*proceedings*

MDPI

*Proceedings*

# Just Machine Test (JMT) †

**Roman M. Krzanowski * and Kamil Trombik**

The Pontifical University of John Paul II, Cracow, ul. Kanonicza 2, Kraków 31-002, Poland

* Correspondence: rmkrzan@gmail.com.

† Presented at the IS4SI 2017 Summit DIGITALISATION FOR A SUSTAINABLE SOCIETY, Gothenburg, Sweden, 12–16 June 2017.

Published: 8 June 2017

**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 'moral' aptitude, an aptitude to make 'good' rather than 'bad' choices. This paper discusses what such a test would consist of. We will call this test the machine ethics test or the Just Machine Test (JMT).The Just Machine Test 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.

**Keywords:** machine ethics; Turing test; machine ethics test; ethics computability; autonomous ethical agents

## 1. Problem

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 [7]. Autonomous machines will enter our lives not as passive devices, as has so far been the case with machines, but as active agents. We will depend on those devices in many, sometimes critical, ways and we will have to learn to live with them. But how can we be sure that those machines will not under any circumstances harm us, and will act as 'reasonable beings'; or, as one would say, that those machines will behave as ethical agents (machines)?

To answer this we need a verification criterion: a test that would verify the autonomous machine's 'moral' aptitude, an aptitude to make 'good' rather than 'bad' choices. This paper discusses what such a test would consist of. We will call this test the ethical machine test or the Just Machine Test (JMT).

## 2. Claim

The Just Machine Test 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 objective of the JMT is only to verify that:

- the product we develop, i.e., an autonomous agent, a machine, is 'safe' for intended use, and
- the propensity of the autonomous machine to do harm, by 'intention', design or neglect, is limited to as a narrow margin as reasonably possible.

## 3. What Should It Be?

The proposed test is thought of as the test verifying the general 'moral' or ethical aptitude of a machine, not an aptitude in specific circumstances. Passing the test for example for a housemaid or a

sex worker or a nurse would not count as passing the just machine test, as these tests are too narrow, too specific to be counted as a verification of the general moral capacity of an agent.

This assumption is similar to the assumption made in the original Turing test for a "thinking machine" [1]. 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 requirements.

## 4. What Should It Contain?

The JMT should not be theoretical or primarily theoretical (theoretical as involving only reasoning verified in a dialogue). We want autonomous machines to be able to make just decisions in specific life situations. Of course, an artificial agent will have to possess some abstract knowledge of ethics. But we would prefer that a machine makes just choices in concrete life situations rather than be able to respond to complex ethical questions. The complexity of life is not the same as complexity within theoretical cases; the difference is in quality not quantity.

Acknowledging the limitations of this analogy, we may compare the JMT to the skipper patent test or an airline pilot test. These tests include theoretical and practical components. What is interesting is that these tests include also 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 can be recognized as a pilot; likewise a skipper permit. Thus, it seems that the proper test of the artificial moral agent test should consist of a theoretical part, a series of practical problems, and maybe include a period of apprenticeship during which we verify an agent's capacity to think morally in real-life situations.

## 5. How We Would Evaluate Results?

How would we know that the machine passed the test? A panel of judges would have to review the results of the test and develop a test-passing score. Should we accept the Turing criterion of a 70% fail score? Possibly. But what would it mean to have a 30% ethical agent? Or, would we accept a 30% ethical machine to be among us? It is easier, it seems, to use a 70% fail score to judge that a machine functions reasonably, but not whether it has a 30% moral aptitude (in the TT the machine 'cheats' the judges in 30% of cases). It seems that any number, short of 100%, as the criterion of acceptance of the JMT results would be, in this case, an arguable qualification.

## 6. Suggestions

### *6.1. Situational Test*

It seems that the bulk of the JMT test should be situational. Examples of situational tests from which we may take some guidance about our JMT design include:

- Milgram Experiment [2]
- Phone booth and Dime experiment [3]
- Stanford Prison Experiment [4]
- Cornel Experiment [3]

Of course, this is in no way a closed list, and nor is its order of any import.

The scenario from the TT in which a machine is engaged in a conversation (actually an imitation game) with people who are trying to guess the machine's identity solely based on its responses to questions, is not the right way to test the moral aptitude of an artificial agent. Such a situation does not have the complexity of a situational test, and thus does not present an opportunity to verify an agent's moral capacities.

### *6.2. The Reasonable Person Test*

The JMT test may also be based on the concept of a reasonable person. The reasonable person is a rule used in the legal profession to assess whether a person violates the duty of due care in specific

circumstances. The rule (in one of its formulations) states that "*…an individual breaches her duty if she fails to obey certain norms of reasonable behavior (of a reasonable person)*" [5]. The reasonable person rule has been assailed on all sides. Yet, as a possible test for machine ethics it may provide some guidelines on how to look at the autonomous machine behavior and how to evaluate it. For this purpose we could even coin a term "the reasonable machine rule", a rule of applying a standard of the reasonable person to machine behavior.

### *6.3. The Just Person Test*

Another approach to the JMT would be to base such a test on the idea of a just person. The Just Person test would be based on the definition of what a just person is. The standard definition states that:

> *The "just" person is one who typically "does what is morally right" and is disposed to "giving everyone his or her due," offering the word "fair" as a synonym*[6].

The definition of a just person is complex with a long history and its exact interpretation depends on the ethical school discussing it. Still, as the concept of justice and a just person are at the core of many ethical disputes [6], they may be used to some degree in defining what the just machine should be, what it cannot be and how it should behave.

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

A. Theoretical verification of ethical aptitudes and reasoning—possibly a qualifying interview rather than a Q&A session plus a white box option.
B. 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:
    a. Staged tests;
    b. Controlled life situations;
    c. Open-ended situations.
C. A period of apprenticeship in which an artificial agent acts in the real conditions under close supervision.

## 7. 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 man 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 behavioral rules for machines not ethics in the human context. And the JMT is supposed to test just this, not to test the presence of some kind of metaphysical moral fiber in hardware. *It is rather difficult to image that machines will have the same complex of values that people have and thus the same responsibilities towards us. Thus, the JMT will verify not how close computing machines come to us but how close they come to our expectations about autonomous machines. Commander Data is just but a dream.*

If history teaches us anything, in this case it may indicate that the ethics of autonomous artificial agents may go the way internet security has gone: it means that as software companies do not take responsibility for damage caused by their faulty software 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.

**Author Contributions:** Both authors contributed to the idea presented in the paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Turing, A.M. Computing Machinery and Intelligence. *Mind* **1950**, *49*, 433–460.
2. Milgram, S. Behavioral Study of Obedience. *J. Abnorm. Soc. Psychol.* **1963**, *67*, 371–378.
3. Doris, J.M. *Lack of Character: Personality and Moral Behavior*; Cambridge University Press: Cambridge, UK, 2002.
4. Stanford Prison Experiment. Available online: http:// www.prisonexp.org/ (accessed on 28 December 2016).
5. Miller, A.D. The Reasonable Man and Other Legal Standards. Available online: http://people.hss.caltech.edu/~alan/ReasonableMan.pdf (accessed on 28 December 2016).
6. Pomerleau, W.P. Western Theories of Justice. Available online: http://www.iep.utm.edu/justwest/ (accessed on 28 December2016).
7. Ford, M. *The Rise of the Robots: Technology and the Threat of Mass Unemployment*; Oneworld Publications: London, UK, 2016.