The Future of Teaching? Asimov's Three Laws and the Hypothetical Robot Teacher

Nicola Robertson

School of Education, University of Strathclyde, Glasgow, UK (n.robertson@strath.ac.uk)

Received: 03/12/2020
Accepted for publication: 14/06/2021
Published: 10/10/2021

Abstract

There is no denying that the influence and use of technology in relation to teaching and learning increased significantly during the Co-Vid-19 periods of isolation and lockdown. The screen became the classroom; the teacher (and the students), rendered as apparitions of virtuality. Nevertheless, despite the barriers of distance and screen, there remained (and indeed remains) something distinctly human about these interactions. What if the teacher on the screen – and, indeed, in the classroom – was not human? Remotely controlled robotic teachers have been trialled in China, with positive feedback from students; yet teaching remains a profession that has been deemed at low risk of automation. This paper will consider Isaac Asimov's three laws of robotics as a foundational base for predicting the behaviour of a potential, autonomous, robot teacher; comparing the predictions in relation to behaviours deemed as necessary for the successful practice of teaching. To do this, the paper will set out the three hypothetical scenarios, in order to explore – and hopefully determine – whether a 'robot' could effectively carry out key teaching activities. The speculative responses to these questions will hopefully inspire further discussion and discourse.

Keywords: Teaching, robotics, technology, Asimov

1. Introduction

The ubiquity and use of digital technology has been on a steady, and increasing, trajectory. Gordon Moore famously predicted that this would be the case, when he argued in 1965, that machines would continue to get smaller, cheaper and more powerful every two years (Rotman, 2020). Already immersed in a world of education technology, the drastic Co-Vid-19 related events that unfolded in 2020 saw us migrate further (in fact entirely) to a digital educational world; an ersatz, alternative to the more traditional and physical worlds of which we were more familiar. The effect that this rapid transition had on our relationship with the digital remains to be fully understood; however, it is safe to confidently assert that digital technology played a dominant role in our professional and educational lives during this time.

In education, this has been keenly felt, and continues to so, as some institutions incorporate models of blended learning, or entirely online teaching. The educational places that we recognised prior to Co-Vid-19 were (during the 2020 and 2021

lockdowns) replaced by screens; students and colleagues rendered as pixelated apparitions on those screens. Nevertheless, provided that we do not descend into solipsism or Matrix-inspired conceptualisations of the universe, we can be fairly confident that the human beings that we are interacting with via the screen were real: machines became mediators of human connection.

The question that I will consider in this paper concerns this removal of direct human contact. Taking the Co-Vid-19 technology based developments in a more speculative direction, I will consider whether it could be possible for technology – in the form of an educating robot – to take the place of a real human teacher? While teaching remains an occupation at low risk of automation (ONS, 2019), trials of robot teachers have taken place in Japan (Hashimoto, Verner and Kobayashi, 2012), and South Korea (Rebora, 2011); although in both of these cases the robots were remotely controlled by a human teacher located nearby. Setting my sights on potential future developments in both the fields of education and robotics, my intention is to discuss the possibility of autonomous robots – not under direct human control – being developed to assume a teaching role.

2. Isaac Asimov and the Three Laws of robotics

Since fully autonomous robots do not yet exist, we can must rely on speculation to consider and explore how they may behave; at most, extrapolating from the machines that we currently have in use. Robot / machine ethics is a continuing and developing field of inquiry within academia, and inspires debate and discourse as wildly diverse as that found in (human) ethics. For this paper, I have chosen to supplement academic discourse with the imaginative offerings of science-fiction. Given that we are dealing with a scenario that we have yet to – and indeed may never – encounter, speculation using science-fiction literature in relation to the educative problem seems a worthy and useful proxy. Indeed, if we were looking to make a comparison, this approach could be seen as similar to using philosophy to add new dimensions to understanding literature. Such distinctions are not necessarily easily distinguishable, and are certainly not opposing entities (Latini, 2019). Besides which, the earliest incarnation of the idea of machine ethics came itself from the pen of Isaac Asimov, in the 1942 short story Runaround (Asimov, 1995), and finds a place in academia as a framework to be considered, moulded and accepted/rejected (as in the case of Anderson, 2005), rather than a literary trifle with no merit.

Furthermore, Asimov himself had a significant academic reputation as a biochemist before he made the decision to become an author full time (Touponce, 1991). It would not be unfair to suggest that his brand of science fiction was rooted as much in his empirical scientific experience as much as it was in his imagination. Indeed, he viewed it as the first duty of a scientist not engaged in research to make science and its principles accessible to the general population – writing science fiction was a way for him to continue to be a working scientist while leaving behind the conventions of academia (Touponce, 1991). Thus, his invention of the words “robotics” and “positronic” could seamlessly make the transition from fiction to regular use in academia as they were intended as scientific terms.

In spite of his background, his success and his passion, Asimov remained surprised to find that he had inspired others to build robots in the fashion that he had so lovingly created them in his stories. As he notes himself, ‘When I wrote my robot stories I had no thought that robots would come into existence in my lifetime’ (Asimov, 1995, p.10) so when he came across these “robots-in-reality” (as he referred to them), he was astounded that they resembled the industrial machinery, built with purpose, of his imagination – while not as intelligent or humanoid as his creations.

Asimov’s writings are perhaps not peer-reviewed, in as much as he had no peers at the time of his writing, and do not appear in any academic journal. However, they continue to be inspirational in a field which bears a title of his invention: Robotics. Therefore, I suggest that it is acceptable for us to look to Asimov’s Three Laws to inspire an imaginative speculation on how an autonomous robot might behave in given situations. In his own words: ‘people who work in the field of artificial intelligence sometimes take occasion to tell me that they think

the Three Laws will serve as a good guide' (1995, p. 10). With this in mind, let us articulate the Three Laws before continuing on to give definitions of the main concepts used in this paper:

Law 1: A robot may not harm a human being, or through inaction, allow a human being to come to harm.

Law 2: A robot must obey the orders given to it by human beings except where such orders would conflict with the first law.

Law 3: A robot must protect its own existence, as long as such protection does not conflict with the first two laws.

3. Definitions

Before we can discuss the behaviour of the hypothetical robot teacher in any detail, it is necessary to define the main concepts mentioned in the laws: human being, teacher, robot, and harm. Such definitions ensure the risk of misinterpretation remains low; besides which, a robot itself would necessarily have to hold sharply defined concepts of, basically, everything in order to identify it and interact with it. If we say, for example, that a robot must not harm a human being, we need to be sure, as a robot would, of what a robot is, and what a human being is to identify those instances where this may occur/has occurred, so that potential infractions of the three laws can be avoided/rectified. So, let us begin with the concept of the human being.

4. Human Being

There are numerous interpretations of the concept of human being, and the task here will be to decide upon and justify the definition to be made. Taking Evans' (2016) argument that we, both as members of the general public and as academics, refine our concept of human using particular [academic] anthropologies – choosing those which closely align with our worldviews – there are four broad perspectives we could take here. We could view humans as biological entities, and that our status as human is nothing more than the result of genetics. There is the theological anthropology – which Evans suggests is specifically Christian – which recognises the biological view but extends it to include the notion of soul, and the idea that humans were created in the image of God and, for that reason, humans can take primacy over other animals, but are equal in respect to each other. The philosophical anthropology identifies humans as those entities which hold important traits; for example, consciousness, higher intelligence, the ability to rationalise and communicate, and a sense of past and future. Finally, a socially conferred anthropology – a definition as given by the human respondents, and members of the general public, of an investigation carried out by Evans - which sees a human as a communicative being which learns human norms and values.

Besides the results of Evans' own investigations, each of his theoretical anthropologies is presented as a summation of centuries old discussions, none of which offer a definition that could not be argued against, some of which could be identifiable with animals as well as humans. We can thus identify problems with each of these anthropologies: the biological view confers fundamental importance to genetics; what then of people with genetic mutations? Should we think of them as not human? The theological view tells us that a human is only that being who was created in the image of God, yet the image of the human race is so diverse that one could easily find a person who did not fit this slim criterion, besides which we have no definitive image of God with which to compare ourselves. In philosophy, neither a baby without an ability to rationalise would be considered human, nor would an adult suffering mental ill health, and the socially conferred anthropology sets the criteria in an even narrower sense, with no real explanation of what those mandatory human norms and values, to which we should align ourselves, might be.

Morriss-Kay (2010) describes one of the defining characteristics of the human species as the ability to create (visual) art. This is again problematic given, as Morriss-Kay herself admits, there is little agreement on how to define art. Without a definition of art, we are left with the possibility of defining any creation as

"art", and any creator of "art" as human. The hypothesis does not hold if, for example, we tie a paintbrush to a dog's tail and aim it toward a canvas. Neither would it work conversely: it would be possible for many people to point to creations by humans we would never consider to be "art". This apparently defining characteristic is further indicative of the difficulty we have delineating what exactly is human.

Such muddy definitions, however, do not support our hypothetical robot to decide what is a human and what is not. It would seem that the most efficient method for our machine, and probably for us, is to categorise humans as those creatures which share the common traits of the genus "homo": namely our large brains; the ability to walk upright; unrivalled fine motor skills and the intelligence to apply them; and our ability to learn and form social structures (Harari, 2011). It may be prudent to add consciousness, and/or empathy, to this recipe given that these are oft cited as markers of humanity. Yet with these, as with any of the aforementioned traits, we are likely to be able to point to a being we would describe as human who was missing at least one of them. Since a sharper definition of the concept of human lies outside the scope of this paper, a view must be taken that is defined enough to suit the needs of the hypothetical robot, who will be required to identify the human on sight, and the needs of the reader, while acknowledging that it cannot be all encompassing. I suggest that such a view could be taken as follows (although I also understand that it remains highly problematic in itself): a human being is that which bears the physically recognisable traits of the genus "homo".

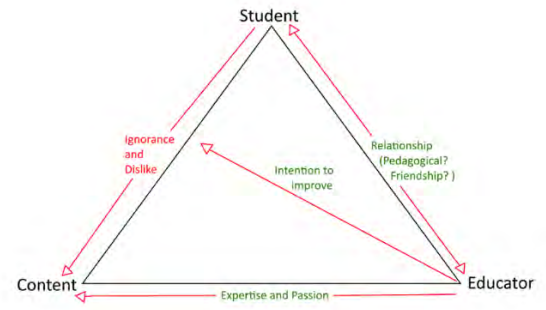
5. Teacher

Much like the notion of human being, the notion of teacher remains an elusive concept to describe given that it would require a thorough investigation of each individual's notion of "teacher". Here, I will use a version of the pedagogical triangle (Figure 1) which features strongly in the German didaktik tradition and is inspired by incarnations given in various works by Friesen & Osguthorpe, 2018 and Kenklies, 2019. As a scholar, my research is closely aligned with the discussions of the pedagogical relation as it presented in works of continental pedagogy such as those by Klaus Mollenhaeur (2013), Max van Manen (1991; 2012) and Norm Friesen (2017). The model of relations as offered by the pedagogical triangle offers a neat parallel to examine the implications of the three laws on relations between robot (teacher) and human (student)

I am, however, aware through observation, discussion, and tutoring prospective teachers that this is but one perspective on pedagogy. As with any and all speculations offered in this paper, the reader is encouraged to consider along the same lines as I have using a definition of pedagogy which is natural to them.

The version of the triangle I have used is more elaborate than those featured in the texts from which I was inspired. The annotations I have included with regards to the qualities that could be found in each of the relations have arisen from inferences made from engagement with discourse and observations in practice; they are by no means fixed. I am confident that they are in keeping with the tradition of the relational model in continental pedagogy from which the concept of the pedagogical triangle was born.

Figure 1



The Pedagogical Triangle

Here, we can see the relation of the educator/teacher to their student and to the subject matter, allowing us to, thus, extrapolate the essential characteristics of the teacher. Firstly, there must exist an intention to improve a student's relation to content, which carries with it the intention to establish a relation with the student – this pedagogical relation is thus the intention of the

educator directed towards the student (Nohl, 1933, as cited in Friesen, 2017).

The intention to improve a student's relation to content, assumes that the educator must also have some relation to the content itself. In the triangle, this relation is characterised as expertise and passion for what is being taught. Friesen and Osguthorpe (2018) tell us that the teacher aims to effect a change in the student's relation to the content, from (perhaps) confusion to clarity, by preparing and teaching it. However, a student's relation to content can have different qualities. For example, in my version of the triangle I have suggested that the student may be ignorant of, or dislike, the content, based on my observations in practice that this can sometimes be the case even before confusion sets in. In order to change the student's relation to content effectively, we can infer that the teacher must have passion to drive the intention to teach the subject initially, and expertise in order to identify and communicate the important elements of it. Furthermore, they should have sufficient ability to apply these in their practice. I suggest that without passion and expertise, the relation between educator and content becomes the same as that between student and content, thereby compromising the structure of the triangle, and of the pedagogical relation itself.

Of course, ideas of passion and relation have notable emotional, and human, connotations; intention arguably less so given that it is the root of all action – a machine, in Asimov's world, is built with the intention of fulfilling a purpose. That the intention lies initially with the engineer is of no consequence; the machine becomes the embodiment of such an intention. Since our robot cannot be characterised as a human, we must craft a definition of teacher which accounts for the essentials of passion, relation, and intention in non-humans. I have, therefore, paraphrased from, and remodelled, a definition of non-human teacher as provided by Caro and Hauser (1992) in order to create a brief definition here: Actor A can be said to teach actor B where its behaviour (whether organic - in which Actor A's actions are not solely contingent on input from a third party - or programmed – where Actor A cannot act without such input) denotes an intention to facilitate actor B's gaining of knowledge and experience, or sets an example for B which results in a changed (or improved) relation between actor B and the subject matter.

6. Robot

To define the concept of robot, it is logical that we return to Asimov himself to inform our definition given that we will be discussing the application of his laws based on the machines of his imagination. He makes a clear definition of robot as an industrial product built by engineers for a specific purpose, and he writes of autonomous robots as machines capable of judgement with no human input after the hardwiring of the three laws (Asimov, 1995). This is somewhat removed from the original intended meaning of the word "robot" – an invention by Karel Capek (printed 1920, edition used 2011) for his play R.U.R in which a robot was a manufactured human, indistinguishable from the real thing. Since we have already provided a definition of human being for the purposes of this investigation, and we have identified the need for as sharp a definition as is practicably possible, it would be senseless to refer to Capek's idea of robot simply because it renders our less than perfect distinction between robot and human with even more blurred lines. Therefore, we can take Asimov's definition of robot as reasonably appropriate for this paper.

However, before we can do so, we must address the issue of autonomy. Asimov tells us that our hypothetical robot is a machine capable of making judgements without human input, and this is supported by definitions of autonomy such as that offered by Bartneck and Forlizzi (2004), in which it is succinctly given that autonomy is "having the technological capabilities to act on behalf of humans without direct input from humans." (p. 593). It is prudent, nonetheless, to examine this against other conceptualisations of autonomy. Autonomy as described by Kupfer is, most basically, "a concept of oneself as a purposeful, self-determining, responsible agent" (1987, p. 82). In order to get there, he argues, privacy is essential as this conveys to the autonomous being that they alone are responsible for determining how much of themselves to reveal to others and it

gives opportunities for self-scrutiny and evaluation. Using Kupfer's definition, we may then say that it is impossible for a machine ever to reach this level of self-concept given that it is not afforded any privacy. The entire inner workings of the machine – including the code used to create its "brain" - are known to the engineer(s) from its inception. Yet, in Asimov's definition, after the robot's build and initial coding, there is no further human input[1]. Could this offer the privacy required for the machine to reach a state of autonomy in the way that Kupfer describes? Even if not, how much would that be required given that, in this world of hypotheticals, we could perhaps code such a concept of oneself as an autonomous agent, thereby creating a short cut to autonomy?

In any case, can we even justify the use of Kupfer's idea of autonomy when it is explicitly referring to humans? We once again risk muddying the waters of our distinctions. Hexmoor, Castelfranchi and Falcone (2003) refer to Karl Popper's response to Alan Turing's challenge that a computer could do whatever a human could, in which he asserted that a computer (or artificial intelligence, or indeed robot) did not hold the initiative, the ability to reason pro-actively, that is necessary for autonomy. This is, according to the authors, not entirely accurate as the agents and robotic systems being built at the time of their writing were showing "nontrivial initiative" (p. 2) which derives from the sharing of initiative with human beings and this, alongside the ability to interact with humans and other machines, is a core component of autonomy as they describe it. Autonomy, when applied to machines in use currently, is derived from a relationship between the robotic agent and the human agent (Castelfranchi and Falcone, 2003). We are, therefore, not yet at the stage where complete autonomy of a machine is being speculated upon; indeed, David et. al (2016) as part of the Defense Science Board, make the assertion that no machine, and, furthermore, no person, is truly autonomous in the strictest sense of the word. It is important to note that the strict sense of the word to which they allude is curiously absent from their assertion.

With this in mind, a definition of autonomous robot, if not fully autonomous, is required to allow us to continue on our hypothetical journey. From Asimov, it has been suggested that robots should be purposeful machines built by engineers capable of non-human led judgement. What we have learned from later academics is that there is a necessary relationship between the human agent and robotic agent – even if this is just at the beginning during the machine's build – and that robot autonomy is characterised by an ability to interact with humans and other machines. The ethical dimension of the autonomous robot is described by Anderson (2005) when she notes that an autonomous machine is one preloaded with ideal ethical principles, or some examples of ethical dilemmas with "correct" answers (which, in my view, wrongly assumes a universally accepted correctness), and a learning procedure from which these so-called ideal ethical principles can be abstracted in order to be used to guide the robot's own actions.

From all of this, we can make a brief definition of autonomous robot as follows: An autonomous robot is a machine built by engineers for a specific purpose, which has the ability to interact with humans and other machines, and in doing so is guided by ethical principles and judgements; some coded and some (potentially) learned.

7. Harm

The final concept that Asimov refers to in his Laws is harm. This is a concept to which one may be tempted to apply a common sense approach; however, given that it is applied in the Three Laws to both humans and robots (although it is done in an implicit way when it is said that a robot must protect its own existence) we need a definition which suits both, and common sense may implore us to view it from a purely human perspective. Thus, here follows a short consideration.

[1] An interesting analogy could be drawn with education here: how much could we compare the building of an autonomous machine with the Bildung – formation - of an autonomous human being?

An initial look to the legal definition – which is said to be ill treatment; the impairment of physical or mental health; or the impairment of physical, intellectual, emotional, social or behavioural development [of a young person] (Thomson Reuters, n.d.) – shows that it could be sufficient to both human and machine, depending on what is meant by ill treatment. It would also be necessary for legal rights to be inferred onto the machine before the legal definition could be said to have any weight. Until the machine is viewed as an entity in its own right, it would be considered a piece of property with legal rights to it being held by its owner. Therefore, any ill treatment given out to the machine would result in damages paid out to the owner with no judicial restitution for the machine alone. This is an insufficient definition for our inquiry, where our autonomous machine may not be thought to necessarily have an owner.

Mill offers a philosophical position on harm, in which he posits that a harmful action must violate, or risk violation of, the important interests of others in which they have a right (Brink, 2018), which is once more sadly insufficient. Again, it assumes rights which may not necessarily be placed onto the machine. While both the legal and philosophical definitions set out the criteria for how a human could come to harm (by either a robot, a human or the inaction of either), it is less suitable for recognising harm in reference to the machine. For this reason, I find that, outside of conducting a larger scale conceptual inquiry, we might appropriate a semantic definition here, rather than a systematic one, as it suits our needs in reference to both human and robot. I paraphrase a definition from the Collins English Dictionary (n.d.): Harm is the damage caused to something which is the result of a particular course of action; to harm a thing, or person, means to damage them or make them less effective or successful than they were.

Having set out the definitions of the four main concepts involved, these will now be applied to three hypothetical scenarios, which play out the laws, to illustrate the potential behaviour of the autonomous robot when coded with Asimov's Three Laws.

8. Law 1: A robot may not harm a human being, or through inaction, allow a human being to come to harm.

In this first scenario, the robot teacher has a set a task for its students. One of the students is feeling particularly anxious about this task and is displaying a number of the common physical symptoms of anxiety: sweaty palm, shallow breathing, fast heart rate and gastric distress. These symptoms may, however, also be applicable to a number of other physical illnesses which would require medical attention. Here, we assume that the robot – like a human teacher – is not necessarily equipped with the ability to diagnose medical ailments and so, based on a hard coding of this first law, must act to avoid this student coming to harm from the symptoms she is experiencing.

It seems likely that the first course of action would be for the robot to remove the student from the situation, thereby lessening their anxiety, and symptoms thereof, but not allowing the student to complete the given task. The robot must take action to avoid any harm coming to the human, and whether it can recognise this as anxiety or not, for the human to remain in the situation with anxiety running high is likely to make them less effective, as in the definition of harm given above. This displays a limitation that the robot has over the human teacher: the ability to weigh risk and sacrifice, as well as a lack of future mindedness.

Education as a means of achieving, or at least aiming for, a changed future (better, perhaps) relies on the ability to be future-minded (Kenklies, 2020). The educator, as we have seen on the pedagogical triangle, begins with an intention to improve a student's relation to content, and this is an intention rooted in the desire to change the future. In order to do this, it is sometimes necessary to endure an unpleasant present: think of a woman suffering the apocalyptic pain of childbirth to feel the love of her child in the years to follow, as an example. In the scenario concerning our student, the anxiety that they feel for the given task offers present discomfort but working through and completing the task offers potential future gain. It is a gamble that the human

teacher makes in predicting that the long term benefits to the student outweigh the risks taken in the short term.

This, however, is not the only risk inherent in the pedagogical relation. Biesta describes, as he sets out the "Beautiful Risk of Education" (2013), that the teacher, in teaching, offers up a gift to the student and the student opens themselves up to receiving this gift as something radically new. The gift given by the teacher, I suggest, is encased in a fragile wrapping of self-consciousness – as is so often the case when one exposes and offers up something that they care, or are passionate, about. The teacher, thus, is opening up a part of him/herself to the student ready to give, and the student is opening up a part of him/herself ready to receive – there is a mutual vulnerability here.

In this scenario, the human teacher would benefit from the flexible judgment of knowing that the student may need some support, reassurance and empathy to see their way through the task, where a robot teacher could not take the risk that the student may come to potential harm.

9. A robot must obey the orders given to it by human beings except where such orders would conflict with the first law.

In this second scenario, the robot teacher asks a loud and disruptive student to quieten down so that the rest of the class can work peacefully. The student, in all of his youthful belligerence, declines and tells the teacher to go away. The robot, seeing no conflict with the first law, in that its going away would not cause harm to the students in the class, dutifully retreats to the base destination as coded into its circuitry, and can no longer support or influence any of the students for whom it was responsible.

This scenario presents us with the question of authority, and this runs beyond the idea of teacher as disciplinarian, that one-dimensional concept of authority so offered by education policy and professional literature (MacLeod, MacAllister and Pirrie, 2012). As a multi-dimensional concept, MacLeod et al offer different notions of authority but note that it is personal authority, deriving from the personal qualities of the teacher, which students seem to recognise and respond to and not so much their expertise or any perceived power. Whether expertise is a necessary presence when communicating authority or not, we know from the framework offered by the pedagogical triangle that it is a necessary presence in any teacher, regardless of the status of their authority. I suggest that authority and expertise run in tandem: without authority there cannot exist any trust that the teacher has expertise in the subject matter above that of the student, something which we have already inferred from the model of the pedagogical triangle should be present for improving a student's relation to content. There are, of course, those who may disagree with this claim. Rancière's (1991) examination of the methods of Joseph Jacotot's Universal Teaching advocates for the ignorant schoolmaster who aims to teach students by teaching them nothing and claims no expertise (or intelligence) above that of the students. I argue, however, that any proponent of this method, to execute it effectively, must at least have expertise in its tenets.

Authority, of course, is not inferred onto any teacher automatically. It is thought that in human teachers it relies somewhat on personal disposition, somewhat on the nature of their professional education (MacLeod, MacAllister and Pirrie, 2012), which, by my interpretation, might arguably be considered expertise by another name. A machine could be coded to display a particular disposition, and be programmed to hold enough expertise, thereby potentially gaining authority via an alternate route. What the second law precludes, however, is the ability of the robot to exercise any authority they may gain as the students start to become aware that the power that they hold over the machine renders its authority moot.

10. Law 3: A robot must protect its own existence, as long as such protection does not conflict with the first two laws.

In this final scenario, inspired by a scene in Asimov's short story Bicentennial Man (Asimov, 1995), the robot is ordered by the same belligerent student from scenario two to dismantle itself. Given that the second law takes precedence over the third,

the robot must obey the human command, or it would be in contravention of law two. A further dilemma ensues if the robot calculates that, by dismantling itself, it cannot act to prevent harm to the students in the class. The hierarchical nature of the laws means that now the first law should always take precedence, but an infinite loop is easily created if the student repeats his command ad nauseum, and each of the laws finds itself in conflict with the other two. An infinite loop in any computer program results in a crash which would render the internal programming, and thus the machine itself, dysfunctional.

This scenario exposes the shortcomings of the laws in general as rigid, dogmatic principles governing behaviour. What hope for an entity in an ever changing world when they hold a limited propensity to change, if they hold such a propensity at all? I would be remiss to negate the idea that there are some humans – and by extension human teachers – who also live by what can be perceived as a rigid set of rules with a reluctance to change. I can understand how parallels could be drawn here, but what is inferred by education, and the pedagogical relation (as detailed in fig. 1), is that there is a capacity for change: the educator identifies this potentiality in their intention to change the student's relation to content. However, this capacity to change in the student is a reflection of the educator's potential for change: the teacher brings to life the student's capacity to change by offering themselves as an example, and that which is brought to life in the student is thus reflected back at the teacher (Fromm, 1956). A robot without such a capacity for change would be unable to project it, and recognise it, in their student.

Of course, it could be argued that a robot can change if it is programmed to do so or has been programmed to learn to do so. We could also even suggest that there is only a slight difference between the modification of internal circuitry and what happens when a person is educated. While such a discussion would be a worthwhile one to conduct, it is, again, not one which can take place here.

11. Conclusion

Using Asimov's Three Laws as a foundation for predicting the behaviour of an autonomous robot in three hypothetical scenarios, I suggest that robots coded with these laws could not be considered teachers in any way comparable to our understanding of teacher as offered by the framework of the pedagogical triangle. It is the inability to exercise risk; a lack of future-mindedness; a hardwired lack of authority; and the rigid adherence to an inflexible set of laws with no capacity for change which work to the robot teacher's detriment. This is, of course, my speculative view based on my understanding of the notion of teacher as needing all of the latter to be able to teach effectively.

As I have mentioned previously, the notion of teacher varies according to the individual, as well as their definition of education, and as such there are others whom I expect will take a different view from my own. For example, there are some who hold the view that the only role of the teacher is to impart knowledge, and this viewpoint is evident in the kinds of technology currently in use, and being created, for this very purpose – I think of the example of Massive Open Online Courses (MOOCs) in which the relation between teacher and student is mediated by not only a screen, but also the physical and metaphorical distance between potentially thousands of students to one educator.

Of course, we can say that a robot need not be coded with the Three Laws; and indeed, as we find the development of robotics gathering apace, many academics in that field are working towards a code of meta-ethics which will offer a little more flexibility than that of Asimov's (Anderson, 2005; McCauley, 2007). Asimov acknowledged himself that what he had set out in his work of fiction was probably not infallible (Anderson, 2005). The question then would become, what ethical system, if any, would be sufficient for a robot to be coded with in order for it to teach effectively?

Furthermore, if we ever did develop a machine which could teach, can we say that there is a machine that could be a teacher? Is there such a distinction between doing and being? If so, when does the newly

trained teacher – instructed with abundance in the doing – make that qualitative leap into being a teacher? These are questions worth asking if we do not wish to risk our universities and colleges becoming "teacher factories" concerned only with the manufacture of people for a specific educational purpose. Perhaps, if we were to take this viewpoint, there would not be so much difference between human and robot teachers after all.

12. Disclosure statement

The author(s) declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

13. Open Access Policy

14. References

Anderson, S.L. (2005, November 4-6). Asimov's "Three Laws of Robotics" and Machine Metaethics [Symposium paper]. *AAAI Fall Symposia, Arlington, VA, United States*. https://www.aaai.org/Papers/Symposia/Fall/2005/FS-05-06/FS05-06-002.pdf

Asimov, I. (1995). *The Complete Robot*. London: Voyager.

Bartneck, C. & Forlizzi, J. (2004). A Design-Centred Framework for Social Human-Robot Interaction. *Proceedings of the Ro-Man2004: 13th IEEE International Workshop on Robot and Human Interactive Communication*, Kurashiki, Japan, 591-594. https://ieeexplore.ieee.org/document/1374827

Brink, D. (2018). Mill's Moral and Political Philosophy. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/mill-moral-political/#HarPri

Capek, K. (2011). R.U.R.: *Rossum's Universal Robots*. [Kindle version]. https://www.amazon.co.uk/R-U-R-Rossums-Universal-Karel-Capek/dp/1557422559/ref=sr_1_2?dchild=1&keywords=rur&qid=1603457927&sr=8-2

Castelfranchi, C. & Falcone, R. (2003). From Automaticity to Autonomy: The Frontier of Artificial Agents. In H. Hexmoor, C. Castelfranchi & R. Falcone (Eds.) *Agent Autonomy* (pp. 103-136). New York: Springer.

Collins English Dictionary (n.d.). Harm. In *Collins dictionary.com dictionary*. Retrieved October 23, 2020 from https://www.collinsdictionary.com/dictionary/english/harm

David, R., Nielsen, P., Allard, J., Alving, A., Anastasio, M., Appleby, B., Austin, W., Bayer, M., Bradshaw, J., Cappuccio, F., Carns, M., Chakraborty, A., Chevillet, M., Chu, D., Coleman, V., Day, C., Evans, E., Fallon, W., Fields, C. & Zacharias, G. (2016). *Final Report of the Defense Science Board Summer Study on Autonomy. Publicly-Releasable Version.*

https://www.researchgate.net/publication/306286423_Final_Report_of_the_Defense_Science_Board_Summer_Study_on_Autonomy_Publicly-Releasable_Version

Evans, J.H. (2016). *What is a Human?: What the Answers Mean for Human Rights.* New York: Oxford University Press.

Friesen, N. (2017). The pedagogical relation past and present: experience, subjectivity and failure. *Journal of Curriculum Studies, 48*(6), 743–756. https://doi.org/10.1080/00220272.2017.1320427

Friesen, N. & Osguthorpe, R. (2018). Tact and the pedagogical triangle: The authenticity of teachers in relation. *Teaching and Teacher Education, 70*, 255-264. http://dx.doi.org/10.1016/j.tate.2017.11.023

Fromm, E. (1956). *The Art of Loving*. New York: Harper & Row.

Harari, Y.N. (2011). *Sapiens: A Brief History of Humankind.* London: Penguin.

Hashimoto, T., Verner, I.M. & Kobayashi, H. (2012). Human-Like Robot as Teacher's Representative in a Science Lesson: An Elementary School Experiment. In J. Kim., E.T. Matson, H. Myung & P. Xu (Eds.) *Robot Intelligence Technology and Applications 2012: An Edition of the Presented Papers from the 1st International Conference on Robot Intelligence Technology and Applications* (pp. 775 – 786). Berlin: Springer. https://doi.org/10.1007/978-3-642-37374-9_74

Hexmoor, H., Castelfranchi, C. & Falcone, R. (2003). A Prospectus on Agent Autonomy. In H. Hexmoor, C. Castelfranchi & R. Falcone (Eds.) *Agent Autonomy* (pp. 1-10). New York: Springer.

Kenklies, K. (2019). The Struggle to Love: Pedagogical Eros and the Gift of Transformation. *Journal of Philosophy of Education, 53*(3), 547-559. https://doi.org/10.1111/1467-9752.12376

Kenklies, K. (2020). Dogen's Time and the Flow of Otiosity – Exiting the Educational Rat Race. *Journal of Philosophy of Education, 54*(3), 617-630. https://doi.org/10.1111/1467-9752.12410

Kupfer, J. (1987). Privacy, Autonomy and Self-Concept. *American Philosophical Quarterly, 24*(1), 81-89. https://www.jstor.org/stable/20014176

MacLeod, G., MacAllister, J. and Pirrie, A. (2012). Towards a broader understanding of authority in student-teacher relationships. *Oxford Review of Education, 38*(4), 493-508. https://doi.org/10.1080/03054985.2012.716006

McCauley, L. (2007). AI Armageddon and the Three Laws of Robotics. *Ethics and Information Technology, 9*, 153-164. https://doi.org/10.1007/s10676-007-9138-2

Mollenhauer, K. (2013). *Forgotten Connections: On Culture and Upbringing*. (N. Friesen, Trans.). New York: Routledge.

Morriss-Kay, G.M. (2010). The evolution of human artistic creativity. *Journal of Anatomy, 216*(2), 158-176.

Latini, M. (2019). Second Variation. Philosophy and Literature: A Hypothetical Comparison Between Different Approaches. *Rivista di estetica, 70*, 11-18. https://doi.org/10.4000/estetica.5020

Office for National Statistics (ONS). (2019). *Which occupations are at highest risk of being automated?* https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/articles/whichoccupationsareathighestriskofbeingautomated/2019-03-25

Rancière, J. (1991*). The Ignorant Schoolmaster: Five Lessons in Intellectual Emancipation.* (K. Ross, Trans.). Stanford, CA: Stanford University Press.

Rebora, A. (2011). Your Future Colleagues? *Teacher Professional Development Sourcebook, 4*(2), 12.

Rotman, D. (2020, February 24). We're not prepared for the end of Moore's Law. *MIT Technology Review*. https://www.technologyreview.com/2020/02/24/905789/were-not-prepared-for-the-end-of-moores-law/

Thomson Reuters Practical Law. (n.d.). Significant Harm. In *Thomson Reuters Practical Law Glossary*. Retrieved October 23, 2020, from https://uk.practicallaw.thomsonreuters.com/8-538-0246?transitionType=Default&contextData=(sc.Default)&firstPage=true

Touponce, W. (1991). *Isaac Asimov*. Twayne Publishers: Boston.

van Manen, M. (1991). *The Tact of Teaching: The Meaning of Pedagogical Thoughtfulness*. New York: SUNY Press.

van Manen, M. (2012). The Call of Pedagogy as the Call of Contact. *Phenomenology & Practice, 6*(12), 8-34.