Chapter 1

Who Should Obey Asimov's Laws of Robotics? A Question of Responsibility

Maria Hedlund and Erik Persson

Lund University, Sweden

Abstract

The aim of this chapter is to explore the safety value of implementing Asimov's Laws of Robotics as a future general framework that humans should obey. Asimov formulated laws to make explicit the safeguards of the robots in his stories: (1) A robot may not injure or harm a human being or, through inaction, allow a human being to come to harm; (2) A robot must obey the orders given to it by human beings except where such orders would conflict with the First Law; (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law. In Asimov's stories, it is always assumed that the laws are built into the robots to govern the behaviour of the robots. As his stories clearly demonstrate, the Laws can be ambiguous. Moreover, the laws are not very specific. General rules as a guide for robot behaviour may not be a very good method to achieve robot safety – if we expect the robots to follow them. But would it work for humans? In this chapter, we ask whether it would make as much, or more, sense to implement the laws in human legislation with the purpose of governing the behaviour of people or companies that develop, build, market or use AI, embodied in robots or in the form of software, now and in the future.

Keywords: The laws of robotics; Asimov's laws; robot ethics; AI ethics; safety; responsibility; democracy

Introduction

The aim of this chapter is to explore the value of implementing Asimov's Laws of Robotics as a general framework for humans with the purpose of governing the behaviour of people or companies that develop, build, market or use artificial

The Ethics Gap in the Engineering of the Future, 9–25

doi:10.1108/978-1-83797-635-520241002

intelligence (AI), embodied in robots or in the form of software, now, and in the future. In 1942, the science fiction author Isaac Asimov introduced the Laws of Robotics to make explicit the safeguards of the robots in his stories (Asimov, 1942). Generally, safety refers to the prevention of harm or other non-desirable outcomes (Hansson, 2012). Asimov's laws aimed to do exactly that:

(1) A robot may not injure a human being or, through inaction, allow a human being to come to harm.
(2) A robot must obey orders given to it by humans except where such orders would conflict with the First Law.
(3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law (Asimov, 1942).

Later, Asimov added a fourth, or 'Zeroth' Law, that preceded the other laws in terms of priority:

(0) A robot may not injure humanity or, through inaction, allow humanity to come to harm (Asimov, 1985).

The laws are intuitively appealing: they are simple and straightforward, and they embrace essential ethical principles of many societies. However, as his stories clearly demonstrate, the Laws can be ambiguous, making it difficult for the robot to follow them. Moreover, the laws are not very specific. For instance, should a robot obey orders from any human? Or how would a robot act if information is kept from it? In addition, these laws were hardwired into the robots' 'positronic' brains. It is doubtful whether it would be possible to build general laws like these into a real robot. And if it would, who should be responsible for the actions of the robot?

Responsibility presumes the capability to make a difference and awareness of what you are doing. Currently, robots lack these attributes and cannot be responsible for their actions, which is a good reason not to hand over complex decisions to robots. Adding context specific considerations to general rules, such as the Laws, are necessary to make reasonable judgements in real-world situations (Persson & Hedlund, 2024). Although self-learning machines get increasingly better at reading the environment in which they act, they do not understand what they are doing, why they are doing it, or the consequences of what they are doing. Thus, general rules for robot behaviour may not be a very good method to achieve robot safety – if we expect robots to follow them. But would they work for humans?

We are not the first to ask this question. In light of the rapid AI development that we are currently witnessing, there is concern for the human relationship to more and more capable robots. And Asimov's laws seem to have encouraged political actors as well as roboticists and software developers: a search on 'three laws of robotics' in Google scholar in December 2023 gives more than 607,000 hits.

Two examples of how political actors have been inspired by Asimov's laws are South Korea and the European Union. In 2007, the South Korean Government drew up a Robot Ethics Charter that would cover ethical standards to be programmed into robots. According to an official from the Ministry's Code of Ethics, the charter would reflect Asimov's Laws of Robotics. However, in a revised version of the South Korean Robot Ethics Charter from 2018, Asimov's laws are not mentioned (Choi et al., 2019).

In 2016, the European Parliament suggested acknowledging robots as 'electronic persons' and that Asimov's Laws of Robotics should be guiding ethical principles for robots (EP, 2016a). The suggestion met hard criticism for the misinterpretation that fictional laws intended for robots could protect humanity (EP, 2016b, 2017), and did not recur later in the process of developing ethical guidelines for AI. These examples from the political sphere illustrate that Asimov's laws have an impact outside the fictional world.

Also in the scholarly literature, we find the idea to apply Asimov's Laws. One example is Clarke (1993, 1994), who sees Asimov's Laws as a set of principal guidelines, or lessons, to be applied during design, development and use of robotic systems. One such lesson is that Asimov's Laws do not designate any particular class of humans as more deserving, that is, they should benefit all humans equally. Another lesson is that we must not focus only on the technology as such but also on how the technology is used.

Another example of the impact of Asimov's Laws on scholarly literature is Murphy and Woods (2009), who suggest three alternative laws. Their alternative laws place the responsibility for robot safety on humans. They also suggest that the hierarchy with humans as superior and robots as subordinate is not always suitable. For instance, we might prefer that the robot ignores a hacker.

In a more recent article that takes Asimov's Laws as an explicit point of departure, Balkin (2017) argues that what we need is laws directed at the people who programme and use robots, AI agents and algorithms. Balkin's proposal focuses on trust and fairness and states that AI businesses should have an obligation to be trustworthy towards their end users and to the public, and that algorithm operators have a duty not to externalise the costs of algorithmic decision-making onto others.

The connection of this proposal to Asimov's Laws is emphasised by Pasquale (2017), who suggests a Zeroth Law to 'ensure the viability of Balkin's three laws'. With reference to Microsoft's chatbot Tay, which quickly adapted its messages in a racist and misogynist direction, Pasquale suggests that the creator of a robot should be obliged to build in constraints on the code's evolution. This has some similarity to Asimov's Laws, which were hardwired into the robots.

Another kind of reference to Asimov's Laws in the scholarly literature is how they are included or excluded from reviews of ethical guidelines on AI. In the context of reviewing current frameworks for regulating AI, all published between 2016 and 2019, Torres and Penman (2021) include Asimov's Laws from the 1940s as one of these frameworks of AI. They found that Asimov's Laws 'easily [...] matched the frameworks of today's AI mainstream'.

In another review, Hagendorff (2020) explicitly excludes Asimov's Laws from the ethical guidelines under scrutiny. This review only includes guidelines published within the last five years, and the fact that the author justifies why he excludes Asimov's fictional laws from the 1940s is a further sign of their influence in thinking about AI development.

There are also examples of practices that claim to consider Asimov's Laws. According to Abdullah et al. (2021), medical bioethical research 'has always considered the Asimov laws, no matter how primitive they were, in the bioethical design of medical AI', and Kaminka et al. (2017) use Asimov's laws to examine safety and autonomy in molecular robots fabricated from a technique called DNA origami.

It is clear that we are not the first to ask whether Asimov's Laws could work for humans. However, we believe that it is necessary to try them out in some real-case situations. We also believe that it is necessary to incorporate the concept of 'responsibility' to do so. The remainder of this chapter is structured as follows. First, we will outline an understanding of responsibility that put emphasis on control and awareness. After that, we present our proposed idea of Asimov's Laws directed at humans, and apply it on a hypothesised real-case contemporary situation. This analysis helps us to finetune the Laws directed at humans. Next, we apply the revised version of the proposition on the Laws directed at humans on three significantly different and morally relevant directions of AI development. Finally, we present our conclusions and discuss potential ways forward.

Responsibility

In this chapter, we will focus on two aspects of responsibility that has particular relevance for our thought experiment: awareness and control. Normally, we are morally responsible for something we have caused as long as we are not acting under coercion or ignorance and are aware of the moral nature of the action (or inaction), that is, that we have moral agency (Held, 1970; Sneddon, 2005; Thompson, 1987).

Adequate knowledge about the causal relations and consequences of an action is necessary for any understanding of responsibility (Adam & Groves, 2011; Thompson, 1987). Of relevance is to what extent the agent is able to realise the effects of an action, the effects of not to act, or the effects of acting differently (Fischer & Ravizza, 1998). This is also valid for knowledge about right and wrong. In some situations, what is right and wrong is not contestable, but in many cases, there are no definite rights or wrongs. However, even without any definite answer on the question of right and wrong, there are always context dependent norms of what constitute a right or a good behaviour (Hedlund, 2012). Bad actions resulting from (real or alleged) ignorance of moral norms are blameworthy and can be seen as paradigm cases of moral responsibility (FitzPatrick, 2008). An actor who unintentionally or unvoluntary has caused a bad situation can be causally but not morally responsible, as she has not done anything blameworthy or objectionable (Talbert, 2008; Thompson, 1987).

To be responsible, agents must be able to control their actions (or non-actions) (Fischer, 1982). 'Control' is often referred to as the power to determine whether or not something occurs (Kane, 2002). In relation to AI algorithms that may develop in directions that is difficult or impossible to predict, even for the designer of the algorithm, the concept of 'responsibility gap' has been suggested to denote situations in which human control is undermined (Matthias, 2004). A responsibility gap is however an unwelcome situation, both practically and conceptually. We do not want to find ourselves in a situation in which AI gives rise to harm without being able to make someone accept that they are responsible for this harm. Perhaps it is too much to demand that the designer should be able to control every causal step in what the AI does. Conceptually, 'responsibility' does not seem to do its job here. Perhaps 'control' is too strong a requirement for responsibility, at least in cases involving AI algorithms and other autonomous systems.

To avoid a responsibility gap, Himmelreich and Köhler (2022) suggest that we instead build responsibility not on control, but on another kind of causal-like relation such as supervision. Drawing on Nyholm (2018), who argues that the relation between the developer and the AI is analogous to the relationship of supervision that is ongoing between a parent and a small child, Himmelreich and Köhler (2022) contend that developers stand in a supervisory relation to AI and autonomous systems. They have control in the sense that they 'maintain, improve, and teach the AI system what to do and how to behave' (Himmelreich & Köhler, 2022). With this weaker relationship between the developer and the AI, the developer can be responsible for what the AI does, even though she cannot fully predict or control its exact course of action (Nyholm, 2018). Supervision places the incentive correctly, as the developer is the one who has influence over the AI by training it, and by this weaker causal-like relation, the developer can be responsible for a harm that an AI causes because she trained it (Himmelreich & Köhler, 2022).

Asimov's Laws of Robotics Directed at Humans

We suggest that a reasonable way to make Asimov's Laws apply to humans is to phrase them in terms of responsibility. Given these premises, Asimov's Law's directed at humans could look like this:

(1) AI developers have a responsibility to see to it that an algorithm may not injure a human being, or, through inaction, allow a human being to come to harm.
(2) AI developers have a responsibility to see to it that an algorithm obeys the orders given to it by humans except where such orders would conflict with the First Law.
(3) AI developers have a responsibility to see to it that an algorithm protects its own existence as long as such protection does not conflict with the First or Second Law.

And the Zeroth law:

(0) AI developers have a responsibility to see to it that an algorithm may not injure humanity or, through inaction, allow humanity to come to harm.

Could the laws, formulated like this, help remedy the damage that recommender algorithms have on the functioning of democracy and thereby avoid severely harming humans? Asimov's Laws aimed at preventing harms to humans and humanity. As pointed out by several scholars (e.g., Clarke, 1993; Schurr et al., 2007), 'harm' is notoriously vague and need to be specified. Schurr et al. (2007) chooses to operationalise 'harm' as 'a "significant" negative loss in utility', capturing as well physical as mental harm. Harm can be direct or indirect. One way of severely harming humans is to damage the democratic system. To try out our idea of applying Asimov's Laws on humans, we consider one particular harm, namely how recommender algorithms may damage the functioning of democracy.

Democracy here refers to a system in which citizens have equal rights and real possibilities to decide on common matters. In its ideal form, democracy is a clever way to reach agreements or make compromises when we do not agree on matters. Instead of fighting, we vote, and we discuss and deliberate. Even though a sound democracy requires that we accept that people have different views, some common ground is necessary for meaningful democratic debate. However, recommender algorithms contribute to giving us very different world views. Based on what we 'like' or share on social media platforms, or on what we search for in search engines, recommender algorithms direct us to content that we are assumed to prefer. The effect is that we are confronted with different world views, which make democratic debate and mutual tolerance difficult (Hedlund & Persson, 2024). Could Asimov's Laws directed at humans help remedy this damage to democracy and thereby protect humans from harm?

Like we stated above, our discussion refers to AI, embodied in robots or in the form of software (algorithms). Later, in the imagined future scenarios, we will primarily talk about super-intelligent AI. In this first case, we will only mention algorithms. We assume that relevant humans are those who design, develop and provide the algorithms. Humans can have several tasks in relation to an AI system, such as designers, developers, operators, deployers or users. The point here is that we are talking about humans involved in development of robots and systems built on AI technology, irrespective of the precise role they play. For the sake of brevity, we use 'developer' to denote any relevant human involved in development of AI, and 'user' for relevant humans that somehow make use of the technology.

Application on a Hypothesised Real-Case Contemporary Situation

Consider a developer of a recommender algorithm. What could these laws imply for her? We need to keep in mind that there is a hierarchy between the laws, with

the Zeroth law as the superior law. Hence, first and foremost, the developer has the responsibility not to harm humanity.

We assume that damaging the function of democracy is a harm to humanity. We also assume that it is important for all humans, for different reasons, including a well-functioning democracy, to be exposed to other opinions and not only be exposed to the same opinions that you already have, which recommender algorithms tend to do. Thus, according to the Zeroth Law, the responsibility of the developer is to see to it that the algorithm does not cause different world views for different people, which would damage the function of democracy and thereby harm humanity. That would require that the algorithm does not give dissimilar recommendations to different people who are making the same search.

How would this align with the First, Second and the Third Law?

The First Law states that a human being may not be harmed. Hence, the developer has the responsibility to see to it that individual humans are not harmed. While individual human beings may benefit from a functioning democracy, they may suffer from getting search results that are not individualised. This is a harm that the developer (possibly) is responsible to avoid.

In 'Liar', Asimov (1941) discusses a similar but less dramatic case when a telepathic robot lies to people, since it has realised that the truth would be painful for them, but, in fact, the robot causes more harm for the people by lying to them. This parallels our algorithm case. By unreflectingly providing those who use the algorithm with what they want, they are probably instantly happy, but in the long run, they will get harmed by not taking part of ideas that differs from their own. Recall J. S. Mill's idea of the importance to expose one's opinions to critical scrutiny both for the good of the society that otherwise would stagnate in old habits and for individuals to thrive (Mill, 1859/2011).

But the First Law is subordinate to the Zeroth Law, aimed at protecting humanity from harm, and in this case, the harm to the individual human being cannot be avoided if humanity should be protected (again, assuming that damage to democracy is a harm to humanity). Harming the individual may however have effects that is detrimental for democracy in other ways. For instance, this individual may lose her confidence in the digital infrastructure that aims to serve a favourable discussion atmosphere and thereby the functioning of democracy.

Confidence, or trust, is a highly prioritised value in discussions on human-technology interaction and include aspects such as safety and transparency of AI systems (AI HLEG, 2019; Buruk et al., 2020; Fjeld et al., 2020), but as Duenser and Douglas (2023) argue, it is important to acknowledge that trust in AI involves not only reliance on the system itself, 'but also trust in the developers of the AI system'. In Balkin's (2017) words, AI developers 'have duties of good faith and trust towards their end-users'.

In our case, the individual's loss of trust in the AI system may, potentially, be outweighed by the individual's trust in the developer, given that it is known to the individual that the developer prioritises the superior law with the aim at protecting humanity. This is however not a very strong claim, as individuals tend to be self-interested and short-sighted. On the other hand, would it not be for the protection of humanity (that is, democracy), there would be no public debate at all, which

would give the individual at least a good reason to trust the developer. Thus, as the hierarchies between the Laws imply, the developer has the responsibility to prioritise the long-term and collective over the short-term and individual.

How about the Second Law? According to this law, the responsibility of the developer is to see to it that the algorithm obeys human orders. While the law does not specify which human the algorithm should obey, we do not know whether it should obey the developer or the user.

In some sense, the algorithm has no choice but to follow the instructions of the developer. On the other hand, machine-learning algorithms may develop in a way that the developer cannot predict. Depending on user input, the algorithm adapts. This adaptation could perhaps be seen as a kind of obedience of the user. But if this adaptation of search results is the root to the damage to democracy, then the developer, according to the Zeroth Law, is responsible not to design adaptive algorithms, *or* to design adaptive algorithms that develop in another direction than creating diverse world views for different users. In the former case, we have ruled out the user as a human that the algorithm should obey. In the latter case, the user is still a human that the algorithm should obey. However, this means we will have to consider also users that are malevolent and for some reason want to utilise the algorithm to make harm or to destroy the algorithm.

We do not want the algorithm to obey all users equally. To avoid that, we suggest that we discriminate what we could call authorised users, that is, users who apply the algorithm as intended, and unauthorised users, such as hackers. But does not the premise in the Second Law that the orders should not be obeyed if they conflict with the First Law already rule out hackers? Theoretically, it does, but hackers are often very creative and good at concealing what they are doing (Scroxton, 2023), and it may not be immediately obvious that the hacker is doing something harmful. To allow for algorithms that adapt (which is, in fact, a key characteristic of AI algorithms), we would have to reconsider the formulation of our Second Law. Drawing on Murphy and Woods (2009), who argue that robots must be built to fit the roles that individuals have, we suggest that the Second Law is revised:

(2) AI developers have a responsibility to see to it that an algorithm obey the orders given to it by *authorised* humans except where such orders would conflict with the First Law.

This revised Second Law takes into account which user the algorithm is obligated to obey and allows the disposal of orders exceeding the authorisation of the user. A difficulty is of course how to discriminate between authorised and unauthorised users, but to solve this primarily technical issue is a responsibility that our developer will have to take on.

The Third Law refers to protection of the algorithm. But for machine-learning algorithms that develop over time, what is it that should be protected? The developer's original blueprint of the algorithm as it was first put to work in the search engine? The algorithm as it is being developed by interaction with user data? And if so, at what point of time?

Considering the superior Second Law, which could be interpreted as the developer's responsibility not to design adaptive algorithms, it should be the original blueprint that the designer is responsible to protect. But as we indicate above, we do not want to rule out adaptation as such, but rather have a kind of adaptation that is not damaging for democracy. Our chosen interpretation of the Second Law assumes that the adaption is integral to the algorithm, and thereby should be protected. But regardless whether we want to protect the original or the adapted version of the algorithm, the algorithm needs to be protected from hackers, which the revised version of the Second Law should warrant.

There could, of course, be other causes of a destroyed algorithm than the act of hackers, which the Third Law directed at humans aims to take care of. For instance, suppose that the algorithm adapts to the extent that it not only exposes the individual to other world views than she already has but also to world views that are intrinsically damaging to humanity, say, world views that divide groups of people into superior and subordinate and are not worthy of equal treatment. Without doubt, that would be detrimental for democracy and for humanity. As Clarke (1993, 1994) reminds us, the laws should benefit all humans equally. Pasquale's (2017) idea on constraints on code evolution would be useful in this regard. Paraphrasing Pasquale, we propose that the developer is responsible to build in constraints to the algorithm such that to prevent outcomes that are damaging to democracy and harmful for humanity. This leads to the following specification of the Third Law:

(3) AI developers have a responsibility to see to it that an algorithm protects its own existence, *as it was intended by the developer*, as long as such protection does not conflict with the First or Second Law.

It appears that Asimov's Laws directed at humans, with some revisions, will be able to protect the functioning of democracy considering the context of technological development.

As it seems from this premature conjecture, the hierarchy between the First and the Zeroth Laws directed at humans solves the potential conflict between short-term individual interest and long-term societal interest. Hence, our developer needs to think more long-term than the robot did in 'Liar' and conclude that a well-functioning democracy is more important for individuals than to be protected from search results that go against their prevailing view. This will follow from prioritising the Zeroth Law before the First Law.

Regarding the Second Law directed at humans, there is a need of a specification of which humans that the algorithm should obey, which we achieve by making a distinction between authorised and unauthorised users. By that, we protect the algorithm from deliberate destruction by hackers.

To protect the algorithm from unintended destruction by its own adaptation, we add a specification to the Third Law that assigns responsibility to the developer to ensure that adaptation is constrained to the extent that the original intention of the algorithm is kept.

The revised version of Asimov's Laws of Robotics directed at humans taken together looks like this:

(1) AI developers have a responsibility to see to it that an algorithm may not injure a human being or, through inaction, allow a human being to come to harm.
(2) AI developers have a responsibility to see to it that an algorithm obeys the orders given to it by authorised humans except where such orders would conflict with the First Law.
(3) AI developers have a responsibility to see to it that an algorithm protects its own existence, as it was intended by the developer, as long as such protection does not conflict with the First or Second Law.

And the Zeroth Law:

(0) Developers have a responsibility to see to it that an algorithm may not injure humanity or, through inaction, allow humanity to come to harm.

Application on Imagined Future Scenarios

How would the revised version of Asimov's Laws of Robotics directed at humans work on future AI technology? In this section, we will tentatively apply the Laws to three imagined future scenarios in which AI has been developed well beyond today's level of sophistication. We take our departure in discussions on Super-intelligence and Artificial General Intelligence (AGI) (Bostrom, 2014; Russell, 2019; Tegmark, 2017) and prospects of conscious AI (Schneider, 2019), and imagine three significantly different and morally relevant directions of AI development. To try out how the revised Laws directed at humans apply in these cases, we adjust them to apply to the technology in question, but make no further changes.

Scenario 1: Super-intelligent AI that does *not* understand what it is doing, why it is doing it, or consequences of what it is doing.

With 'superintelligence' we follow Bostrom and refer to 'any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest' (2014: 22). The super-intelligent artificial systems that Bostrom discusses in his famous paperclip example[1] does not have any kind of 'understanding', but a brute force rationality kind of intelligence and ability to act towards its goal to maximise paperclip production. Hence, the super-intelligent AI that we imagine here differs from today's AI systems in scope and efficiency, but not in kind, as it

[1]Bostrom's paperclip example refers to a thought experiment in which 'someone programs and switches on an AI that has the goal of producing paperclips. The AI is given the ability to learn, so that is can invent ways to achieve its goal better. As the AI is super-intelligent, if there is a way of turning something into paperclips, it will find it. [. . .] Soon the world will be inundated with paperclips' (Gans, 2018).

does not have any kind of moral awareness. What would that mean for the responsibility of our developer as it is stated in the Laws directed at humans?

The relation between the developer and the super-intelligent AI would be the same as in our contemporary hypothesised case. As this super-intelligent AI lacks moral agency, the principal responsibility of the developer will not be challenged. However, as this super-intelligent AI is far more efficient than contemporary AI and on practically all areas, the extent of the developer's responsibility to protect humanity and individual human beings will be dramatically larger. Still, this is not a principal, but a practical issue, yet a huge one.

Regarding the Third Law directed at humans, aimed at protecting the developer's intention for the super-intelligent AI, one practical issue that the developer will have to handle with particular care is how to ensure that these intentions (to protect humans and humanity) is formulated in a way that cannot be misunderstood by the super-intelligent AI. To align the goals of a potential future super-intelligent AI with human and societal values is a problem that is increasingly getting attention in the scholarly literature (Bostrom, 2014; Christian, 2020; Mechergui & Sreedharan, 2023; Russell, 2019; Søvik, 2022; Taylor et al., 2020, pp. 342–382). Besides technical questions on how values could be translated into machine code and incorporated into AI technology (Dignum, 2019), 'value alignment' involves both philosophical and societal aspects such as which values should be promoted and how to decide that (Savulesc et al., 2019; Smits et al., 2022; Hedlund, 2022).

For our future developer, we can only hope that these issues have been sufficiently solved. However, the point here is not to speculate whether that would be the case, but to illustrate the massive task ahead for our developer. While in principle, the developer's responsibility is the same as in the contemporary case, in practice, it magnifies with the scope of the intelligence of the AI.

Scenario 2: Super-intelligent AI that can act *as if* it understands what it is doing, why it is doing it, and consequences of what it is doing.

This super-intelligent AI shares the qualities with the super-intelligent AI in the first scenario, with the addition that it also has the ability to act as if it has moral agency. From a functional perspective, this is sometimes used as an argument for machine responsibility (Laukyte, 2017; Sullins, 2006). However, the fact that our future super-intelligent AI by interaction with the social world has learnt to imitate human interaction patterns does not lead us to the conclusion that it is a moral agent. Compelling biological, historical, and logical arguments speak against that (Gunkel, 2017; Hakli & Mäkelä, 2019; Sharkey, 2017). Hence, the fact that a super-intelligent AI can perform *as if* it is a moral agent does not make any morally relevant difference as compared to the first scenario, that is, the relation between our developer and this super-intelligent AI is qualitatively the same.

Like in the first scenario, the developer is responsible to see to it that the AI does not harm humanity or individual humans, and to do that, make sure that the AI continues to act according to the intentions of the developer. However, in addition to the demanding task of aligning the goals of the super-intelligent AI with human and societal values, the ability of this AI to act more humanlike will

put even more arduous requirements on the developer. There is a risk that the behaviour of this super-intelligent AI might fool our developer in different ways.

For instance, the developer might believe that the super-intelligent AI in fact have moral agency and delegate responsibility to it in ways that she already does to other humans. (Remember the extremely heavy burden that by now rests on the shoulders of our developer.) Given that the work with value alignment that we introduced in the first scenario have been perfectly successful, this need not be a danger, but any loophole in that work is something that the super-intelligent AI would exploit if that would increase its chances to reach whatever goals it has been given. This could jeopardise compliance of all the Laws.

Another risk is that the capacity of this super-intelligent AI to imitate humans may make it capable to disobey human orders, but act in a way that makes the developer believe that it has in fact followed her orders. Obviously, in this setting, the developer does not have control in any strong sense over the AI. But could she have the weaker form of control in the sense of supervision? Considering that it is the developer who maintains, improves, and teaches the AI, she is arguably responsible for this in a way comparable to the responsibility of a parent for its child. However, if the AI's disobeying leads to harm for humanity or individual humans, then the developer has not taken appropriate responsibility according to the Third Law directed at humans, and failed in regard to the hierarchy between the Laws directed at humans.

Scenario 3: Super-intelligent AI that *understands* what it is doing, why it is doing it, and consequences of what it is doing.

A super-intelligent AI that understands what it is doing, why it is doing it, and consequences of what it is doing is a system that is conscious. There is no undisputed definition of 'consciousness', but sentiments, inner mental life, inner experiences, and subjective experience approximately captures what it is about (Schneider, 2019; Tegmark, 2017). As we indicated above, consciousness is also a requirement for moral agency. In the current discussion on conscious AI, two main positions are discernible, one claiming that consciousness is unique for biological beings, the other that consciousness is substrate independent (Schneider, 2019; Tegmark, 2017). In this imagined future scenario, we assume that substrate independent consciousness is possible, meaning that also silicon-based artefacts like our super-intelligent AI could be conscious. What would that mean for the responsibility of our developer to see to it that the super-intelligent AI does not harm humanity or human individuals?

As a moral agent, the super-intelligent AI has the capacity to be responsible. Would that somehow interfere with the responsibility of the developer? After all, it is the developer who has created this super-intelligent AI, and if it is conscious, that would be a result of this creation. But as we know already from contemporary times, an intrinsic feature of AI is its learning and adaptation, and the consciousness of this super-intelligent AI might be the result of such adaption and not of the intention of the developer. However, the Third Law entails that adaptation should be constrained to the extent that the intention of the developer is protected. This would probably rule out the option that the developer has not

intended the super-intelligent AI to be conscious. Or would it? Perhaps superintelligence itself stands in the way for that option?

If the continuously increasing level of intelligence of the AI is accelerating exponentially, as Bostrom (2014) suggests, every new step of increasing intelligence is considerably larger than the previous one. Somewhere along this journey the developer might lose track. If consciousness emerges at a certain level of molecular complexity, as some 'techno-optimists' reason (Schneider, 2019), then consciousness could emerge even if that is not the intention of our developer. Then this scenario, with a conscious super-intelligent AI, is not the result of the developer's intention. The Third Law directed at humans never kicked in. Or, rather, the developer was unable to foresee the development and was thereby not able to take responsibility according to the Third Law directed at humans. For the same reason, she was not able to regard the hierarchy between the Laws directed at humans.

This tentative exposé of imagined future super-intelligent AI suggests that Asimov's Laws directed at humans would not be sufficient to protect humanity and individual humans in scenario 3, with a conscious super-intelligent AI, and, to some extent, in scenario 2, with a super-intelligent AI that can act as if it is conscious. The Laws would need to be complemented with some kind of precautionary principle. A precautionary principle is preemptive, tries to foresee the risks, and 'imposes some limits or outright bans on certain applications due to their potential risks' (Pesapane et al., 2018). However, the nature of AI systems, which adapt in an unforeseeable way according to their experiences and learning, makes this difficult, especially with regard to long-term development. For contemporary AI, the Laws directed at humans are more promising, at least for the particular case of protecting the functioning of democracy. This is, on the other hand, not a bad undertaking.

Conclusions

By taking Asimov's Laws of robotics as our point of departure, we tried out the option to direct them at humans and to incorporate the concept of 'responsibility' to do so. We applied our reformulation of the Laws directed at humans on a hypothesised real-case contemporary situation, namely, how recommender algorithms may damage the functioning of democracy.

This worked pretty well, given two specifying adjustments. In the Second Law, stating that a robot (or an algorithm) must obey human orders, a distinction between authorised and unauthorised humans was needed, and in the Third Law, we specified that the intention of the human creator of the robot (algorithm) must be protected. With these adjustments, we then applied the Laws to three imagined future scenarios, based on three significantly different and morally relevant directions of AI development: super-intelligent AI with no moral awareness, super-intelligent AI with the capacity to act as if is morally aware, and super-intelligent AI that is conscious and has moral awareness.

We found that the super-intelligent AI that lacks consciousness implies no principal problems for the human, albeit significantly larger responsibility in practical terms. As for the super-intelligent AI with the capability to act as if it has moral awareness, there is no morally relevant difference per se compared to the former case. However, the capability of this super-intelligence to act humanlike involves some risks, as the human developer might be deceived to believe that the super-intelligent AI is a moral agent, or that it is obeying the human when it in fact is not. If any of these outcomes materialises, then the Laws have not done their job. With the conscious super-intelligence, we come to the conclusion that the super-intelligence as such, due to is exponential development, makes the human unable to foresee the consequences, which makes her unable to take responsibility as stated in the Laws.

This tentative look into imagined future scenarios show that Asimov's Laws directed at humans seem to have a hard time in protecting humans and humanity when super-intelligent AI is or appear to be conscious, at least when the human is the only part responsible. Perhaps a viable way forward could be to hand over some of the responsibility to the AI? As it has moral awareness, or can act as if it has moral awareness, it would at least have the capacity to take responsibility. Could we imagine a future in which humans and the super-intelligent AI are responsible together? While this might look promising at a first sight, it would be a kind of collective responsibility, which gives rise to the problem of many hands. The problem of many hands refers to a situation when many agents contribute some part to an outcome that is dependent on each agent's contribution taken together. In such a situation, each of them is responsible for their own part, but none of the contributing actors is responsible for the entirety (Hedlund & Persson, 2024; van de Poel et al., 2015). This is a complicated situation when all involved agents are humans. When one of the agents is an artificial entity and many times more intelligent than the other, as in our imagined future scenarios, the involved agents stand in an asymmetrical relationship to each other, which would amplify the difficulties. Again, unless value alignment will be perfectly worked out, we cannot rule out the risk that the human will be deceived by the super-intelligent AI.

Finally, we would like to anticipate a possible objection to our approach to direct the Laws at humans. Considering the chosen examples of super-intelligent AI in our imagined future scenarios, it could be argued that we would not expect the Laws directed at humans to work. Given that our super-intelligence not only refers to intelligence that greatly exceeds the cognitive performance of humans but also to intelligence that is conscient, it is perhaps in the nature of things that humans would not have the capability to be responsible to protect humans and humanity from harm. The work with value alignment will have to start soon enough to be complete before we have a potential super-intelligence around us.

References

Abdullah, Y. I., Shuman, J. S., Shabsigh, R., Caplan, A., & Al-Aswad, L. A. (2021). Ethics of artificial intelligence in medicine and ophthalmology. *Asia-Pacific Journal of Ophthalmology*, *10*(3), 298.

Adam, B., & Groves, C. (2011). Futures tended: Care and future-oriented responsibility. *Bulletin of Science, Technology & Society*, *31*(1), 17–27.

AI HLEG. (2019). *Ethics guidelines for trustworthy AI.* Brussels: European Commission: High-Level Group on AI. Report No. B-1049. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Asimov, I. (1941). "Liar". *Astounding Science Fiction*. May issue.

Asimov, I. (1942). "Runaround". *Astounding Science Fiction*. March issue.

Asimov, I. (1985). *Robots and empire*. Doubleday.

Balkin, J. (2017). 2016 Sidley Austin distinguished lecture on big data law and policy: The three laws of robotics in the age of big data. *Ohio State Law Journal*, *78*(5), 1217–1242.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Buruk, B., Ekmekci, P. E., & Arda, B. (2020). A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine, Healthcare & Philosophy*, *23*(3), 387–399.

Choi, Y. L., Choi, E. C., Chien, D. V., Tin, T. T., & Kim, J.-W. (2019). Making of South Korean robot ethics charter: Revised proposition in 2018. In *ICRES 2019: International Conference on Robot Ethics and Standards*. London, UK, 29–30 July 2019. https://doi.org/10.13180/icres.2019.29-30.07.004

Christian, B. (2020). *The alignment problem: Machine learning and human values*. W.W. Norton & Company.

Clarke, R. (1993). Asimov's laws of robotics: Implications for information technology, part 1. *Computer*, *27*(1), 57–66.

Clarke, R. (1994). Asimovs laws of robotics: Implications for information technology, part 2. *Computer*, *27*(2), 57–66.

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer.

Duenser, A., & Douglas, D. M. (2023). *Who to trust, how and why? Untangling AI ethics principles, trustworthiness and trust.arXiv:2309.10318*.

EP. (2016a). European Parliament, Committee on Legal Affairs. *Draft report with recommendations to the Commission of civil Law Rules on Robotics*. 2015/2103(INL).

EP. (2016b). European Parliament, Legal Affairs. *European Civil Law Rules in Robotics*. PE 571.379.

EP. (2017). *European Parliament resolution "Civil Law Rules on Robotics"*. P8_TA (2017)0051.

Fischer, J. M. (1982). Responsibility and control. *The Journal of Philosophy*, *79*(1), 24–40.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

FitzPatrick, W. J. (2008). Moral responsibility and normative ignorance: Answering a new skeptical challenge. *Ethics*, *118*(4), 589–614.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. C., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI*. Berkman Klein Center for Internet & Society.

Gans, J. (2018). AI and the paperclip problem. *VoxEU* 10 June, 2018.

Gunkel, D. J. (2017). Mind the gap: Responsible robots and the problem of responsibility. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-017-9428-2

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *AI & Society*, *30*, 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hakli, R., & Mäkelä, P. (2019). Moral responsibility of robots and hybrid agents. *The Monist*, *102*, 259–275.

Hansson, S. O. (2012). Safety is an inherently inconsistent concept. *Safety Science*, *50*(7), 1522–1527.

Hedlund, M. (2012). Epigenetic responsibility. *Medicine Studies*, *3*, 171–183.

Hedlund, M. (2022). Distribution of forward-looking responsibility in the EU process on AI regulation. *Frontiers in Human Dynamics*, *4*, *703510*. https://doaj.org/article/1f56de7f628e405da7118151635101d3

Hedlund, M., & Persson, E. (2024). Expert responsibility in AI development. *AI & Society*, *39*(2), 453–464. https://doi.org.10.1007/s00146-022-01498-9

Held, V. (1970). Can a random collection of individuals be morally responsible? In May, L., & Hoffman, S. (Eds.) *Five decades of debate in theoretical and applied ethics*. Rowman & Littlefield Publishers, Inc., pp. 89–100.

Himmelreich, J. & Köhler, S. (2022). Responsible AI through conceptual engineering. *Philosophy and Technology*, *35*, 1. https://doi.org/10.1007/s13347-022-00542-2

Kaminka, G. A., Spokoini-Stern, R., Amir, Y., Agmon, N., & Bachelet, I. (2017). Molecular robots obeying Asimov's three laws of robotics. *Artificial Life*, *23*, 343–350.

Kane, R. (2002). Responsibility, reactive attitudes and free will: Reflections on Wallace's theory. *Philosophy and Phenomenological Research*, *64*(3), 693–698.

Laukyte, M. (2017). Artificial agents among us: Should we recognize them as agents proper? *Ethics and Information Technology*, *19*, 1–17.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*, 175–183.

Mechergui, M., & Sreedharan, S. (2023). *Goal alignment: A human-aware account of value alignment problem*. *arXiv:230200813v2*.

Mill, J. S. (1859/2011). *On liberty*. Cambridge University Press.

Murphy, R. R., & Woods, D. D. (2009). Beyond Asimov: The three laws of responsible robotics. *IEEE Intelligent Systems*, *24*(4), 14–20.

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human-robot collaborations and responsibility. *Science and Engineering Ethics*, *24*, 1201–1219.

Pasquale, F. (2017). Toward fourth law of robotics: Preserving attribution, responsibility, and explainability in an algorithmic society. *Ohio State Law Journal*, *78*(5), 1245–[iv].

Persson, E., & Hedlund, M. (2024). The Trolley problem and Isaac Asimov's first law of robotics. *Journal of Science Fiction and Philosophy*, *7*. https://jsfphil.org/volume-7-2024-androids-vs-robots/asimovs-first-law-and-the-trolley-problem/

Pesapane, F., Volonté, C., Codari, M., & Sardanelli, F. (2018). Artificial intelligence as a medical device in radiology: Ethical and regulatory issues in Europe and the United States. *Insigths into Imagining*, *9*(5), 745–753.

Russell, S. (2019). *Human compatible: Artificial Intelligence and the problem of control*. Viking.

Savulescu, J., Kahane, G., & Gyngell, C. (2019). From public preferences to ethical policy. *Nature Human Behaviour*, *3*, 1241–1243.

Schneider, S. (2019). *Artificial you: AI and the future of your mind*. Princeton University Press.

Schurr, N., Varakantham, P., Bowring, E., Tambe, M., & Grosz, B. (2007). Applying laws of robotics to teams of humans and agents. In R. H. Bordoni, M. Dastani, & J. Dix (Eds.), *Programming multi-agent systems* (pp. 41–55). Springer.

Scroxton, A. (2023). Hackers: We won't let AI get the better of us. *Computer Weekly*, *8*(1), 3–7.

Sharkey, A. (2017). Can robots be responsible moral agents? And why should we care? *Connection Science*, *29*(3), 210–216.

Smits, M., Ludden, G., Peters, R., Bredie, S. J. H., van Goor, H., & Verbeek, P.-P. (2022). Values that matter: A new method to design and assess moral mediation of technology. *Design Issues*, *38*(1), 39–54.

Sneddon, A. (2005). Moral responsibility: The difference of Strawson, and the difference it should make. *Ethical Theory & Moral Practice*, *8*(3), 239–264.

Søvik, A. O. (2022). What overarching ethical principle shoulc a superintelligent AI follow? *AI & Society*, *37*, 1505–1518.

Sullins, J. P. (2006). When is a robot a moral agent? *International Review of Information Ethics*, *6*(12), 23–36.

Talbert, M. (2008). Blame and responsiveness to moral reasons: Are psychopaths blameworthy?. *Pacific Philosophical Quarterly*, *89*(4), 516–535.

Taylor, J., Yudkowsky, E., LaVictoire, P., & Critch, A. (2020). Alignment for advanced machine learning systems, *Ethics in Artificial Intelligence* (pp. 342–382). Oxford University Press.

Tegmark, M. (2017). *Life 3.0: Being human in the age of Artificial Intelligence*. Penguin Random House.

Thompson, D. F. (1987). *Political ethics and public office*. Harvard University Press.

Torres, E., & Penman, W. (2021). An emerging AI mainstream: Deepening our comparisons of AI frameworks through rhetorical analysis. *AI & Society*, *36*, 597–608.

Van de Poel, I., Royakkers, L., & Zwart, S. D. (Eds.) (2015), *Moral responsibility and the problem of many hands*. Routledge.