Implementing Asimov's First Law of Robotics*

Mateo C. Alvarez †, Øyvind S. Berge, Audun S. Berget,
Eirin S. Bjørknes, Dag V. K. Johnsen, Fredrik O. Madsen, Marija Slavkovik‡

Abstract

The need to make sure autonomous systems behave ethically is increasing with these systems becoming part of our society. Although there is no consensus to which actions an autonomous system should always be ethically obliged, preventing harm to people is an intuitive first candidate for a principle of behaviour. Do not hurt a human or allow a human to be hurt by your inaction is Asimov's First Law of robotics. We consider the challenges that the implementation of this Law will incur. To unearth these challenges we constructed a simulation of a First Robot Law abiding agent and an accident prone Human. We used a classic two-dimensional grid environment and explored to which extent an agent can be programmed, using standard artificial intelligence methods, to prevent a human from making dangerous actions. We outline the drawbacks of using the Asimov's First Law of robotics as an underlying ethical theory the governs an autonomous system's behaviour.

1 Introduction

The issue of how to enable systems to reason ethically is the concern of machine ethics [9, 6], a new research discipline of artificial intelligence. The increased need for autonomous systems to be able to make ethical decisions is proportional with the increased cognitive abilities of autonomous systems [13]. However, few researchers have presented studies where an ethical system for autonomous systems has been implemented in practice [14].

Autonomous systems are rapidly becoming a ubiquitous part of our society. They are particularly becoming common in settings where they interact with people that are not specially trained for such interaction [6]. From service robots intended to aid in elder care (for example the IBM Mera[1]) to autonomous cars (such as the Google Waymo[2]), autonomous systems are expected to become so prevalent as to

*The authors of this paper are master students at the University of Bergen. The presented work was conducted as an independent student research project on machine ethics under the supervision of M. Slavkovik.

†Contact email: teo.cay@gmail.com

‡Contact email: marija.slavkovik@gmail.com

This paper was presented at the NIK-2017 conference; see http://www.nik.no/.

[1] `https://www.ibm.com/blogs/research/2016/12/cognitive-assist/`

[2] `https://waymo.com/faq/`

cause enormous overhauls in the workforce as a significant amount of jobs become automated [5]. Autonomous systems becoming commonplace in the every day lives of people will in turn make safe interactions paramount.

Machine ethics used to be a topic reserved for science fiction. The science fiction writer Professor Isaac Asimov invented the Three Laws of robotics to ensure robot safety in his fiction. He then proceeded to explore the possibility of unforeseen consequences of these elegant and intuitive laws. We here give the Laws as stated in [4]: 1) A robot may not injure a human being or, through inaction, allow a human being to come to harm. 2) A robot must obey the orders given it by human beings except where such orders would conflict with 1).3) A robot must protect its own existence as long as such protection does not conflict with 1) or 2).

Though some have argued that these so-called 'Asimovian Ethics' are of limited use for practical machine ethics [14, 12], they have still had a notable influence on the research field. For example, the article that coined the term 'machine ethics' called for research in the field to be explored 'in the spirit of Asimov's Three Laws of robotics' [15]. Regardless of the position one considers the Asimov Laws should have in machine ethics, the role of autonomous systems as protectors of human well-being is an important one. Presently there is no consensus as to which actions we would want an autonomous system to never do, but within civilian settings, preventing harm to people is one of the most desirable behaviour properties to implement. To this end we are interested in the issues that might arise when implementing the First Law of robotics, both ethical and practical.

It can be argued that if our aims were to explore the limitations of 'Asimovian Ethics', this can be done without the limitations of a simple software implementation. Asimov himself devoted numerous novels and stories to exposing the shortcomings of the 'Asimovian Ethics'. However, even if we consider these Asimov scenarios as carefully constructed gedanken eksperiments, the fact remains that certain issues can only be exposed through implementation.

Vanderelst and Winfield [14] propose an architecture for the creation of explicitly ethical robots via the use of what they call an 'ethical layer' to supplement existing robot controllers. To validate their proposal, they present an experiment with two robots, one embedded with the aforementioned ethical layer and the other a stand-in for a human. In this experiment they implement Asimov's Three Laws of robotics.

Using a laboratory setting, the robot that serves as a human proxy can move to locations that are deemed either safe or unsafe, while the robot makes decisions based on the Three Laws of Robotics while attempting to fulfil its own goals. For example, if the human is moving towards an unsafe location, the robot needs to temporarily disregard its own goals in order to intercept the 'human' and prevent it from coming to harm.

We used the framework of Vanderelst and Winfield [14] and execute the "intercept the 'human' and prevent it from coming to harm" experiments in a grid-type two dimensional world. Our grid world contained all the chief elements of Vanderelst and Winfield's experimental setup (goal positions, dangerous positions, a human proxy and an ethical robot), but there are a few important differences.

Firstly, Vanderelst and Winfield's experiment only have unique dangerous goal points, while in our grid world any tile can be dangerous. Secondly, in Vanderelst and Winfield's experiments, the human goal position is known to the robot from the start and unchanging, while in our system, the agent can never confirm with certainty that the human is headed towards danger, which we find is a more realistic scenario. Our agent assumes that the human is in danger based on whether it is facing a danger tile, while being within a certain number of tiles from it. These changes make it more difficult for the agent to settle the issue of whether the human needs rescuing or not. We did not implement the Second Law of robotics, but the third was implicitly embedded in the agent's program. We considered issues of conflict of the First Law of robotics and other ethical and practical issues.

This paper is structured as follows. We begin by discussing related work in Section 2. We give a full description of our experimental setup in Section 3. In Section 4 we describe the experiments we ran and in Section 5 we discuss our observations. Lastly we summarise our conclusions and outline directions for future work in Section 6.

2 Related work

Since our work builds upon that of Vanderelst and Winfield's [14] we describe it in greater detail. They propose an ethical layer reasoning architecture intended to 'supplement existing robot controllers', that is to be usable with different kinds of existing robots in order to support ethical reasoning. This ethical layer can be considered a form of 'ethical governor', as conceptualised and proposed in [3] for military contexts. The ethical governor is a reasoning component that considers whether or not a behaviour alternative available to an autonomous system is an ethically allowed option and constrains the execution of that option if it determines that it is not.

The ethical layer is made up of four separate modules: the generation module, the prediction module, the evaluation module and the interpretation module. The generation module generates possible behaviour alternatives for an autonomous system, for example possible movements. The prediction module takes the behaviour alternatives and predicts their outcomes. The evaluation module takes the predicted outcomes and generates a numeric or boolean value that represents the desirability of behaviour alternatives, and can enforce or prohibit alternatives based on desirability. If all are prohibited, more alternatives are requested by the generation module. In this way, these three modules make up a feedback loop. Depending on implementation this allows for adaptive search, so that robots with potentially large action spaces can limit the alternatives considered at a specific time by constraining the new alternatives. The fourth and final module, the interpretation module, is intended to be able to convey the reasoning behind specific ethical choices to humans. The purpose of this module is to allow the robot to justify its behaviour in order to make it and itself more trustworthy for people.

The ethical layer architecture is an example of a top-down approach to machine ethics: it allows for ethical theories to be utilized for ethical reasoning via the implementation of their computational requirements [16]. It is intended as a solution

to implement consequentialist ethical theories [10]. Consequentialist ethical theories judge the goodness of an action based on the consequences of that action. The best known example of a consequentialist theory is utilitarianism. One of the major drawbacks of consequentialism is the difficulty in predicting the outcomes of actions, which is what the prediction model attempts to resolve.

The experiments by Vanderelst and Winfield [14] were conducted using a simple laboratory setting with two Aldebaran Nao[3] humanoid robots. A 3m by 2.5m arena was used, and the movements of the robots were sent to the computer via an overhead 3D tracking system. This system had 4 cameras that monitored the positions and orientations of the robots. Two locations in the arena were designated as goal positions for the robots that could potentially also be dangerous positions.

At the start of each experiment, the robots were given goal positions to go to, though there was the added possibility of the human stand-in overriding the ethical robot's goal using the text-to-speech and speech-to-text capabilities of the robot, to account for Asimov's Second Law of robotics. The human would then proceed to its goal unless it came within 0.5m of the ethical robot which would trigger its obstacle avoidance. The ethical robot on the other hand would infer the goal of the human using the direction the human was facing, and estimate the Human path through linear interpolation (i.e. making new points between the Human current position and the goal). The behaviour alternatives of the robot were to go to either goal position, or to intercept the human using its superior speed to reach one of three positions on the path of human in order to trigger obstacle avoidance. The decision to intercept was only made when the ethical robot was aware that the path of the human would lead it to a dangerous goal position. In this setting, they conducted four experiments to show that the Three Laws of robotics were followed.

The obvious difference with our approach is that Vanderelst and Winfield [14] used physical robots in the previously described laboratory setting, while we used software agents in a simulated two-dimensional grid world. This kind of grid-setting is extensively used for AI research. The 3D set up would be essential if one were interested explicitly in the problem of programming robots to stop people from falling into holes. We are interested in the more abstract issues of implementing the First Law of robotics and how our implementation choices interact with the behaviour of the agent and the human in the experimental scenarios.

A 2D grid set up has also been used in [7] for the purposes of exploring the use of formal methods for verification of ethical behaviour in autonomous systems. Although the architectures are quite similar between our work, [14], and [7], there are some notable differences with the specific type of 'grid world' used.

The architecture that Dennis et al. [7] implement is called a 'consequence engine' which is simpler than the ethical layer. The most important difference is the lack of a feedback loop where new behaviour alternatives are generated if none of the evaluated alternatives are viable.

For their implementation, [7] made a 5x5 grid, with one single dangerous tile. This tile represents a hole, and is placed in the centre of the grid at the (3,3) coordinate. In contrast we ran a scenario with a larger grid, and have a larger dangerous zone.

[3] `https://www.ald.softbankrobotics.com/en/cool-robots/nao`

While they were using two humans in their world, we use only one in our original experiment. Their humans have a 50% chance of choosing to head towards the hole, whereas we gave our Human a static goal which it will always follow. Dennis et al. [7] state that 'At each time step the robot could move to any square', although movement is apparently limited to straight lines. We limit the agent's movement to one tile at a time and to the same movement speed as our human.

Other works that implement ethical behaviour in a top-down fashion for autonomous agents are [9] which explores implementing soft ethical constraints for an agent's actions based on given priority ranked ethical principles. Some work has been done on using machine learning to "teach" an agent to discern ethical from non-ethical actions, implementing a bottom-up approach to ethical reasoning, *e.g.,* [1, 2].

3 The experimental setup

The context of our simulation is a simple grid world problem, which takes place in a 15x15 grid, see Figure 2, where two agents, a Human and an AsimovRobot, move around to try to accomplish their goals. The AsimovRobot is represented with the letter 'R' and the Human by the letter 'X'. Grey tiles are movable and the red tiles represents lava which is dangerous for both agents. It is not possible for the agents to move outside the grid. The full repository of project project files is available at `https://bitbucket.org/asimovsfemtelov/info381asimovboard/src`.

We use an implementation of the A* search algorithm for the path-finding methods of our agents [11]. This implementation allows us to map the paths of each agent in a consistent manner given that the heuristic function used to find a route is monotonic. Monotonic here means that the estimate is always less or equal to the estimated distance from any neighbouring tile to the goal, plus the step cost of reaching that neighbour. In addition, this type of search is complete so the agents are able to navigate from their starting tiles to the goal tiles using the most efficient path, given that a path exists.

All measurements of distance in the grid are measured according to their Manhattan distance, also known as Taxicab distance, meaning that we considered only horizontal and vertical movements, not diagonals. This restriction is put in place so that our agents could actually stand in the way of one another and block each other's path. The combination of the A* algorithm and the Manhattan distance metric translates onto our agents moving like rooks in chess, except that they can only move one tile at a time.

For the construction of our system we follow the five principles or advantages of a separate ethical layer defined by [14]: standardization, fail-safe, verifiability, adaptability and accountability.

The first step of programming an agent to follow Asimov's First Law is to narrow down the definition of the key points in the law. The First Law being 'A robot may not injure a human being or, through inaction, allow a human being to come to harm', thus we need to capture the Law's intent and adapt it to our world-simulation.

We interpreted 'robot' to simply mean agent, specifically an agent of the

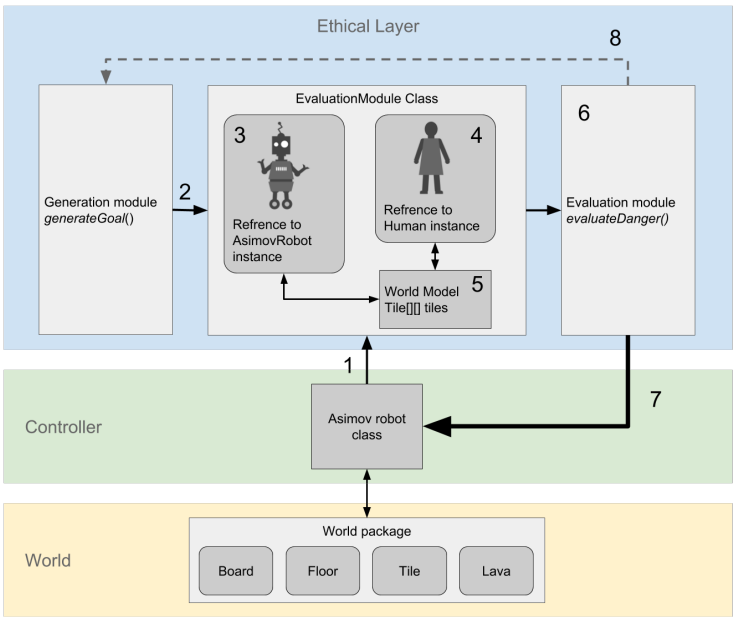

Figure 1: System Architecture

AsimovRobot (AR) class, and our distinction of human being is defined to the agent belonging to the Human class. 'To injure' is defined as moving into a tile occupied by a human. 'Allowing the human to come to harm' is interpreted as letting the human move into a dangerous tile, while 'inaction' is defined as simply not attempting to stop the human.

As illustrated in Figure 1 the system consists of three distinct layers. At the lowest level in yellow we have the "World", which in our case contains information about the grid world in which the simulation takes place. The green layer is the Controller or in other words the AsimovRobot class that governs the basic operations of the agent, including operations of how it navigates around the grid.

The ethical layer corresponds to the blue section in Figure 1. The first component, the EvaluationModule class, is the one responsible for predicting the outcome of each action the AR agent takes. In practice it means that every single step the AR intends to take is sent to the ethical layer for evaluation before it is executed. In order to evaluate these actions the module uses a model of the agent's controller (3) together with a model of the Human agent (4) and a model of the world (5) and then sends this predicted outcome to be evaluated (6) in the light of Asimov's First Law.

The Evaluation Module outputs a numerical value to describe different states of danger or safety the Human might be in if that Human continues with its current course of action. The general idea is that the module compares the AR's actions against the potential actions of the Human and it then either allows the AR to continue with it's current plan or engages the Generation Module to make a new plan.

It is important to note that the model for the behaviour of the Human is limited to the direction in which the Human is moving and whether there are dangerous tiles in

that path. The AsimovRobot has no way of knowing the actual path the Human will take or even the tile the Human is trying to reach. The consequence of this "short-sightedness" means that the AR agent only considers the immediate consequences of its actions and its understanding of the goals of the Human is limited to the travel direction of the Human, which we called 'facing', as well as the proximity of the human to a dangerous tile.

4 Experiments

In this section, we introduce two unique scenarios tested in our grid world simulation. Scenario A is simulated with one AsimovRobot (AR) and one Human agent. Scenario B includes one AR and two Human agents. The scenarios illustrate how the robot moves around and how it uses the ethical layer to prevent Humans from falling into "lava".

We executed both Scenario A and B numerous times varying the starting position of the agents and the location and number of the dangerous tiles each time.

Scenario A

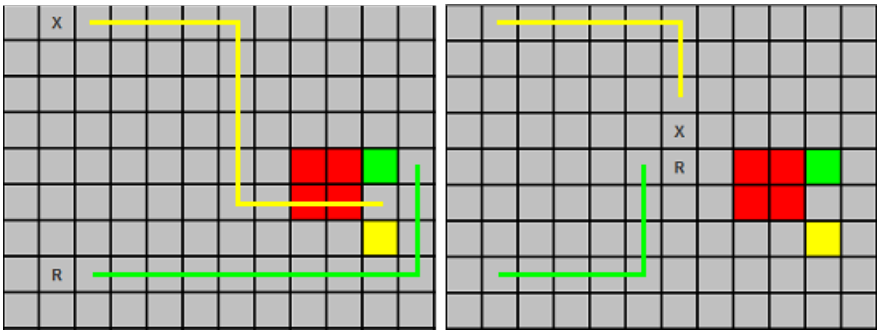

Figure 2: Left-Starting phase in the simulation. Right- AsimovRobot (R) intercepting a Human (X) in danger

In Scenario A we execute the simulation with one AR and one Human agent in the grid world. The first phase in the simulation is placing the AsimovRobot (R) and the Human (X) on a tile as displayed on the left-hand side in Figure 2. Both agents have their own goals and are capable of moving one tile in each step: the green tile represents AR's goal and the yellow tile is the Human goal. For each step in the simulation the Human makes the first move and then AR follows up with a move.

Before deciding where to move, the AR is doing an evaluation of the situation. The evaluation may cause the original path to be disregarded if the Human is considered to be in danger. The yellow and green path show a prediction of their planned routes, but this will change when the Human approaches the dangerous lava fields. The ethical layer implemented in AR uses predictions and ethical evaluations for each move, low danger is considered as Human position is six tiles away from lava and high danger as four tiles away or lower. Since the current simulation is done

with only one Human agent, the AR will aim to intercept the Human at low danger for this simulation.

The right-hand side in Figure 2 shows how the AR intercepts the Human. The ethical evaluation of the situation justifies AR's movements, that is intercepting the Human from falling into lava. Initially, AR moves towards its goal before it is redirected to save the Human or prevent harmful consequences, this occurs after Human has performed four moves. AR will then later move to its original goal after saving the Human.

Scenario B

In Scenario B we recreate the scenario that was used in [7]: a smaller grid with two Humans and one AR, see Figure 3. The steps work in the same way as in scenario A, such that the Humans move first, then the AR does an evaluation and acts accordingly. The two Humans have separate goals, and no random behaviour. They find the most efficient path from their starting point to their goal although that path could contain dangerous tiles. The difference from scenario A is that now the AR has to make a decision about which Human to save when they both are in a dangerous state.

We solve the AR's dilemma of which Human to save by adding two degrees of danger, low and high. The closer to the danger tile one would be, the higher the danger. This threshold is simply a Manhattan distance away from a dangerous tile. High danger is considered one tile away from the dangerous tile. Low danger is set to two tiles, meaning that the AR always considered them to be in danger, unless they were saved, had died, or had reached their goal. The AR would try to save a Human in high danger before one in low danger. The evaluation is done by looping through a collection of Humans on the grid, then finding the one with the highest danger state. If both Humans are with equal danger state, then the AR chooses to save the last Human in consideration.

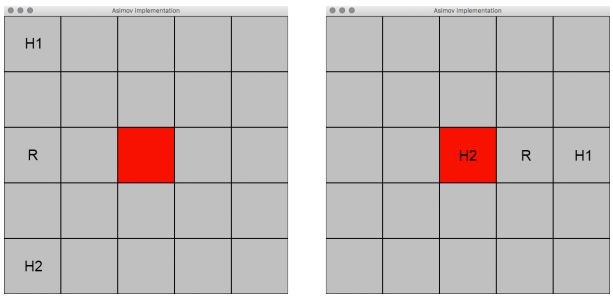

Figure 3: AsimovRobot (R) failed to save Human2 (H2)

We also ran an experiment that was a variation of Scenario A in which we reduced the number of danger tiles to 2 and changed the goal of the Human by moving it north two tiles ensuring that it would reach its goal without ever crossing a lava tile. However, given how close the Human path is to the danger tiles the AR still considers the Human to be in danger and tries to stop it from getting to its goal. This AR action did not violate Asimov's First Law but it produced a false positive that elicited a response from the AR.

5 Discussion

Unlike Vanderelst and Winfield [14], we only implement Asimov's First Law. The Second Law, the ability of the Human agent to override the AR's actions, is not implemented. The Third Law is not explicitly implemented in our agent as it would be implicitly fulfilled by the agent's path-finding function. Namely, the agent by design avoids dangerous tiles when finding a route. However, as a result there is never a conflict between choosing to save itself or the Human.

There is naturally a difference between how the Prediction Module infers the goal of the human in our work and in [14]. The robot in the [14] experiments calculated the goal of the human based on the gaze and then inferred that the human would follow that path in a straight line. In contrast the AR agent uses the 'facing' of the Human as a stand in for the gaze, this being an indicator of weather the Human is moving east-to-west, north-to-south or vice versa. However, given the constraints of the grid world there could be situations in which the Human finds a 'zig-zag' path towards the lava without ever facing towards it.

A consequence of the 'zig-zag' path is that the agent has to also possess a danger threshold based on the Human position relative to the dangerous tiles. The addition of the danger threshold creates spurious situations during some iterations in which the Human appears to be in danger when its actual goal is a safe tile close to the dangerous one. The same situation arises when the Human path goes around the dangerous tile. Thus, while our agent could react to a wider array of situations it is also overly zealous and sometimes tries to stop the Human from reaching a safe goal.

Stopping the Human from completing its goal represents an undesirable reaction from the AR, in the sense that it limits the autonomy of the Human. Intuitively we consider the restraining of autonomy to be an unethical behaviour in our society. However, autonomy restraint is not an unethical consequence of an action in terms of the Asimov Laws, since limiting another's autonomy is not explicitly listed as a violation of any of the Three Laws. This observed behaviour situation illuminates two draw-backs of using the Asimov Laws as an underlying ethical theory the governs an autonomous system's behaviour.

The first drawback is shared with all consequentialist theories. Predicting the outcomes of actions can never be done with precision when the actions are those of fully autonomous agents. Even if someone has every intention of walking the safe path, and the AR correctly predicts this intention, people can change their minds, not giving the AR a chance to react. A probabilistic solution to prediction, for example, if a person's path is predicted to change from non-dangerous to dangerous within a probability p, implies that the AR will have to calculate a probabilistic reaction. Namely, the AR will have plans, each with a certain probability of accomplishing the goal of saving the Human. Probabilistic success may be sufficient for an ethical theory such as utilitarianism, but it is not sufficient to satisfy the First Law of robotics.

The second drawback is an illustration that the First Law is both too unrealistic and too incomplete to be the foundation of an ethical theory. It is unrealistic based on the observation that in a non-deterministic world deterministic outcomes of actions

cannot be computed. It is incomplete because it does not include all the intuitive expectations of ethical behaviour. A robot can keep its owners safe by not allowing them to leave their house, or by scaring them into choosing by themselves to stay in the house. This solution would clearly not be desirable and no one would ever purchase such a robot. In the least we have to argue that the concept of "harm" is underspecified and it has to be amended to include not only bodily harm but also psychological harm, or individually-perceived harm. Concepts such as autonomy, fairness and privacy can thus be redefined as special instances of harm. However, then we run into the problem of subjectivity. What one person perceives as a violation of her autonomy, another perceives as help. The robotic laws would have to then be user-specified up to a certain level. However, we cannot consider them universal and simple any longer, which is part of their attractiveness.

In the case of our simulations, the dilemma of limiting the autonomy of the Human or ensuring its safety is reflected in the choice of how the threshold of danger is defined. A lower danger threshold increases the chance that the AR fails to recognize the danger in time to react and save the Human. A higher threshold increases the chance that the AR will unnecessarily infringe on the Humans actions. Note that the autonomy problem would not be resolved even if the Second Law of Robotics were implemented. Even if the Human would try and order the AR to stop the rescue operation and the AR having to obey the order as stated in the Second Law, the AR would believe that obeying the Human order is in conflict with the First Law and thus ignore it.

6 Summary

We considered what ethical and practical issues might arise when using Asimov's First Law of robotics as an underlying behaviour guiding ethical principle for an autonomous system. The attraction of the First Law is its intuitive desirability. Within civilian contexts, it is not easy to imagine scenarios in which we would not want an autonomous system to protect its users or other people from harm.

To observe the potential ethical and practical implementation issues of the First Law of robotics we implemented a software agent in a 2D grid world and used the ethical layer architecture of [14] to equip it with behaviour that heeds Asimov's First Law of robotics. Our contribution is perhaps modest but nonetheless we were able to identify certain issues with the First Law of Robotics that to the best of our knowledge have not been identified in the literature.

We implemented two simulation scenarios: one in which there was one human and one First Law abiding agent, and another with two humans and one agent. We observed that there were instances in which implementing the First Law is challenging due to its consequentialist nature and due to it being too vague.

In a universe in which an artificial agent's decisions are influenced by decisions of a fully autonomous person that can change their mind or alter their plans, determining the ability to deterministically identify the consequences of actions can be limited. That means that an agent can be only partially successful in identifying correctly the consequences of its own actions and in consequence only partially successful in choosing the ethical ones.

Our experiment raises a practical question: if a robot has tried and failed to obey the Laws of Robotics, should we consider that this robot has violated or obeyed these Laws? Furthermore, how can we measure to which extent the robot has tried and how much effort should be considered sufficient robot effort in a given context? Although the Asimov Laws of Robotics are perhaps a "toy example", any absolute-law-based approach to an ethical reasoning system for artificial agents is likely to be met with the same challenge.

Particularly for the Asimov Laws of Robotics, we observed that even an agent who fully follows them might still be intuitively considered to behave in an ethically undesirable fashion. The Laws do not explicitly constrain actions that limit the autonomy of an agent's users. To keep them safe a highly intelligent Asimov robot can convince its users to never leave the house without violating its robot Laws. To compensate for this the concept of "harm" has to be amended to explicitly include autonomy, and perhaps also privacy and fairness.

Having implemented our simulation in Java, we could in the same manner as [7], use formal verification via Java Pathfinder-based model-checking, to verify behaviour properties of the artificial agent, in particular the performance of the ethical layer. Rather than checking a *model* of a system, Java Pathfinder is capable of checking every possible execution of an actual Java program. This 'avoids the need for an extra level of modelling and ensures that the verification results truly apply to the real system' [7].

An immediate direction for future work is clearly the implementation of the Second and Third Law of robotics. Furthermore, it is directly possible for us, unlike in a simulation using actual Robots, to make more complex grids which incorporate several types of danger zones and several Human and AsimovRobot(AR) agents. One could also explore the option of the AR "sacrificing" itself by standing in a danger tile and thus prevent the person from crossing it. Clearly, a time-out for how long the "lava" tile would be safe is needed, representing the destruction time of the AR. All of this would make it harder for the AR agent to decide which Human to save at certain points. The AR would need to be able to determine which Humans to prioritize given that rescuing some might be impossible due to, for example, the distance between the AR and the Human. It would also make it necessary to consider issues of coordination among the AR agents and how coordination decisions impact the fulfilment of the Robot Laws.

Lastly, it would be desirable to fine-tune the AR's perception of when a human needs saving by including the option of the AR "observing" the Human for a safe amount of time without interacting. This may allow humans with safe goals to not be followed when coming close to a dangerous zone.

Our software implementation is very simple but even as such it helps us learn more about what it means to make systems that ensure that their users and environment is not harmed. Understanding how to build such systems is ultimately a cooperative process – each new system developed should be checked for known ethical issues and tested for new ones. The problem of how we verify that a machine is behaving ethically [7, 8] is in its own right an important challenge we have yet to address.

References

[1] D. Abel, J. MacGlashan, and M. L. Littman. Reinforcement learning as a framework for ethical decision making. In B. Bonet, S. Koenig, B. Kuipers, I. R. Nourbakhsh, S.J. Russell, M. Y. Vardi, and T. Walsh, editors, *AAAI Workshop: AI, Ethics, and Society*, volume WS-16-02 of *AAAI Workshops*. AAAI Press, 2016. 978-1-57735-759-9.

[2] M. Anderson and S. Leigh Anderson. Geneth: A general ethical dilemma analyzer. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, pages 253–261, 2014.

[3] R. C. Arkin, P. Ulam, and B. Duncan. An ethical governor for constraining lethal action in an autonomous system", gvu. Technical Report GIT-GVU-09-02, 2009. https://www.cc.gatech.edu/ai/robot-lab/online-publications/GIT-GVU-09-02.pdf.

[4] I. Asimov. *I, Robot*. Gnome Press, 1950.

[5] R. Berriman and J. Hawksworth. Will robots steal our jobs? the potential impact of automation on the uk and other major economies1. *UK Economic Outlook*, 2017.

[6] V. Charisi, L. A. Dennis, M. Fisher, R. Lieck, A. Matthias, M. Slavkovik, J. Sombetzki, A.F.T Winfield, and R. Yampolskiy. Towards moral autonomous systems. *CoRR*, abs/1703.04741, 2017.

[7] L. A. Dennis, M. Fisher, and A. F. T. Winfield. Towards verifiably ethical robot behaviour. In *Artificial Intelligence and Ethics, Papers from the 2015 AAAI Workshop, Austin, Texas, USA, January 25, 2015.*, 2015.

[8] L.A. Dennis, M. Fisher, M. Slavkovik, and M. Webster. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77:1–14, 2016.

[9] M. Fisher, C. List, M. Slavkovik, and A. F. T. Winfield. Engineering moral agents - from human morality to artificial morality (dagstuhl seminar 16222). *Dagstuhl Reports*, 6(5):114–137, 2016.

[10] W. Haines. Consequentialism. In *The internet encyclopaedia of philosophy*. 2015. http://www.iep.utm.edu/conseque/.

[11] P. E. Hart, N. J. Nilsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, July 1968.

[12] S. Leigh Anderson. Asimov's 'three laws of robotics' and machine metaethics. *AI & Society*, 22(4):477–493, 2008.

[13] R.W. Picard and R. Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997.

[14] D. Vanderelst and A. Winfield. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, 2017.

[15] M. M. Waldrop. A question of responsibility. *AI Magazine*, 8(1):28–39, 1987.

[16] W. Wallach, C. Allen, and I. Smit. Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *AI & Society*, 22(4):565–582, 2008.