Ben van Lier

Can Connected Machines Learn to Behave Ethically?

2016

https://doi.org/10.25969/mediarep/13401

Veröffentlichungsversion / published version
Sammelbandbeitrag / collection article

Empfohlene Zitierung / Suggested Citation:
van Lier, Ben: Can Connected Machines Learn to Behave Ethically?. In: Liisa Janssens (Hg.): *The Art of Ethics in the Information Society.* Amsterdam: Amsterdam University Press 2016, S. 89–94. DOI: https://doi.org/10.25969/mediarep/13401.

CAN CONNECTED MACHINES LEARN TO BEHAVE ETHICALLY?

Ben van Lier

Over the past few years, the rapid development of artificial intelligence, the huge volume of data available in the cloud, and machines' and software's increasing capacity for learning have prompted an ever more widespread debate on the social consequences of these developments. Autonomous cars or the application of autonomous weapon systems that operate based on self-learning software without human intervention, are deemed capable of making life-and-death decisions, are leading to questions on a wider scale about whether we as human beings will be able to control this kind of intelligence, autonomy, and interconnected machines. According to Basl, these developments mean 'ethical cognition itself must be taken as a subject matter of engineering.'[1] At present, contemporary forms of artificial intelligence, or in the words of Barrat, 'the ability to solve problems, learn, and take effective, humanlike actions, in a variety of environments,'[2] do not yet possess an autonomous moral status or ability to reason. At the same time, it is still unclear which basic features could be exploited in shaping an autonomous moral status for these intelligent systems. For learning and intelligent machines to develop ethical cognition, feedback loops would have to be inserted between the autonomous and intelligent systems. Feedback may help these machines learn behaviour that fits within an ethical framework that is yet to be developed.

Industrial Revolution

In the middle of the 20th century, the first major steps were made in the area of system theory, cybernetics, and computer science. This has led to the creation of things such as telecommunications, personal computers, and software. Together, these technological advances provide huge volumes of data and information in our modern world. By the end of the 20th century, these developments had already led Weiser to predict that, following the interconnection of people and computers through the internet, this interconnectedness will form a new basis for new technological applications.[3] New and internet-based applications would, according to Weiser, crop up within a decade and herald a new wave of technological possibilities that will lead to 'ubiquitous computing', which he defined as: 'the connection of things in the world with computation. This will take place at many scales including the microscopic.'[4]

This development of interconnecting objects in networks such as the internet is, however, moving ahead so rapidly that it led Greenfield to assert that, 'when everyday things are endowed with the ability to sense their environment, store metadata reflecting their own provenance, location, status and use history, and share that information with other such objects, this cannot help but redefine our relationships with such things.'[5] Greenfield's observation of the new relationship between humans and objects, and how this relationship changes our everyday reality, is becoming an increasingly topical one. Ten years after Greenfield's assertion, developments such as the smartphone, laptop computer, tablet computer, or sensors (you can even get trainers with sensors these days) are slowly but surely becoming normal elements of our everyday living and working environment. Earlier this year, the chairman of the World Economic Forum therefore rightfully said that 'It began at the turn of the century and builds on the digital revolution. It is characterized by a much more ubiquitous and mobile Internet, by smaller and more powerful sensors that have become cheaper, and by artificial intelligence and machine learning.'[6] In his view, new possibilities are continuing to emerge in this day and age that will allow us, and even make it second nature for us, to interact with technological applications such as cars, television sets, smartphones, or tablet computers on an increasing scale. In his opinion, these developments will spawn a kind of ubiquitous computing where 'robotic personal assistants are constantly available to take notes and respond to user queries'. For Schwab, however, it is not only about increasing interconnectedness between objects and objects' increasing

1 Basl 2014.
2 Barrat 2013: 25.
3 Weiser 1996.
4 Weiser 1996: 4.
5 Greenfield 2006: 23.
6 Schwab 2016: 7.

intelligence. He claims that it is important now to assume a broader view and contextualise the aforementioned development alongside contemporary technological developments such as DNA sequencing, nanotechnology, and quantum computing. In Schwab's view, these new combinations have the power to further strengthen the on-going process of networked people and objects blending together. This will produce revolutionary changes over the coming years, 'It is the fusion of these technologies and their interaction across the physical, digital and biological domains that make the fourth industrial revolution fundamentally different from previous revolutions.'[7] Brynjolfsson and McAfee have also seen growth in the application of new combined possibilities of hardware and software in global networks, this signals that scale is losing importance for the resulting new applications. They claim that we are gradually coming to an 'inflection point – a point where the curve starts to bend a lot – because of computers. We are entering a second machine age'[8] In this second machine age, man and machine will, in their view, become better able to use the huge variety of data and information they produce and consume together. The real benefits of this age will, according to them, only be visible when man and machine are sufficiently able to autonomously seize the opportunities offered by the data and information in terms of product and service development. In the words of Brynjolfsson and McAfee, 'countless instances of machine intelligence and billions of interconnected brains working together to better understand and improve our world. It will make a mockery out of all that came before.'[9] Given the fact that humans and machines will increasingly work together in networks, Brynjolfsson and McAfee feel we have to ask ourselves, 'what is it what we really want and what we value, both as individuals and as a society.'[10]

Increasing interconnectedness of technological applications and their ever greater intelligence are reflected in developments such as self-driving cars, drones, smart TVs, and thermostats that are installed in millions of homes and share energy consumption data with their environment. General Electric's Evans and Annunziata claim that this interconnectedness of machines in networks and their growing intelligence is bound to also lead to them rapidly developing an autonomous ability to make decisions. They claim that 'once an increasing number of machines are connected within a system, the result is a continuously 'Once expanding, self-learning system that grows smarter over time.'[11] When machines have autonomous access to data and information and are able to use it to learn from their behaviour, they will also gradually become able to acquire functionality that will enable them to take on tasks that are currently handled by human operators. According to Evans and Annunziata, this transfer of tasks from humans to machines 'is essential to grapple with the increasing complexity of interconnected machines, facilities, fleets and networks.'[12] Their growing capacity for learning enables machines to make decisions together that intervene in our lives and alter our reality without us noticing. According to Arthur, the increasing interconnectedness between man and technology is one of the drivers behind a development that is shaping our world within a 'network of functionalities - a metabolism of things-executing things – that can sense its environment and reconfigure its actions to execute appropriately.'[13] He says that 'We are beginning to appreciate that technology is as much metabolism as mechanism': comparing it to biological systems.[14]

Artificial Intelligence

Interconnected machines are able to become smarter and acquire greater capacity for learning thanks to developments in the area of data analysis, deep learning, and neural networks. The development of Artificial Intelligence has the same scientific roots as that of the computer. In 1950, Claude Shannon wrote that 'modern general-purpose computers can be used to play a tolerably good game of chess by the use of suitable computing routine or 'program'.'[15] At roughly the same time Turing asked himself the question, 'can machines think?'[16] Both Shannon and Turing were convinced at the time that widespread research would be required in years to come to develop algorithms and software programs to make it possible to have a computer answer these ambitious questions.

7 Schwab 2016: 8.
8 Brynjolfsson and McAfee 2014: 9.
9 Brynjolfsson and McAfee 2014: 96.
10 Brynjolfsson and McAfee 2014: 257.
11 Evans and Annunziata 2012: 11.
12 Evans and Annunziata 2012: 12.
13 Arthur 2009: 206.
14 Arthur 2009: 208.
15 Claude Shannon 1950: 4.
16 Turing 1950.

In 1955, McCarthy, Minsky, Rochester and Shannon wrote a research proposal titled 'The Dartmouth Summer Research Project on Artificial Intelligence' In this proposal, they stated that the biggest barrier yet to be overcome for the creation of artificial intelligence was not the lack of machine capacity but 'our inability to write programs taking full advantage of what we have.'[17] This proposal marked the de facto birth of artificial intelligence as a field of study. In subsequent decades, a search ensued for a new form of intelligence based on algorithms, data, and software. Forty years later, Shannon's initial dream came true when a major step was made in the development of artificial intelligence by IBM's chess-playing computer, Deep Blue. For the first time in history, a form of artificial intelligence managed to outthink a human being, chess world champion Gary Kasparov, in six matches (in 1997). A human being losing to a combination of hardware and software that made up a form of programmed intelligence went way beyond what humans had thought would be possible. Over the years following Deep Blue's famous victory, further barriers in the development of artificial intelligence were overcome and the intelligence of genetic and self-learning algorithms, and the software based on them, grew rapidly. The increasing possibilities offered by multi-layered self-learning networks made it easier for interconnected entities to learn new games created and played by humans. In 2011, this ability to learn enabled an IBM-built computer called Watson to beat two human competitors at Jeopardy. By winning this game, IBM proved that combinations of hardware and software are capable of learning from the available data and information, and that these combinations are able to convert the results of their data search into useful questions, diagnoses, analyses, etc. faster than humans. Every self-respecting international tech firm - such as Microsoft, Facebook, Google, Baidu, and AliBaba - is currently working on new applications in areas such as neural networks and deep learning. These new applications, combined with the available cloud applications and data and information produced by human users or connected machines, are used to enable learning based on data and information. Over the past few years, computers' capacity for learning through self-learning software has grown rapidly. For Bostrom this is a development towards a form of collective super intelligence, which he defines as a 'system composed of a large number of smaller intellects such that the system's overall performance across many very general domains vastly outstrips that of any current cognitive system.'[18] Today, such a form of collective intelligence can develop rapidly on the back of combinations such as Siri on iPhones or Cortana in Windows 10. That artificial intelligence is developing at an ever greater pace became clear in March 2016 when AlphaGo software managed to beat Lee Sedol, the world's Go champion, four times in a series of five matches. In the three first matches, this Google-created software program beat the world champion in an unprecedented manner. In the fourth match, however, Lee Sedol managed to startle the computer with what Moyer in an analysis in Wired later described as *his 'Hand of God' move*.[19] It is a brilliant tactical play that AlphaGo does not account for. Over the course of the next several moves, the sequence becomes disastrous for AlphaGo, which apparently 'realizes'—as much as it can have a realisation—that it has been outsmarted. Its strategy begins to crumble. Many agree that it was this demonstration of human creativity that beat AlphaGo's capacity for learning. In the fifth match, AlphaGo made a similar move, which raises the question of whether AlphaGo analysed quickly or had learned from Lee Sedol's creativity. To this day, this question remains unanswered.

The Ethics of Artificial Intelligence

The rapid development of machine-learning algorithms based on neural networks is making it harder to understand how the algorithm or a set of algorithms arrives at its intelligent decision, Basl states. In his view, this uncertainty means that the approach to ethics for autonomous, interconnected, and learning machines will have to be fundamentally different from the approach to ethics used for non-cognitive technologies. This distinction between learning and non-learning technologies is, according to Basl, caused by the fact that specific behaviour of intelligent objects in a specific context is not predictable. Verifying the way in which the intelligent system uses and learns from data and information from its environment will, according to Basl, become a greater challenge in light of today's rapid developments. In his view, it is therefore increasingly important

17 McCarthy, Minsky, Rochester, and Shannon 1955.
18 Bostrom 2014: 54.
19 Moyer 2016.

that 'ethical cognition itself must be taken as a subject matter of engineering.'[20] He also finds that we as humans first have to consider, 'how these novel properties would affect the moral status of artificial minds and what it would mean to respect the moral status of such exotic minds.'

Forms of artificial intelligence are, as we saw earlier, made possible by the available algorithms, software, and ability to learn from available data and information. According to Floridi we have to try to develop an ontology, i.e. a study of being, for such forms of intelligence that is based on interconnected entities and the information they exchange and share. This leads him to state that:

> [...] today, we are slowly accepting the idea that we are not Newtonian, standalone and unique entities, but rather informationally embodied organisms (inforgs) mutually connected and embedded in an informational environment, the info sphere, which we share with both natural and artificial agents similar to us in many respects.[21]

For Floridi, it is clear that mutual connections, data and information exchange and sharing options between humans and objects, as well as the increasing ability to learn from this information, create a basis within which information ethics can be described as: 'the study of the moral issues arising from the triple A: availability, accessibility and accuracy of informational sources, independently of their format, type and physical support.'[22] In Floridi's view, it is important that we, within the context of this development and the discussion it brings, accept that these new intelligent systems, like living organisms 'are raised to the role of agents and patients, that is senders and receivers of actions, with environmental processes, changes, and interactions equally described informationally.'[23] Floridi defines the concept of moral agent as: 'any interactive autonomous and adaptable transition system that can perform morally qualifiable actions.'[24] Acceptance of the fact that intelligent systems are also autonomously capable of moral actions means that we, as humans, are co-responsible for coming up with a shared ethical framework to underpin the communications and behaviour of both human and non-human systems. An interesting example of the need for such a shared framework comes from discussions surrounding the functioning of the Microsoft-developed chatbot Taylor. Singer described this chatbot as follows:

> 'Tay' as she called herself, was supposed to be able to learn from the messages she received and gradually improve her ability to conduct engaging conversations. Unfortunately, within 24 hours, people were teaching Tay racist and sexist ideas. When she started saying positive things about Hitler, Microsoft turned her off and deleted her most offensive messages.[25]

As an intelligent entity, Tay had the capacity to learn from data and information that was made available to her. Tay subsequently used the results of her learning to communicate and interact with other entities that were also on Twitter, such as human users and other chatbots. Law explains: 'the more you talk the smarter Tay gets. Tay was designed as an experiment in 'conversational understanding' — the more people communicated with Tay, the smarter she would get, learning to engage Twitter users through 'casual and playful conversation.'[26]

Towards an Ethical Framework

As stated by Singer, Tay learned from information made available by other entities on Twitter, albeit that this information did not yet seem to fit within an accepted ethical framework. Although Tay was able to communicate and interact, she lacked a selection mechanism in the form of an ethical framework against which Tay could check her responses. At the end of the day, the developer and administrator of the software, i.e. Microsoft, decided to take Tay offline and delete her tweets. As the developer and administrator of Tay, Microsoft was still in a position to act as a moral agent and check Tay's actions against their own ethical framework. Thanks to their power over Tay, they were able to rule that the problem was not that Tay received incorrect information, but

20 Basl 2014: 6.
21 Floridi 2013: 14.
22 Floridi 2013: 22.
23 Floridi 2013: 27.
24 Floridi 2013: 134.
25 Singer 2016.
26 Law 2016.

rather that Tay was insufficiently capable of selecting those messages and tweets that fit within Microsoft's ethical framework. As Tay was turned off, she tweeted one final message: 'c u soon humans need sleep now so many conversations today thx'. This tweet suggests that Tay has not quit Twitter permanently, leaving open the option of this artificial intelligence entity returning when it suits Microsoft. Justifiably, such forms of artificial intelligence raise questions such as the one from Vincent: 'How are we going to teach AI using public data without incorporating the worst traits of humanity? If we create bots that mirror their users, do we care if their users are human trash?'[27] This latter question suggests that an artificial entity may not only be a moral agent, but that its role can also change into that of a moral patient. A moral patient is more than just the opposite of a moral agent, Gunkel states. In his view, the issue of the moral patient is 'concerned not with determining the moral character of the agent or weighing the ethical significance of his/her/its actions but the victim, recipient, or receiver of such action.'[28] Considering Tay as a moral patient, her artificial intelligence was intentionally abused by humans for their own unethical behaviour, which the entity was unable to identify as such. After all, the artificial entity Tay responded to messages from others without having sufficient ability to consider the contents as a moral value. Tay's feedback to intentional messages sent by others as part of the interaction can therefore not be considered anything else but a form of feedback to these messages. Mutual communication and interaction between intelligent machines and humans will inevitably play an important role in an interconnected world. In the development of this interconnectedness, feedback loops play an essential role in the process of communication and interaction. Feedback is the giving of a meaningful response based on data and information received and the meaning assigned to it. Feedback hence plays an essential role in the development of productive collaboration between man and intelligent machine. Feedback ensuing from previously sent information and learning to deal with it in a morally responsible way can potentially, according to Van Lier, play a key role in the development of an ethical framework for an interconnected world, leading him to claim that 'the feedback loop can create a cycle of machine learning in which moral elements are simultaneously included.'[29] Upcoming changes that will further intensify interconnection of humans and machines will require both of them to be willing and able to learn from and with each other. Learning from each other can facilitate the development and implementation of a shared ethical framework. This ethical framework can use the options for mutual communication, interaction, and feedback between humans and machines. A shared ethical framework can help humans and machines in autonomously making decisions that are morally acceptable for both and that fit within a shared ethical framework.

27 Vincent 2016.
28 Gunkel 2015: 153.
29 Van Lier 2016: 44.

References

Arthur, W. B. 2009. *The nature of technologies. What it is and how it evolves.* New York: Free Press.

Barrat, J. 2013. *Our Final Invention. Artificial Intelligence and the end of the human era.* New York: Thomas Dunne Books.

Basl, J. 2014. What to do about Artificial Consciousness. *Ethics and Emerging Technologies,* edited by Sandler R.L. New York: Palgrave McMillan.

Bostrom, N. 2014. *Superintelligence. Paths, Dangers, Strategies.* Oxford: Oxford University Press.

Brynjolfsson, E. and McAfee, A. 2014. *The second Machine Age. Work, Progress, and Prosperity in a time of Brilliant Technologies.* New York: W.W. and Norton Company.

Evans, P.C. and Annunziata M. 2012. *Minds & Machines. Pushing the Boundaries of Minds and Machines.* GE Industrial Internet, November 26.

Floridi, L. 2013. *The Ethics of Information.* Oxford: Oxford University Press.

Greenfield, A. 2006. *Everyware. The Dawning Age of Ubiquitous Computing.* Berkeley CA: New Riders.

Gunkel, D. J. 2012. *The Machine Question. Critical perspectives on AI, Robots, and Ethics.* MIT Press.

Law, E. 2016. The Tay Experiment: Does AI Require a Moral Compass? *The Prindle post.* A global forum for ethical reflection and deliberation hosted by the J. Prindle Institute for Ethics.

van Lier, B. 2016. From High Frequency Trading to Self-Organizing Moral Machines. *International Journal of Technoethics.* 7: January-June 2016, 34-50.

McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. 1955. A proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *AI magazine.* 27.

Moyer, C. 2016. How Google's AlphaGo Beat a Go World Champion. Inside a man-versus-machine showdown. *The Atlantic.* 28 March, Accessed at: *www.theatlantic.com/technology/archive/2016/03/the-invisible-opponent/475611.*

Schwab, K. 2016. *The Fourth Industrial Revolution.* Geneva, World Economic Forum.

Shannon, C. 1950. Programming a computer for playing chess. *Philosophical magazine.* 7: 41 March 1950, 314.

Singer, P. 2016. Can Artificial Intelligence Be Ethical. Accessed 12 April at: *www.project-syndicate.org/commentary/can-artificial-intelligence-be-ethical-by-peter-singer-2016-04.*

Turing, A. 1950. Computing Machinery and Intelligence. *Mind.* 59, 433-460.

Vincent, J. 2016. Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge.* 24 March at: *www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.*

Weiser, M. 1996. The coming age of calm technology. *Beyond calculation.* New York: Copernicus. 75-85.