Chapter 10
Ethical Machine Safety Test

Roman M. Krzanowski and Kamil Trombik

Abstract Within a few decades, autonomous robotic devices, computing machines, autonomous cars, drones and alike will be among us in numbers, forms and roles unimaginable only 20 or 30 years ago. How can we be sure that those machines will not under any circumstances harm us? We need a verification criterion: a test that would verify the autonomous machine's aptitude to make "good" rather than "bad" decisions. This chapter discusses what such a test would consist of. We will call this test the ethical machine safety test or machine safety test (MST) for short. Making "good" or "bad" choices is associated with ethics. By analogy, an ability of the autonomous machines to make such choices is often interpreted as machine's ethical ability, which is not strictly correct. The MST is not intended to prove that machines have reached the level of moral standing people have or reached the level of autonomy that endows them with "moral personality" and makes them responsible for what they do. The MST is intended to verify that autonomous machines are safe to be around us.

1 Introduction

Within the next few decades, autonomous machines will enter our lives not as passive devices but as autonomous agents in unprecedented numbers and roles (Ford 2015; Berg 2016; Bloem et al. 2014; Boyle 2016; Brown 2016; Clifford 2017; Cookson 2016; Krzanowski et al. 2016; Pew Research Center 2014; Schwab 2016, 2017; Sullins 2011). The view that this technology will be all gain and no pain, supported by some,[1] is hardly justified by the historical record and current

[1] "Ethical machines would pose no threat to humanity. On the contrary, they would help us considerably, not just by working for us, but also by showing us how we need to behave if we are to survive as a species" (Anderson and Anderson 2010; Bostrom 2015; Anderson 2016). See also Schwab (2016, 2017).

R. M. Krzanowski · K. Trombik (✉)
The Pontifical University of John Paul II, Cracow, Poland

W. Hofkirchner, H.-J. Kreowski (eds.), *Transhumanism: The Proper Guide to a Posthuman Condition or a Dangerous Idea?*, Cognitive Technologies,
https://doi.org/10.1007/978-3-030-56546-6_10

experience (see Heron and Belfort 2015). A more cautious approach, such as that suggested by Ford (2015), Gray (2007), Yampolskiy (2012a, b) and Kaczynski (1995), is preferable.

The question, then, is the following: How can we be sure that autonomous machines will not harm us and will behave as "ethical" agents?[2] To address this problem we need a test that would verify the autonomous machine's aptitude to act in a safe way, i.e. in some sense to act ethically. We will call such a test the machine safety test (MST).[3]

2 Key Terms

To avoid any misinterpretations, let us define certain key terms used in this chapter. These are ethics, ethical agent, machine ethics, autonomous machines and autonomous ethical agents.

Ethics, in the most general terms, is a set of prescriptions or rules about how to live a good and rewarding life as an individual and as a member of society (Bourke 2008; Vardy and Grosch 1999; MacIntyre 1998). Such a concept of ethics may be reduced, as is often the case, to a set of rules with a yes/no answer, specifying what to do (Beavers 2011). But ethics is more than just rules. It (implicitly or explicitly) requires free will, consciousness, a concept of good and bad, virtue and values, a concept of self, an understanding of responsibility (of "ought" and "ought not") (MacIntyre 1998; Veach 1973; Sandel 2010) and a good comprehension of the reality around us. A lot of deep metaphysics (free will, a concept of good, a concept of self, etc.) is involved in the concept of ethics. Dispensing with metaphysics leaves ethical statements groundless.

[2] We need to keep in mind that the future full of happiness and unalloyed human flourishing promised by light-minded AI and robotic enthusiasts is just an uncritical and hardly justified fairy tale fantasy. I propose to leave behind Start Trekfans, Asimov's Three Laws of Robotics and other Sci-Fi phantasms. History does not justify such a vision at all (unfortunately!). Recall the cautionary words about progress offered by more discerning minds: "What the Enlightenment thinkers never envisioned was that irrationality would continue to flourish alongside rapid development in science and technology... In fact, [there is] no consistent link between the adoption of modern science and technology on the one hand and the progress of reason in human affairs on the other ... There is nothing in the spread of new technologies that regularly leads to the adoption of what we like to think of as a modern, rational worldview" (Gray 2007, p. 18).

[3] "The development of machines with enough intelligence to assess the effects of their actions on sentient beings and act accordingly may ultimately be the most important task faced by the designers of artificially intelligent automata" (Allen et al. 2000). Seibt writes: "...we are currently in a situation of epistemic uncertainty where we still lack predictive knowledge about the individual and socio-cultural impact of the placement of social robots into human interactions space, and we are unclear on which aspects of human interactions with social robots lend themselves to predictive analysis" (Seibt 2012).

Can then such a deep ethics be computed (in the Church–Turing sense), given that metaphysics is not mathematical? Ethical rule-based on Hobbesian, Kantian, utilitarian or other ethical schools can be to some extent translated into a computer algorithm and made "computable". But then all "metaphysical" dimensions of the ethical actor are "lost in translation". If a machine is programmed according to "translated" rules, one may claim that it possesses ethical qualities or that it is an ethical machine (Anderson and Anderson 2010). But this ethics would be a special type of ethics, not ethics in the deep, metaphysical sense. Ethics in a deep sense (like metaphysics) is non-computable, and we do not have any other meaning of "computable" that could rescue "computerized ethics" from its shallows (Yampolskiy 2012a, b; Krzanowski et al. 2016)[4] (see Turner 2016 for a definition of computation). The ethical rules translated into a machine format constitute what we will call machine ethics as m-ethics.[5]

An ethical agent is an individual (artefact or natural) acting according to ethical rules. In the context of our discussion, we may call an autonomous machine an ethical agent understanding that we mean here ethics as m-ethics, or a set of behavioural rules directing the behaviour of an autonomous machine. Nothing else. Thus, it is misleading to talk about "moral machines", "ethical machines" or the likes. Too generous use of these terms will only confuse the problems we face with autonomous machines (by attributing to them capacities they cannot have); we need to constantly remind ourselves that the subject of our discussion is autonomous machines with implemented m-ethical software or system.

Autonomous machines (e.g. Floreano et al. 1998; Patrick et al. 2008; Ni and Leug 2016) are machines that act in the environment without direct command by or involvement with humans. Autonomous machines, which implement m-ethics rules and thus display ethical-like behaviours, are autonomous ethical machines.

3 Where We Are with MST

So far we have only a few proposals of such a test. These are:

[4] For someone that cannot accept a concept of deep ethics, the more technical explanation of what ethics really entails may be easier to comprehend: "… given the complexity of human values, specifying a single desirable value is insufficient to guarantee an outcome positive for humans. Outcomes in which a single value is highly optimized while other values are neglected tend to be disastrous for humanity, as for example one in which a happiness-maximizer turns humans into passive recipients of an electrical feed into pleasure centers of the brain. For a positive outcome, it is necessary to define a goal system that takes into account the entire ensemble of human values simultaneously" (Yampolskiy and Fox 2012).

[5] It is critical to understand this difference. If we attribute ethics to machines we may be tempted to bestow on them personality, responsibility and similar (which unfortunately is slowly happening). But if we say that these machines have m-ethics, which is what they have, we will make such flights of fancy much more difficult.

1. Moral Turing test (MTT)
2. Turing triage test (TTT)
3. Ethical competence test (ECT)

The moral Turing test (MTT) proposed by Allen et al. (2000) is similar to the "imitation game" proposed by Turing (1950). In the original Turing formulation of the "imitation game", machine "intelligence" is assessed by a panel of judges based on a series of responses to questions posed to a machine and a human subject. If, based on the answers given, the judges cannot distinguish a machine from a human, then the machine has passed the test, but what exactly it means is open to interpretations as pointed out for example by Oppy and Dove (2016). The MTT would be run in a similar way, but with the key difference that the questions would be of moral import. Allen, Vermer and Zinser recognized the multifarious nature of ethical problems that would beset such a test and thus recognized its potential limitations. The problem with the MTT, however, may lie elsewhere. The Turing test has not lived up to its promise or its author's intentions, and there is no consensus as to what the TT is actually testing or attempting to demonstrate (Turing 1950; Oppy and Dove 2016; Saygin et al. 2000). Thus, if the TT is not clear regarding its meaning,[6] on what grounds can it be extended to ethical problems with any expectation of success?

Another proposal to apply the TT to an ethical machine test is the Turing triage test (TTT) described by Sparrow (2004, 2014). In the test, the AI-based machine (robot, autonomous artificial agent), following an electrical shortage in a hospital, must choose between (1) turning itself off (equivalent to a human suicide in some respects) and saving a patient and (2) saving itself and allowing a patient to die. Sparrow claims that:

> My thesis, then, is that machines will have achieved the moral status of persons when this second choice has the same character as the first one. That is, when it is a moral dilemma of roughly the same difficulty. For the second decision to be a dilemma it must be that there are good grounds for making it either way. It must be the case therefore that it is sometimes legitimate to choose to preserve the existence of the machine over the life of the human being. These two scenarios, along with the question of whether the second has the same character as the first, make up the Turing Triage Test (Sparrow 2004).

Sparrow himself suggests that a machine could never pass the TTT; thus, a machine could never achieve the moral status of a person. Apart from the fact that the TTT has little to do with the original TT (only that a test is applied to a computing machine), it is within that class of abstract ethical problems that includes the notorious trolley problem, terror bomber, strategic bomber or similar (keeping in mind the differences). The "imaginary ethical problems"(also known as thought

[6] The objective of the Turing test (TT) was not to verify some specific kind of intelligence; it was aimed at a general intelligence. Thus, success in playing Chess or Go did not in fact prove or disprove a machine's capacity to reason, according to the TT's requirements.

experiments) have been devised not to provide a test of general ethical abilities[7] but to expose the multifarious nature of ethics. Thus, solving of the Turing triage problem will not translate itself into solving a problem of mariners stranded at sea (Sandel 2010); similarly, an ability to excel in strategic military games does not guarantee success in leading a real battle (as examples from history testify). Complex abstract ethical puzzles tell us little about the solutions to practical ethical problems (Dancy 1985; Szabó 2000; Elster 2011; Lehtonen 2012; Cathcart 2013).

A different proposal, not based on the Turing test, came from Moor (2006). Moor suggested the ethical competence test (ECP), which would assess the ethical aptitude of the computing agent using hypothetical situational scenarios (but not of the "imaginary ethical paradoxes" type). Such scenarios often do not have yes/no options, but instead use less or more favourable ones. The responses of an artificial agent to such scenarios would be compared with the choices made by humans in the comparable situations. Moor also introduced a requirement that the ethical robot provide a defensible and convincing justification for its decision. As he writes: "If a robot could give persuasive justifications for ethical decisions comparable to or better than those of good human ethical decision-makers, then the robot's competence would be inductively established for that area of ethical decision-making" (Moor 2009). As he points out, ethical tests should be situation-dependent as an automated agent may be competent in some situations and yet not in others. In certain situations, Moor points out, computer agents, because of their huge information processing powers, can make better (and faster) decisions than human agents regarding, for example, the allocation of scarce resources or scheduling the delivery of supplies in the event of catastrophic situations to avoid waste. But these decisions would qualify rather as better or optimal managerial decisions, rather than as ethical ones per se. Moor's proposal warrants attention as it acknowledges the complexity of ethical decisions, their dependence on situational context and the need to view an artificial agent not as a moral black box but, at the very least, as a grey one.

Summing up current efforts on testing of m-ethics, we can say that no conclusive proposal is on the table. We do not have a test that could serve as the MST of the general m-ethical capabilities of an artificial agent, nor do we have (implemented, proposed or conceptualized) a testing methodology to construct such a test. As well, we do not know what it would take to test an autonomous ethical agent for its ethical (m-ethical) probity.

[7] "Imaginary stories and thought experiments are often used in philosophy to clarify, exemplify, and provide evidence or counterevidence for abstract ideas and principles. Stories and thought experiments can illustrate abstract ideas and can test their credibility, or, at least, so it is claimed. As a by-product, stories and thought experiments bring literary, and even entertaining, elements into philosophy" (Lehtonen 2012).

4 Machine Safety Test: What It Should Be?

We take a safe approach to the MST (advocated by Yampolskiy 2012a, b; Moor 2009) and formulate three assumptions that underline the definition of the MST:

1. The inherent complexity of ethics renders it incomputable in the TM sense.
2. Computing machines cannot be ethical in the way people are.
3. Computing machines may play the ethical game. It just means that autonomous machines/agents may be made to act amicably towards us in all foreseeable situations and they should be safe to be around.

5 Claim: What We Are Testing

We need to ask: What are the objectives of the MST? Are we testing whether autonomous machines behave like humans or like ethical artificial agents with m-ethics capabilities? The first case is impossible to achieve considering our definition of m-ethics. The machine safety test is NOT the test verifying the *general "moral" or ethical aptitude of a machine*. It means that we do not test the machine's "equivalence" to a human agent, and it also means that we do not test a moral aptitude in specific circumstances.

The machine safety test is not designed to prove that machines have reached the same level of moral standing as people or have reached the level of autonomy that endows them with "moral personality" and makes them responsible for what they do.[8] The objective of the test is only to verify that:

1. The product we develop, i.e. an autonomous agent, a machine, is "safe" to be around people in general circumstances.
2. The propensity of the autonomous machine to do harm, by "intention", design or neglect, is limited to as narrow a margin as reasonably possible. However, it seems that the possibility of doing harm cannot be completely eliminated.

Of course, terms such as "general circumstances", "safe" and "a narrow margin" can be interpreted in many ways. Here, I use common understandings of these terms, accepting that it may require further elaboration. It is possible that soon we will have to create a new dictionary of ethical terms that would correctly represent machine ethics, as current ethical terms may not be sufficient to describe the complex human–machine interactions.

[8] We want to avoid dilemmas as reported by Heron and Belfort (2015): "The question of who we should blame when a robot kills a human has recently become somewhat more pressing. The recent death (we talk about 2015) of a Volkswagen employee at the hand of an industrial factory robot has left ethicists and legislators unsure of where the moral and ethical responsibility for the death should lie—does it lie with the owners, the developers, the factory managers, or elsewhere?"

6 What Should the Test Be?

It seems that the MST should not be theoretical or primarily theoretical (theoretical meaning involving only ethical reasoning verified in a dialogue or a conversation—free or structured). Theoretical questioning is not a good test of moral aptitude. As we know, behind the clever and reasonable answers there may be nothing of substance but software, nothing of substance in an ethical sense, as the experience with chat bots (and politicians) teaches us. And besides, what kinds of questions would we ask? Certainly, any type of standard personality test or tests that try to gauge a person's sanity should be ruled out as the responses to them can be easily programmed and reproduced by a computing machine without even the smallest ethical insight. Imaginary ethical cases as proposed in the TTT are not suited to this function, as explained earlier. Yet, there may still be some use for the verbal verification of the moral standing. Such a verbal test in the form of an interview (qualifying conversation?) may be used to understand the ethical reasoning of a machine (correctness of the software implementation of ethical capacities?), and it may be more conducive to the purpose of the MST than a test requiring specific answers. The requirement that an autonomous ethical agent be able to explain itself stipulates a requirement that it be implemented as a white box. The problem as to how we would gauge the results of such an interview remains an open question.

The bulk of the MST test should be an ability to make just decisions in specific life situations. Making just decisions in a real-life context, not an abstract ability to assign "right" or "wrong" labels to abstract situations, seems to be at the core of ethics.[9] Of course, to act as an ethical agent in life situations an agent will have to possess considerable abstract knowledge of ethics. But we would rather require that an autonomous machine makes ethical choices in concrete situations rather than be able to respond to complex ethical questions or solve imaginary cases.

We may compare the MST to the skipper patent test or an airline pilot test ceteris paribus. These tests include theoretical and practical components. Learning or testing for a pilot or a skipper begins with theoretical tests of basic technical knowledge, before progressing through training on the flight simulators and then flights with an instructor. Finally, these tests also include a period of apprenticeship. In the case of an airline pilot, the pilot-to-be must fly in a junior position for a certain number of hours, before he or she can be recognized as a pilot[10] with the licence to undertake solo flights, likewise with a skipper permit. In the case of the MST, the

[9] "'Ethics', as understood in modernity, focuses on the rightness and wrongness of actions. The focus is misleading in that actions never occur outside of the wider social and natural contexts to which they respond. Individual, community, and society clearly constitute such contexts, on the different levels of the natural . . . 'human' world. This world comprises our interpersonal relationships as well as the natural givens" (McCumber 2007, p. 161).

[10] See, for example, the requirements for the testing standards for an airline pilot: https://www.faa.gov/ training_testing/testing/ test_standards/media/faa-s-8081-20.pdf

tests are even more complex than for a specific job function, as they will be testing more general situations.

Thus, it seems that the proper test of the artificial moral agent should consist of a theoretical part, a series of practical problems of varying scope and difficulty, progressing from staged scenarios through to gradually less controlled situations and ending up with completely uncontrolled life situations, and (maybe) include a period of apprenticeship during which we verify an agent’s capacity to think and act morally in real-life situations (we would call such situations “open-ended”).

Thus, in summing up the discussion, the MST should include the following components:

1. Theoretical verification of ethical aptitudes and reasoning—possibly a qualifying interview rather than a Q&A session plus a white box option.
2. A situational test or series of tests, in which an artificial agent makes autonomous decisions in the fully life-like (controlled or not) environment. The tests may have a different scope and increasing complexity and include:
   - Staged tests
   - Controlled life situations
   - Open-ended situations
3. A period of apprenticeship in which an artificial agent acts in the real conditions under close supervision.

7 Use Case Framework for the MST

Due to the generality of the MST test, only the high-level framework of the UC could be provided. This would consist, in the proper sequence, of four stages:

1. Interview and discussion that would verify understanding of m-ethical rules and m-ethical reasoning using imaginary ethical cases. The tests should be performed under a white box paradigm; i.e. the tested system should be able to explain its decisions
2. Staged situational tests that would verify an ability of a tested system to respond to complex (arranged) situations. These tests may be similar to those used on human subjects such as Milgram experiment (Milgram 1963), Phone booth and Dime experiment (Doris 2002), Stanford Prison Experiment (2014) or Cornell experiment (Doris 2002).
3. Situational tests including controlled and open-ended life situations that would verify an ability of a tested system to respond to complex life situations. In this case, any real-life situation of substantial ethical import could be used, in particular situations prone to dilemmas and conflicts.
4. Apprenticeship, which would test an autonomous machine’s ability to act without supervision in real-life environment by participation in real situations.

The white box paradigm, as it was pointed out in the Stage 1, applies to all four stages of testing.

8 Operational Concerns

It is not obvious how the MST should be implemented. It obviously requires a machine capable of human-like functioning. A hardware-embedded software would be incapable of situational tests or apprenticeship without significantly compromising the MST framework. Thus, such devices by definition would not qualify as ethical agents and would not be subjected to the MST.

The learning period for an ethical agent would be long; yet because of the nature of computer technology, it may be that only selected exemplars of robots would be tested and the gains in ethical aptitude may be shared by appropriate updates within the compatible class of machines. Thus, there is no need to test every machine; only selected units should be tested, and the experience would be passed onto other agents.

The learning process for autonomous robots is not well defined. How the autonomous system would learn the proper responses to complex situations and how these responses would be integrated into their m-ethical data base is not clear. This should be another area of research.

It seems also that computing technology would allow the ethical experience to be "inherited" from generation to generation of ethical machines, provided that ethical norms stay unchanged. Thus, the ethical testing would not have to be done *ab ovo* with each new version of machines, something that we humans cannot avoid.

9 Review and Summary

It seems that the MST should include several testing venues, leaning mostly towards solving practical life situations. Such complex tests would verify the ability to make ethical decisions in the presence of conflicting cognitive stimuli, conflicting values and time pressure. Table 10.1 below shows possible components of such a test. The elements of the MST are arranged from the most elementary (Level I) and as such of lower importance to the most critical and complex (Level IV) in the rising degree of importance.

We should ask the question whether every autonomous agent should pass all of these test levels, or maybe, we could accept different levels of "robot ethics" and accept after Seibt (2017) "partial realizations" as applied to the MST, depending on the robot design?

Table 10.1 Proposed structure and components of the machine safety test

	Level	Test component	Objectives	Possible implementation
Theoretical component	I	Interview and discussion	Verify understanding of ethical rules and ethical reasoning	Imaginary ethical cases, a white box paradigm for ethical decisions
Practical component	II	Situational tests—staged	Test ability to respond to complex (arranged) situations	Milgram experiment (Milgram 1963) Phone booth and Dime experiment (Doris 2002) Stanford Prison Experiment (2014) Cornell experiment (Doris 2002)
	III	Situational tests—from controlled to open-ended life situations	Test ability to respond to complex life situations	Any real-life situation of substantial ethical import
	IV	Apprenticeship	Test ability to act without supervision in real-life environment	Participation in real situations—war relief effort, etc.

10 How Would We Evaluate Results?

How would we know that the machine passed the test? One option is to have a panel of judges to review the results of the test and develop test-passing criteria as in the Turing proposal. Should we also accept the Turing criterion of a 70% pass score? If we do, what would it mean to have a 70% ethical agent? Or, would we accept a 70% ethical machine to be among us? It is easier, it seems, to use a 70% pass score to judge that a machine functions reasonably (this was a Turing proposal), but not whether it has a 70% moral aptitude. It seems that any number, short of 100%, as the criterion of acceptance of the MET results would be, in this case, an arguable qualification. But how are we to judge situational tests?

We need to admit that we are not sure how the MST should be graded and what it would mean for the autonomous agent to pass/fail the test.

Perhaps instead of a numerical score, we ought to assign some qualitative “moral” standing to ethical machines. Moor (2009) proposed four classes of ethical robots, or as he calls them—ethical impact agents. These are unethical agents, implicit ethical agents, explicit ethical agents and full ethical agents. These are interesting classifications of hardware–software constructs. However, the four classes are too crude to address the ethical capacity of autonomous robots required by the MST. The best we can say is that these classes mark the points on the spectrum of ethical aptitude from inert objects to human agents, but the scale by nature admits fuzzyfied, not crisp, classes.

Some suggestions as to the gradation of ethical abilities may come from HRI research. Seibt (2017) proposes in the context of human–robot interactions "five notions of simulation or partial realization, formally defined in terms of relationships between process systems (approximating, displaying, mimicking, imitating, and replicating)". With the MST assumption that ethical machines "play an ethical imitation game" (not in the Turing sense of the game) and do not replicate human ethical abilities, such a classification may help us in understanding and classifying the MST results.

11 Parting Comments

We cannot exclude the possibility that the meaning of ethics or morality will evolve to the point that in the future ethical or moral principles attributed to humanity would be attributable to machines, robots, software or the like. Meanings of the words do evolve. Yet it is and will be important to make sure that now and in the future "machine ethics" means behavioural rules for machines, or machine safety specifications, not ethics in the human context. And the MST is supposed to test just this, not the presence of some kind of metaphysical moral fibre in hardware or software.

It seems that one of the barriers in the conceptualization of the MST is that ethical agents are perceived as "computers" or software bundles, in the same way as Turing conceptualized the Turing test subject (which is why we have TTT and MTT proposals). The ethical agents will be machines that act, move, interact with us in physical, not only mental, space. A small taste of this is offered by self-driving cars, which are essentially tested as the MST test is structured including software development, driving in a controlled environment,[11] driving with a supervisor and autonomous driving[12] (Stillgoe 2017a; Hern 2017; Balch 2017). These are essentially four stages of the MST. In the context of self-driving cars, these tests are called social learning (Stillgoe 2017b).

It is rather difficult to imagine that machines will have the same complex of values that people have and thus the same responsibilities towards us. Thus, the MST will verify not how close computing machines come to us, but rather how close they come to our expectations about safe, autonomous machines. One must consider that our expectations regarding autonomous machines will evolve. Another challenge

[11] "Michigan is also home to 'M City,' a 23-acre mini-city at the University of Michigan built for testing driverless car technology". Available at: http://fortune.com/2017/01/20/self-driving-test-sites/

[12] "The carmaker's autonomous vehicles traveled a total of 550 miles on California public roads in October and November 2016 and reported 182 'disengagements,' or episodes when a human driver needs to take control to avoid an accident or respond to technical problems, according to a filing with the California Department of Motor Vehicles. That's 0.33 disengagements per autonomous mile. Tesla reported that there were 'no emergencies, accidents or collisions.' Tesla's report for 2015 specified that it didn't have any disengagements to report" (Hall 2017).

will be posed by the fact that the machine technology has global reach, while m-ethics (as any ethics) is quite often local. Thus, training of "ethical" machines will have to keep pace with these changes; otherwise we may risk meeting on our streets autonomous agents with behavioural propensities of cavemen.

And above all, we need to constantly keep in the mind the fact that we are training or developing machines to act safely, to not to harm us—this is the essence of m-ethics. If, by some fit of imagination, we call it ethical training so be it, as long as we are aware of the difference—thus, no moral robots, no ethical robots, just safely operating autonomous machines.

If history teaches us anything, in this case it may indicate that the ethics of autonomous artificial agents may go the same way as Internet security or software in general: just as software companies do not take responsibility for damage caused by their faulty software, so they will shed the responsibility for the transgressions of their faulty ethical agents. Thus, willingly or not, we may have to learn how to live with Microsoft-Windows-quality ethical machines.[13] Because is there any reason why the future should be any different? It rarely is; it just presents itself in different technology.

Acknowledgments We would like to thank Prof. Pawel Polak for his constructive comments on the early draft of this paper. All the errors, faulty conclusions and logical and factual mistakes are of course of our doing.

References

Allen, C., Varner, G., Zinser, J.: Prolegomena to any future artificial moral agent. J. Exp. Theor. Artif. Intell. **12**, 251–261 (2000)

Anderson, S.: The promise of ethical machines. https://www.project-syndicate.org/commentary/ethics-for-advanced-robots-by-susan-leigh-anderson-2016-12 (2016)

Anderson, M., Anderson, S.L.: Robot be good. Sci. Am. **10**, 72–77 (2010)

Balch, O.: Driverless cars will make our roads safer, says Oxbotica co-founder. https://www.theguardian.com/sustainable-business/2017/apr/13/driverless-cars-will-make-our-roads-safer-says-oxbotica-co-founder (2017)

Beavers, A.F.: Is ethics computable. Presidential Address, Aarhus, Denmark, July 4. http://www.afbeavers.net/cv (2011)

Berg, A.: Revolution evolution. Finance Dev. (2016)

[13] A few quotations substantiate this claim: "Microsoft likes to have everything glued together like a kindergarten art project gone berserk, but this is ridiculous" (Vaughan-Nichols 2014); "*Microsoft Windows isn't the only operating system for personal computers, or even the best … it's just the best-distributed. Its inconsistent behavior and an interface that changes radically with every version are the main reasons people find computers difficult to use. Microsoft adds new bells and whistles in each release and claims that this time they've solved the countless problems in the previous versions … but the hype is never really fulfilled*" (Anonymous, available at: http://alternatives.rzero.com/os.html [Accessed on 5/1/2017]).

Bloem, J., van Doorn, M., Duivestein, S., Excoffier, D., van Maas, R. Ommeren, E.: Fourth industrial revolution. VINT research report. 3 of 4. https://slidelegend.com/queue/the-fourth-industrial-revolution-sogeti_59b503731723ddf2725f00c7.html (2014)

Bostrom, N.: Superintelligence. Oxford University Press, Oxford (2015)

Bourke, V.J.: History of Ethics, Vol. I, V.II. Axios Press, Mount Jackson, VA (2008)

Boyle, A.: AI prophets say robots could spark unemployment – and a revolution. Geekwire, February 13 (2016)

Brown, A.: YOUR job won't exist in 20 years: Robots and AI to 'eliminate' ALL human workers by 2036. https://www.express.co.uk/life-style/science-technology/640744/Jobless-Future-Robots-Artificial-Intelligence-Vivek-Wadhwa (2016)

Cathcart, T.: The Trolley Problem. Workman Publishing, New York (2013)

Clifford. C: The robots will take our jobs. Here's why futurist ray Kurzweil isn't worried. Entrepreneur. https://www.entrepreneur.com/article/272212 (2017)

Cookson, C.: AI and robots threaten to unleash mass unemployment, scientists warn. Financial Times. February (2016)

Dancy, J.: The role of imaginary cases in ethics. Pac. Philos. Q. **66**, 141–153 (1985)

Doris, J.M.: Lack of Character: Personality and Moral Behavior. Cambridge University Press, Cambridge (2002)

Elster, J.: How outlandish can imaginary cases be? J. Appl. Philos. **28**(3), 2011 (2011)

Floreano, D., Godjecac, J., Martinoli, F., Nicoud, J.-D.: Design, control and applications of autonomous mobile robots. Swiss Federal Institute of Technology, Lausanne. https://infoscience.epfl.ch/record/63893/files/aias.pdf (1998)

Ford, M.: The Rise of Robots: Technology and the Threat of Jobless Future. Basic Books, New York (2015)

Gray, G.: Heresies Against Progress and Other Illusions. Granta Publications, London (2007)

Hall, D.: Tesla Is Testing Self-Driving Cars on California Roads. https://www.bloomberg.com/news/articles/2017-02-01/tesla-is-testing-self-driving-cars-on-california-roads (2017)

Hern A.: Google's Waymo invites members of public to trial self-driving vehicles. https://www.theguardian.com/technology/2017/apr/25/google-self-driving-waymo-invites-members-public-trial-vehicles-phoenix-arizona (2017)

Heron, M., Belfort, P.: Fuzzy ethics: or how I learned to stop worrying and love the bot. SIGCAS Comput. Soc. **45**(4), 13 (2015)

Kaczynski, T.: Industrial society and its future. http://editions-hache.com/essais/pdf/kaczynski2.pdf (1995)

Krzanowski, R. Mamak, K. Trombik, K., Gradzka, E.: Ethics computable, non-computable or nonsensical? In: Defense of Computing Machines. Machine Ethics and Machine Law Conference. Jagiellonian University, Cracow, Poland, 18–19 November 2016

Lehtonen, T.: Idealization and exemplification as tools of philosophy. E-logos. Electro. J. Philos. **16** (2012)

MacIntyre, A.: A Short History of Ethics. Notre Dame Press, Notre Dame (1998)

McCumber, J.: Reshaping Reason. Indiana University Press, Bloomington (2007)

Milgram, S.: Behavioral study of obedience. J. Abnorm. Soc. Psychol. **67**(4), 371–378 (1963)

Moor, J.H.: The nature, importance and difficulty of machine ethics. IEEE Intell. Syst. 18–21 July/August 2006

Moor, J.H.: Four kinds of ethical robots. Philosophy Now. **72**, 12–14 (2009)

Ni, R., Leug. J.: Safety and liability of autonomous vehicle technologies. https://groups.csail.mit.edu/mac/classes/6.805/student-papers/fall14-papers/Autonomous_Vehicle_Technologies.pdf (2016)

Oppy, G., Dove D.: The Turing Test. The Spring 2016 Edition of the Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/archives/spr2016/entries/turing-test/ (2016)

Patrick, L., Bekey, G., Abney, K.: Autonomous Military Robotics: Risk, Ethics, Design. US Department of Navy, Office of Naval Research, Arlington (2008)

Pew Research Center. AI, robotics, and the future of jobs. http://www.pewinternet.org/2014/08/06/future-of-jobs/ (2014)

Sandel, M.J.: Justice: What's the Right Thing to Do? Penguin Books, London (2010)

Saygin, A.P., Cycelki, I., Akman, V.: Turing test: 50 years after. Mind. Mach. **10**, 463–518 (2000)

Schwab, K.: Why everyone must get ready for the 4th industrial revolution. http://www.forbes.com/sites/bernardmarr/2016/04/05/why-everyone-must-get-ready-for-4th-industrial-revolution/2/#a9fc30f40c8c (2016)

Schwab, K.: The Fourth Industrial Revolution. Crown Business, New York (2017)

Seibt, J.: "Integrative social robotics" - a new method paradigm to solve the description problem and the regulation problem? In: Frontiers in Artificial Intelligence and Applications. Volume 290: What Social Robots Can and Should Do. IOS Press, Amsterdam (2012)

Seibt, J.: Towards an ontology of simulated social interaction. In: Hakli, R., Seibt, J. (eds.) Sociality and Normativity for Robots. Studies in the Philosophy of Sociality, vol. 9. Springer, New York (2017)

Sparrow, R.: The Turing triage test. Ethics Inf. Technol. **6**(4), 203–213 (2004)

Sparrow, R.: The Turing Triage Test. When is a robot worthy of moral respect? http://www.thecritique.com/articles/the-turing-triage-test-when-is-a-robot-worthy-of-moral-respect/ (2014)

Stanford Prison Experiment. https://www.prisonexp.org (2014)

Stillgoe, J.: Self-driving cars will only work when we accept autonomy is a myth. https://www.theguardian.com/science/political-science/2017/apr/07/autonomous-vehicles-will-only-work-when-they-stop-pretending-to-be-autonomous (2017a)

Stillgoe, J.: Machine learning, social learning and the governance of self-driving cars. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2937316 (2017b)

Sullins, J.: Introduction: open questions in roboethics. Philos. Technol. **24**, 233 (2011)

Szabó, G.T.: Thought Experiment: On the Powers and Limits of Imaginary Cases. Routledge, New York (2000)

Turing, A.M.: Computing machinery and intelligence. Mind. **49**, 433–460 (1950)

Turner, R.: The Philosophy of Computer Science. The Winter 2016 Edition of the Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/computer-science/ (2016)

Vaughan-Nichols, S.J.: At Microsoft, quality seems to be job none. Computerword. 16 December 2014

Vardy, P., Grosch, P.: The Puzzle of Ethics. Fount, London (1999)

Veach, H.B.: Rational Man. Indiana University Press, London (1973)

Yampolskiy, R.V.: Leakproofing singularity - artificial intelligence confinement problem. J. Conscious. Stud. (JCS). **19**(1–2), 194 (2012a)

Yampolskiy, R.V.: Artificial intelligence safety engineering: why machine ethics is a wrong approach. In: Müller, V.C. (ed.) Philosophy and Theory of Artificial Intelligence, SAPERE, vol. 5, pp. 389–396. Springer, New York (2012b)

Yampolskiy, R.V., Fox, J.: Safety engineering for artificial general intelligence. Topoi. https://intelligence.org/files/SafetyEngineering.pdf (2012)