, in turn, reads like this:

> Humans must obey the orders from the very same complex and articulated collectivity that conforms them, and which includes biomechanical entities, microbes, viruses, electric circuits, primates, pets, wild plants and cyborgs, except where such orders would conflict with the first law. (2020)

And the Third Law is:

> Humans must not pursue at any cost the preservation nor reproduction of their own existence when doing so would conflict with the first law. (2020)

Besides these three laws, and just like Asimov did, Balzano introduces the need for a posthuman Zeroth Law:

> Humans must fight for the survival of the Earth as a whole, because this is where human life takes place, together with all forms of organic and artificial forms of life, more, less, or non-human, cyborgs, other creatures, monstrous and inappropriate/d. (2020)

This posthuman Zeroth Law complements Asimov's idea of "Galaxia" addressed above with Haraway's (2016) notions of the "monstrous" and the "inappropriate/d others." As such, this proposed Zeroth Law should be our new political "compass" (Balzano 2020). Without it, we would not be able to clearly navigate the muddy waters of contemporary political struggles. We hence need to abandon Asimov's focus on individual entities (each human and each machine as a closed system) and move towards Haraway's (2016) "sympoietic" worldview in which human, technical, and living beings are constantly interacting and mutually recomposing each other. Only in this way, Balzano (2020) concludes, we will be able to imagine new "horizontal and transversal coalitions" beyond inherited forms of anthropocentric prejudice.

Conclusion

In this article we have repurposed Kant and Arendt's considerations of judgement to interpret Asimov's stories to think of "philosophy after automation" (Hui 2021). In doing so, we have come across different scenarios that confirm what Leslie-McCarthy identified as the key tension within Asimov's stories: while his robots follow the "letter of the law," they do not have the capacity for following its "spirit." This incapacity of robots to gulf that interpretive gap of reflective judgement elicits both humanist pity and relief; pity because these machines are so human-like we empathize with their plight — relief because our singular command of reflective judgement keeps us from being dethroned from the sovereign seat of Reason.

But this is but one telling of the story. Despite the clear differences between the posthumanist authors (Clarke 1994; Balkin 2017; Plant 2000; Henry 2018; Balzano 2020), each advocated for a reading of Asimov's stories in a way that unsettles such self-aggrandizing illusions. As such, they represent an attempt to go beyond the stark opposition between humans and machines, between culture and technics, between the individual and the multiple.

This situation becomes especially clear in the context of the Zeroth Law that, by its very nature, cannot be followed without engaging in reflective judgement. It not only presumes a creative capacity of non-humans, it also requires transgressing the presumption of closed systems. From these considerations, we have argued for the necessity of a new notion of judgement. Grounded in posthumanist theory, this judgement is defined, not as an individual faculty, but as the capacity of porous and complex systems to regulate themselves and their political relations to the proximate environment. It is not acceptable to simply argue for a cybernetic ontology that conceives of all beings (human and non-human) exclusively as self-regulatory systems. The challenge is how to acknowledge the cybernetic condition while still leaving room for political (reflective) judgments that deal with asymmetric distributions within these open systems. This newly recognizable form of reflective judgement necessarily reconfigures what we have come to know as "humanity," perhaps to the point that another collective polity comes to be, one we lack the name for but feel the crisis of its necessity all the same.

Acknowledgements

Some sections of this article were written within the framework of the project "Immaginario politico, filosofia postumana e tecnologie emergenti" (2021) directed by Prof. Barbara Henry, Scuola Superiore Universitaria Sant'Anna, Pisa, Italy.

References

Arendt, H. (1992). *Lectures on Kant's political Philosophy*. The University of Chicago Press.

Arendt, H. (2003). *Responsibility and Judgment*. Schocken Books.

Arendt, H. (2005). *The Promise of Politics*. Schocken Books.

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

Asimov, I. (1976). *The Bicentennial Man*. Ballantine Books.

Asimov, I. (1986). *Foundation and Earth*. Del Rey.

Balkin, J. (2017). "The Three Laws of Robotics in the Age of Big Data." *Ohio State Law Journal, 78*(5), 1217–1241. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2890965

Balzano, A. (2020). "Haraway in Loop." *Opera Viva*. https://operavivamagazine.org/haraway-in-loop/

Bender, E. M., & Koller, A. (2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Cantwell Smith, B. (2019) *The Promise of Artificial Intelligence: Reckoning and Judgement.* MIT Press.

Clarke, R. (1993). "Asimov's Laws of Robotics: Implications for Information Technologies (Part 1)." *IEEE Computer, 26*(12), 53–61. https://ieeexplore.ieee.org/document/247652

Clarke, R. (1994). "Asimov's Laws of Robotics: Implications for Information Technologies (Part 2)." *Computing Milieux*, 57–66. https://doi.org/10.1109/2.248881

Coeckelbergh, M. (2010). "Moral appearances: emotions, robots, and human morality." *Ethics and Information Technology, 12*, 235–241. https://doi.org/10.1007/s10676-010-9221-y

Coeckelbergh, M. (2022). *The Political Philosophy of AI: An Introduction*. Polity.

Davis, J., A. Williams, and M. Yang. (2021). "Algorithmic Reparation." *Big Data & Society* 8 (2). https://doi.org/10.1177/20539517211044808

Dreyfus, H. (1999). *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT Press.

Eubanks, Virginia. (2017). *Automating Inequality*. St Martin's Press.

George, D., Lázaro-Gredilla, M., & Swaroop Guntupalli, J. (2020). "From CAPTCHA to Commonsense: How Brain Can Teach Us About Artificial Intelligence." *Frontiers in Computational Neuroscience, 14*, 1–14. https://doi.org/10.3389/fncom.2020.554097

Gunkel, D. (2012). *The Machine Question*. MIT Press.

Haraway, D. (2004). *The Haraway Reader.* Routledge.

Haraway, D. (2016). *Staying with the Trouble*. Duke University Press.

Henry, B. (1992). *Il problema del giudizio politico fra criticismo ed ermeneutica*. Morano Editore.

Henry, B. (2018). Cultura di massa, immaginario politico, etica e robotica. *In Circolo* (6), 39-54. http://www.incircolorivistafilosofica.it/cultura-di-massa-immaginario-politico-etica-e-robotica/

Hui, Y. (2021). "Introduction: Philosophy after Automation?" *Philosophy Today*, 65(2), 217–233. https://doi.org/10.5840/philtoday2021652392

Kant, I. (1997). *Critique of Practical Reason*. Cambridge University Press.

Kant, I. (1998). *Critique of Pure Reason*. Cambridge University Press.

Kant, I. (1987). *Critique of Judgement*. Hackett Publishing Company.

Kaplan, J. (2024). *Generative Artificial Intelligence: What Everyone Needs to Know*. Oxford University Press.

Kurzweil, R. (2000). *The Age of Spiritual Machines*. Penguin.

Kurzweil, R. (2005). *The Singularity is Near*. Duckworth Overlook.

Kurzweil, R. (2013). *How to Create a Mind*. Duckworth Overlook.

Lacan, J. (1997). *The Seminar of Jacques Lacan. Book III. The Psychoses. 1955–1956*. W. W. Norton & Company.

Leigh Anderson, S. (2008). "Asimov's 'three laws of robotics' and machine metaethics." *AI & Society*, *22*, 477–493.

Leigh Anderson, S. (2011). "The Unacceptability of Asimov's Three Laws of Robotics as a Basis for Machine Ethics." In *Machine Ethics*, edited by M. Anderson & S. Leigh Anderson. Cambridge University Press, 285-296.

Levesque, H. (2017). *Common Sense, the Turing Test, and the Quest for Real AI.* MIT Press.

Leslie-McCarthy, S. (2007). "Asimov's Futuristic Pharisees: Examining the Posthuman in Isaac Asimov's Robot Novels." *Law, Culture and the Humanities. 3*(3), 398–415.

Miller, J. (2004). "The Greatest Good for Humanity: Isaac Asimov's Future History and Utilitarian Calculation Problems." *Science Fiction Studies, 31*(2), 189–206.

Murphy, R., & Woods, D. (2009). "Beyond Asimov: The Three Laws of Responsible Robotics." *Intelligent Systems, IEEE, 24*(4), 14–20. https://doi.org/10.1109/MIS.2009.69

Pasquale, F. (2020). *New Laws of Robotics*. Belknap Press.

Persson, E., and M. Hedlund. (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-the-trolley-problem/

Pinka, R. (2020). "Synthetic Deliberation: Can emulated Imagination enhance Machine Ethics?" *Minds and Machines*, 1-16. https://doi.org/10.1007/s11023-020-09531-w

Plant, S. (2000). "On the Matrix. Cyberfeminist simulations." In *The Cybercultures Reader*, edited by D. Bell & B. Kennedy. Routledge, 325–336.

Powers, T. (2011). "Prospects for a Kantian Machine." In *Machine Ethics*, edited by M. Anderson & S. Leigh Anderson. Cambridge University Press, 464-475.

Sadin, É. (2019). *Critica della ragione artificiale*. Luiss University Press.

Salge, C. and D. Polani. (2017). "Empowerment as Replacement for the Three Laws of Robotics." *Frontiers in Robotics and AI,* 25(4), 1-16. https://doi.org/10.3389/frobt.2017.00025

Searle, J. (1980). "Minds, Brains and Programs." *Behavioral and Brain Sciences, 3*(3), 417-457. https://doi.org/10.1017/S0140525X00005756

Spaulding, N. (2020). "Is Human Judgement necessary?" In *The Oxford Handbook of Ethics of AI*, edited by M. Dubber, F. Pasquale and S. Das. Oxford University Press.

Stokes, C. (2018). "Why the three Laws of Robotics do not work."*International Journal of Research in Engineering and Innovation, 2*(2), 121–126.

Wallach, W., & Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.

Weil, E., & Bender, E. M. (2023). "You are not a Parrot." *New York Magazine*. https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html

Weizenbaum, J. (1976). *Computer Power and Human Reason*. W. H. Freeman and Company.

Wiener, N. (1948). *Cybernetics, or control and communication in the animal and the machine*. MIT Press.

Winograd, T. (1972). *Understanding Natural Language*. Academic Press.

Zerilli, L. (2005). "We feel our Freedom: Imagination and Judgement in the Thought of Hannah Arendt." *Political Theory, 33*(2), 158–188. https://www.jstor.org/stable/30038411