THE TRANSFORMATION OF "ARTIFICIAL" SCIENCE INTO ARTIFICIAL INTELLIGENCE: 50 YEARS LATER

Boris Aberšek
University of Maribor, Slovenia

> *"Natural science is knowledge about natural objects and phenomena. We ask whether there cannot also be 'artificial' science – knowledge about artificial objects and phenomena."*
>
> *Herbert Simon (1969)*

In the Beginning

For years, experts have warned against the unanticipated effects of general artificial intelligence (AI) on society. Ray Kurzweil (1998, 2005) predicts that by 2029 intelligent machines will be able to outsmart human beings. Stephen Hawking argues that *"once humans develop full AI; it will take off on its own and redesign itself at an ever-increasing rate"*. Elon Musk warns that AI may constitute a "fundamental risk to the existence of human civilization". If the problems of incorporating AI in manufacture and service operations, i.e. using *smart machines*, are smaller, as the 'faults' can be recognized relatively quickly and they do not have a drastic effect on society, then the *incorporation of AI in society and especially in the educational process* is an extremely risky business that requires a thorough consideration. The consequences of mistakes in this endeavour could be catastrophic and long-term, as the results can be seen only after many years.

AI is ultimately only a computer program, a "simple" optimization algorithm. Such algorithms can contain different ethical constraints (law) in the source code. A well-known historical example in the form of such simple "robotic laws" dates as far back as 1950, when Isaac Asimov proposed the following:

1. A robot may not injure a human being, or, through inaction allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law (Asimov, 1950).

It is clear from these laws that the robot (intelligent machine), or, in today's terminology, AI, must protect humans and put the safety of human beings before its own existence. 50 years later, however, Mark W. Tilden wrote similar, but at the same time different laws[1]:

[1] http://www.botmag.com/the-evolution-of-a-roboticist-mark-tilden/

https://doi.org/10.33225/jbse/20.19.340

1. A robot must protect its existence at all costs.
2. A robot must obtain and maintain access to a power source.
3. A robot must continually search for better power sources.

Tilden's laws suggest that the primary role of the robot (AI) is first and foremost to protect itself from the outside world, including human beings. Because the AI of today learns primarily from the world wide web, where both types of laws can be found, an ethical dilemma could thus be created: *which of these two sets of laws should be considered as guidelines, or, in other words, what is the Categorical Imperative for AI according to Kant (1981)?*

Machine Ethics and/or Machine Behaviour

Machine morality in intelligent systems, whether physical systems with a mind and body or just thinking algorithms somewhere in the cloud, is a recurring issue. Morals demonstrate the relationship of humanity to nature and society and are manifested as a sum of values (rules, norms, principles, categories, ideals, etc.), according to which we make decisions, what is good and what is bad, what is just and what is unjust, what is right and what is wrong, and in line with which we also behave (Kordigel Aberšek, 2012). When it comes to the morality of smart machines, philosophers mostly focus on theoretical questions such as: *does AI have the status of a moral agent, is AI responsible for its actions, is AI a 'being' with a higher moral status*, etc. – rather than on such a specific and practical area as is the usage of AI in education, especially in the field of ensuring social competences and developing emotional intelligence (Aberšek, 2013, Kordigel Aberšek & Aberšek, 2020).

The ethical dilemma related to the understanding and interpretability of the behaviour of AI agents, is one of the pivotal challenges of the next decade of AI. Until today, most of the interpretability techniques have focused on exploring the internal structure of deep neural networks. But *machine behaviour* (Rahwan et al., 2019) relies more on observations than on engineering knowledge in order to understand the behaviour of AI agents. Most of the conclusions obtained from observations in nature are not related to knowledge from biology, but rather to our understanding of social interactions. In the case of AI, scientists who study the behaviours of different virtual and embodied AI agents are predominantly the same scientists who have created the agents themselves. But understanding AI agents must go beyond interpreting a specific algorithm and requires analyzing the interactions between agents and with the surrounding environment. In order to accomplish that, behavioural analysis via simple observations can be used as a powerful tool.

Machine Behaviour

Machine behaviour (Rahwan et al., 2019) is a field that leverages behavioural sciences to understand the behaviour of AI agents. Currently, scientists who most commonly study the behaviour of machines are computer scientists, roboticists and engineers who have created the machines in the first place, but they are typically not trained behaviourists. Similarly, even though behavioural scientists understand those disciplines, they lack the expertise to understand the efficiency of a specific algorithm or technique. From that perspective, machine behaviour sits at the intersection of computer science, engineering, and behavioural sciences, in order to achieve a holistic understanding of the behaviour of AI agents. As AI agents become more sophisticated, analyzing their behaviour is going to be a combination of understanding their internal architecture (the domain of computer scientists), as well as their interaction with other agents and their environment (the domain of behavioural scientists). While the former aspect will be a function of deep learning optimization techniques, the latter will rely partially on behavioural sciences.

As a starting point in the development of a new transdisciplinary science, which could be termed *AI behavioural science,* Nikolaas Tinbergen's work (1963) can be used for identifying the key dimensions of animal behaviour. Tinbergen's thesis was that there were four complementary dimensions to understand animal and human behaviour, these are: *Mechanism, Development, Function,* and *Evolution* (Nesse, 2013).

Despite fundamental differences between AI and animals, machine behaviour borrows some of Tinbergen's ideas to outline the main types of behaviour in AI agents. Machines have *mechanisms* that produce behaviour, undergo *development* that integrates environmental information into behaviour, produce *functional consequences* that cause specific machines to become more or less common in specific environments, and embody *evolutionary histories* through which past environments and human decisions continue to influence machine behaviour.

https://doi.org/10.33225/jbse/20.19.340

These four dimensions provide a holistic model for understanding the behaviour of AI agents. However, these four dimensions do not apply in the same way with respect to whether we are evaluating a classification model with a single agent, or with hundreds of agents. In that sense, machine behaviour applies the previously mentioned four dimensions across three different scales. The first is **Individual Machine Behaviour:** this dimension of machine behaviour attempts to study the behaviour of individual machines by themselves. There are two general approaches to the study of individual machine behaviour. The first focuses on profiling the set of behaviours of any specific machine agent using a within-machine approach, comparing the behaviour of a particular machine across different conditions. The second, a between-machine approach, examines how a variety of individual machine agents behave in the same conditions. The second scale is **Collective Machine Behaviour:** unlike the individual dimension, this area looks to understand the behaviour of AI agents by studying the interactions in a group. The collective dimension of machine behaviour attempts to spot behaviours of AI agents that do not surface at an individual level. And finally, the scale of **Hybrid Human-Machine Behaviour:** there are many scenarios in which the behaviour of AI agents is influenced by their interactions with humans. This dimension of machine behaviour focuses on analyzing behavioural patterns in AI agents triggered by the interaction with humans.

Solution

What can be done? In trying to provide a solution, a simple example related to the notion of *proprioception* (Aberšek, 2018) can be considered. What does proprioception really mean? Proprioception could also be called *self-perception of thought*, or *self-awareness of thought*, i.e., thought, which is able to perceive its own flow, be aware of its own movement. Alongside proprioception, the emotional intelligence of a person also develops, and the changes that occur in this process will affect, step by step, the human historical memory, and add new elements to this historical memory on the level of intuitive thinking. By way of analogy, a similar philosophy for proprioception in AI can be developed. It is important that this kind of awareness be developed in every individual – human or AI; "changes" or adaptations must be made to the specific way of thinking (creative, critical, and conscious thinking), and it is very important to begin this process with agents (human or AI) of the "youngest" possible age. In this sense, competences should be developed gradually, step by step, to enable dealing with the day-to-day needs of others, and help raise the awareness.

Machine Behaviour and Education

Before any kind of learning environment is given some sort of intelligence, machine ethics and/or machine behaviour must be built into this learning environment, in order to ensure that the cognitive, social, and emotional competences of students are defined in a way that will allow them to be formalized or translated into a scientific language, into a language familiar to the machine. Additionally, methods have to be defined for assessing whether such intelligent systems work correctly in the long-term, since either noticing or removing the consequences which their failure or irregular operations have on the moral development of individuals, is not possible in real time. And since these methods, as mentioned earlier, are not in the domain of computer scientists, roboticists and engineers who have created the machines, but rather in the hands of experts from the field of behavioural science, the roles of the evaluator and the auditor must take over the role of teachers. For this reason, teachers must be able to acquire some kind of knowledge from the area of AI behavioural science in order to become competent observers and evaluators of such intelligent learning environments (Balogh & Kucharik, 2019, Kordigel Aberšek, 2012).

The general question to be answered could therefore be formulated thus: *"What are the moral problems of using advanced learning systems and modern learning environments supported by AI methods*?", with the concrete goal of the research being *the development of a test, on the basis of which teachers could assess whether an intelligent accessory (program or algorithm) for learning is such that it ensures the acquisition of all cognitive, social, and emotional competences in students*, i.e., whether it is 'safe' to use in the educational process. The development of such a test, as well as the related knowledge and skills, could encourage the development of various other similar 'security' tests for AI usage in other areas.

https://doi.org/10.33225/jbse/20.19.340

Summing-up

Machine behaviour is one of the most intriguing, nascent fields and AI. Behavioural sciences can support traditional interpretability methods in developing new methods that will help to better understand and explain the behaviour of AI. As the interactions between humans and AI become more sophisticated, machine behaviour might play a crucial role to enable the next level of hybrid intelligence. From all of the above it can be concluded that at least the following three guidelines should be taken into consideration, especially with respect to using intelligent learning environments in education:

1. Not every kind of AI is a benefit to mankind, and not all uses of AI are ethical and moral.
2. The ethical use of AI should be judged not only by computer scientists, roboticists and engineers, but (especially) by behavioural scientists.
3. Teachers need to be trained (empowered) and provided with appropriate competences to assess the usefulness and ethical use of AI.

References

Aberšek, B. (2018). *Problem-based learning and proprioception*. Cambridge Scholars Publishing.
Aberšek, B. (2013). Cogito ergo sum homomachine? *Journal of Baltic Science Education*, *12* (3), 268-270.
Asimov, I. (1950). *I robot*. Gnome Press.
Balogh, Z., & Kucharik, M. (2019). Predicting student grades based on their usage of LMS Moodle using Petri Nets. *Applied Sciences, 9*(20), 4211, 1-16. https://doi.org/10.3390/app9204211
Kant, I. (1981). *Grounding for the metaphysic of morals*. Heckett Publishing Company.
Kordigel Aberšek, M., & Aberšek, B. (2020). *Society 5.0 and literacy 4.0 for 21st century.* Nova Science Publishers.
Kordigel Aberšek, M. (2012). Neuroscience, world wide web and reading curriculum. *Problems of Education in the 21st Century, 46*, 66-73. http://www.scientiasocialis.lt/pec/node/files/pdf/vol46/66-73.Kordigel-Abersek_Vol.46.pdf
Kurzweil, R. (2005). *The singularity is near*. Viking Press.
Kurzweil, R. (1998). *The age of spiritual machines: When computers exceed human intelligence*. Viking Press.
Nesse, R. M. (2013). Tinbergen's four questions, organized: A response to Bateson and Laland. *Trends in Ecology & Evolution*, *28*, 681–682. https://doi.org/10.1016/j.tree.2013.10.008
Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., Roberts, M. E., Shariff, A., Tenenbaum, J. B., & Wellman, M. (2019). Machine behaviour. *Nature*, *568,* 477 - 286. https://doi.org/10.1038/s41586-019-1138-y
Simon, H. A. (1969). *The sciences of the artificial.* MIT Press.
Tinbergen, N. (1963). On aims and methods of ethology. *Ethology*, *20*, 410–433.

Received: *March 28, 2020*

Accepted: *May 20, 2020*

Cite as: Aberšek, B. (2020). The transformation of "artificial" science into artificial intelligence: 50 years later. *Journal of Baltic Science Education*, *19*(3), 340-343. https://doi.org/10.33225/jbse/20.19.340

Boris Aberšek PhD, Professor, University of Maribor, Faculty of Natural Sciences and Mathematics, 2000 Maribor, Slovenia.
E-mail: boris.abersek@um.si
Website: https://scholar.google.com/citations?user=aRid0w4AAAAJ&hl=en
ORCID: http://orcid.org/0000-0003-0563-7329