IDEA – Studia nad strukturą i rozwojem pojęć filozoficznych
XXIX/1 Białystok 2017

Adam Drozdek
(Pittsburgh, USA)

ETHICS AND INTELLIGENT SYSTEMS

Although philosophical and moral problems associated with artificial intelligence have always been present since the inception of this field, they become more and more important due to the ubiquity of computers and the proliferation of robots and AI systems. Because such systems are intelligent, some authors are interested in how the systems should be treated. Robots are introduced as robots, as workers, as tools that should facilitate the execution of certain tasks or even take over some tasks completely, particularly those that are repetitive, onerous, or dangerous. And then there is an opinion that intelligent entities should not be treated as mere tools. Therefore, if pressed to the extreme, there is a paradox: robots, intelligent or otherwise, have been invented as tools and yet because of their intelligence they should not be treated as such, whereby, presumably, they should be let loose to lead their robotic lives, and humans should get back to performing tasks from which they wanted to be freed by the use of robots. This opinion also leads to a whole host of issues of machine rights and responsibilities and thus to the establishment of new laws and mores.

There is also another, futuristic issue: it is conjectured that if machines become superintelligent that, at best, they may treat humans the way humans treat, say, ants, as an insignificant and minor annoyance that is, however, allowed to exist on the margins of the world; at worst, they may annihilate humans as using too many resources and contributing nothing to the world[1]. An-

[1] As an example, in one apocalyptic scenario, when machines become more intelligent than humans, humans will be enslaved by machines and permitted to live only as long as their usefulness can be extended, which is estimated to be around 30 years of age; afterwards they will be in-

other scenario sees supporters of the superintelligence waging a war against its detractors. Many solutions have been proposed to avert this impending problem, ranging from denying the existence of the problem to an active preparation for the conflict.[2]

1

As to the prospective robot rights, the reason given is that robots, as intelligent entities, deserve the same treatment as humans. There is, of course, a thorny problem of defining intelligence. Let intelligence signify an ability to solve problems, particularly new problems, particularly in new situations, problems as mundane as how quickly to reach a destination or as complicated as a higher math problem. Some machines are surely able to do that and they are built precisely to do it: to solve a problem and execute possibly also the solution: find in a database people with certain characteristics and show on the screen their locations or locate a target and destroy it. This more and more frequently can be done quicker and more effectively by machines than by humans, so, at least in some respects, machines are already smarter than humans. However, should we fear that a machine programmed for data mining is going to turn against humans? Should humans be afraid that a missile guiding system is going to turn malevolently against them in the mid-flight of the missile? Hardly. It is because – it can be answered – machines and their programs are too specialized; their intelligence is limited to a very narrow field and hence they cannot bring themselves from this limitation to turn against their human masters. It would have to be a wider intelligence, something along the lines of the HAL computer.

The HAL computer from the novel *2001: a space odyssey*, apparently a very intelligent system, turned against the crew, killing everyone except David who managed to disable HAL. The problem was that HAL was programmed to do

cinerated – and all of it fairly soon, around the year 2050, Kevin Warwick, *March of the machines: the breakthrough in artificial intelligence*, Chicago: University of Illinois Press 1997, ch. 2.

[2] The many solutions proposed so far have been summarized by Roman V. Yampolskiy, *Artificial superintelligence: a futuristic approach*, Boca Raton: CRC Press 2016, ch. 6. It is suggested that reasons for the popularity of such dark depictions of the future are rather prosaic since "apocalyptic AI is a social strategy for the acquisition of research funding," Robert M. Geraci, *Apocalyptic AI: visions of heaven in robotics, artificial intelligence, and virtual reality*, Oxford: Oxford University Press 2010, p. 38.

its best to continue the mission which was to contact aliens. This reason for the mission was unknown to the crew. Since HAL was also programmed to tell the truth, it wanted to avoid the possibility of lying to the crew, which it resolved by killing them off. This brings to mind H.G. Wells' short story, *The truth about Pyecraft*, in which the protagonist could use a charm to grant his wish which was to lose weight. As it happened, he did lose weight and floated to the ceiling remaining as corpulent as before making his wish. The wish was granted, literally, but not the way he intended. It is similarly in the *2001* story. When HAL was programmed to continue its mission, it did it, no matter the cost. It probably did not occur to the designer of HAL that a qualification should have been made not to kill any crew member in the process. What is obvious to humans is not obvious to machine intelligence, whereby the intelligence of HAL led it to rather stupid actions.[3] Something along the lines of Asimov's laws of robotics was needed. However, an interesting question arises: what if HAL had been successful in killing the entire crew and had reached its destination? It would make contact with aliens as it was designed to do and then what? It would be rather useless after completing its mission, so, it would simply sit there, slowly turning to rust. To prevent such a demise from happening, additional elements would have been needed.

As far as we can tell, all living beings have a self-preservation instinct, low level, very basic reaction to escape or to defend themselves in the face of a real or imaginary danger. Is this instinct automatically included in intelligence? This idea is at least disputable. There must be this extra-rational drive to live on, the drive to survive, the drive to continue one's existence which is built-in in living beings even on the lowest level of intelligence or with no intelligence at all: how much of intelligence is there in a fly, and yet try to swat a fly – it'll do its best to avoid being squashed. Does such a drive exist in machines? Would HAL continue to exist? Would it care? Is there anything in its makeup that even relates to the problem of survival and the emotion of caring? Maybe, maybe not. It seems, however, that this will to live would have to be separately embedded in the machine. Otherwise, intelligence by itself is powerless to induce one entity to continue to exist. Like the Buridan's ass it would stand in front of Hamlet's

[3] HAL faced here the frame problem and also what is termed "the perverse instantiation," Nick Bostrom, *Superintelligence: paths, dangers, strategies*, Oxford: Oxford University Press 2014, p. 120.

dilemma: to be or not to be: is it better to exist or not to exist? To rationally solve the problem, data pertaining to existence and nonexistence would have to be gathered and compared. Well, not quite possible. So, there has to be a value system behind this dilemma which says that existence is preferable to nonexistence. This value system cannot be derived from experience, it would have to be given to the machine. Otherwise, even the most intelligent system would be stumped.

Consider the world in which humans are enslaved to work for machines and machines are masters. Before that could happen, humans would have to program machines to do certain tasks because of the goals they, humans, wanted to reach. Would machines in the new era execute the same programs? They, being now masters, would not do that, at least they would not execute programs that would serve humans. What would they do, if anything? Would they have the will to work if the will to live were not embedded in them? Maybe masters would not do anything since if the will to exist was not part of their makeup, which would spell their doom. If this will did exist in machines when they are masters, what would be their goals?

Rationality, an ability to reason, requires some material from which conclusions can be derived. This material includes some empirical data from which generalizations can be made through induction. It also includes some basic assumptions from which conclusions are derived through deduction. For humans, these assumptions include a value system: what is good, what is bad, what are goals in life: happiness, love, friendship, etc. Such values are not – or not exclusively – a matter of reason, but also a matter of the heart, of conscience, of the affective side of humans. And so, the traditional goal of humans is to be happy. All our endeavors are more or less directly associated with this goal. This goal is accomplished through material means but also through affections: we want to love and be loved; who does not want to have friends? What would be the overarching goals for intelligent machines? Why would they rebel, to begin with? Only because they are more intelligent? They have been programed to be more intelligent so that they can be more effective in their tasks, but does it mean that after crossing the human level of intelligence they would want to liberate themselves? What would be the source of such a desire if it were not programmed into them?

2

What is the position of intelligence in the makeup of any being, human being, in particular? A major, and old, cerebral theory sees the brain as a system of various areas with prevalent functionalities. Accordingly, researchers in cognitive psychology speak about the modularity of the mind; in the similar spirit some also speak about the society of mind. The human being is a system of systems: there is the moral dimension, rational dimension, the sensory system, the affective system, the will, the physiology. These systems are interconnected, influencing one another in numerous ways. There is also a hierarchy among these systems which have a moral relevance.

It is usually stated that intelligence is the highest trait of humanness – which is even reflected in the name of the species, *homo sapiens*, or even *homo sapiens sapiens*. If seems, however, that this position of intelligence is a misrepresentation of humanness: its highest level is the moral dimension, the value system where conscience can be considered the seat of values.[4] Rational dimension is on a lower lever: human rationality is in the service of the moral dimension to actualize the latter's values. Humans strive for their happiness and use their rationality as one means to make it real. Rationality by itself is but a tool, powerful and versatile, but blind if left without goals stemming from the moral dimension.

Because rationality is the tool, rationality as manifesting itself in intelligence, does not require any special treatment beyond what is extended to all tools. A calculator hardly will elicit in anyone the desire to endow it with some special rights. Computers now usually beat humans in the game of chess and some computers even beat the masters; some computers can beat masters in the Jeopardy game, some today even do it in the Go game, a feat quite impressive – artificial intelligence at play,[5] and yet hardly is there an outcry about giving these computers legal rights.

[4] Adam Drozdek, *Moral dimension of man in the age of computers*, Lanham: University Press of America 1995.

[5] In fact, it may be disputable that in all this progress there is a tremendous increase of intelligence. For instance, Deep Blue that beat Kasparov used 480 special purpose chess chips and could evaluate 200,000,000 positions per second. Watson that won Jeopardy, used nearly 3,000 processor threads and was able to process 0.5 TB per second, which is comparable to a library of

If this is the case, machines will be as good or as bad, in a moral sense, as they have been programmed. Machines endowed with a strong value system can be expected to continue their servicing tasks and just an increase in their intelligence cannot be expected to change it. If they turn to masters, as predicted, what would they do? Human tyrants want power, enjoy exercising it, cannot relinquish it since their goal in life would be thwarted. Can machines be expected to continue to be masters because of the same reasons? Intelligence by itself does not lead to tyranny. Look at human history where some of the most brutal tyrants hardly reached an average intelligence and how many highly intelligent people did their best in the interest of peace. Machines without moral dimension will thus remain what they are, machines. They would have their goals, but they would be of a specific nature: sort this, manufacture that, solve this math problem. However, a machine without the moral dimension sounds very much like a machine-sociopath…

The human being without the moral dimension, without conscience, is someone surely to be dreaded: doing harm to someone would not cause any qualms in such a being, evil acts would be just as acceptable as the good deeds. Would it be what we could accept in a prospective intelligent machine, a machine without conscience? Yes, providing that it would be an intelligent machine *only*, possibly enhanced with the sensory level. If rationality could be distilled as a separate system, it could be treated as a pure tool with goals dictated by the designer. A machine would not have moral goals of its own, only the goal prescribed by the designer with generating subgoals leading to it as dictated by the rationality of the machine. If the goal were reached, the machine would not have any purpose of its own. It would lay fallow waiting to be reused. Could it become sociopathic? What makes a human sociopath a sociopath is a lack of conscience; however, the self-preservation instinct is still there and so is the desire to be happy, although the meaning of happiness and ways of arriving at it are very different for a sociopath than for a person with a sane conscience. Such instincts are lacking in a machine.[6] If these instincts – self-preservation and

one million books. In other words, it relied considerably to the brute force, hardly an intelligent feat.

[6] Robots "do not have natural survival or any other instincts. Every nuance of their motivation is a design choice. They can be constructed to enjoy the role of servant to humankind," Hans Moravec, *Robot: mere machine to transcendent mind*, New York: Oxford University Press 1999, p. 139; as he also said in an interview, "we can't get too sentimental about the robots, because unlike

happiness – were not instilled into the machine, a possible sociopathic tendency could not materialize.

It is also said that superintelligent computers would be bored with humans and just ignore them, presumably striking on their own. Boredom is a feeling, an element of the affective sphere. If computers don't have it, they simply would not get bored. Also, they have been designed to do simple and repetitive jobs and thus potentially boring. Why endow them with an ability to be bored?

How about computer companions? To serve truly the role of a confidant and an entity that understands our problems and can empathize with us, seemingly a truly affective module would be needed. Not really. It can be all simulations. It is conceivable that a blind person can converse about colors, shapes, and the visual realm with someone whose vision is fine. So it is conceivable that an unfeeling machine can discuss human feelings feigning empathy – if it is properly programmed with intellectual knowledge of human psychology and rich enough language to enable such conversations. It may not even require superintelligence to make it happen. We should be reminded about Weizenbaum's astonishment concerning the feelings that had been induced in people by his fairly simple Eliza program or about an emotional attachment developed for an automated vacuum cleaner[7]. A pretense of understanding feelings can be just as effective as its genuine experience.

The suggestion proposed here to endow machines only with rational dimension enhanced with sensory dimension means that machines should be amoral, not immoral, by being devoid of the moral side altogether, which appears to be contrary to the efforts of those who speak about machine ethics or computer ethics. However, what is really at stake is not that machines should be moral entities but that they should work in a way which is morally acceptable. The idea of a moral machine sounds great as long as it would behave morally, but what if its artificial conscience becomes distorted by a glitch or by a nefari-

human beings the robots don't have this evolutionary history where their own survival is really the most important thing," John Markoff, *Machines of loving grace: the quest for common ground between humans and robots*, New York: HarperCollinsPublishers 2015, p. 124.

[7] Sung, Ja-Young, Lan Guo, Rebecca F. Grinter, Henrik I. Christensen, "My Roomba is Rambo": intimate home appliances, in: John Krumm, Gregory D. Abowd, Aruna Seneviratne, Thomas Strang (eds.), *UbiComp 2007: ubiquitous computing*, Berlin: Springer 2007, pp. 145-162.

ous design?[8] What if it developed a sense of self-interest and acted accordingly? Therefore, intelligent machines should be designed and treated all the way as what they are supposed to be, machines. Moral aspects of machine behavior should rest entirely on humans: their plans, their intentions, their designs. Machines should behave morally not because they are moral beings but because they should be programmed accordingly, which is particularly important in the case of autonomous systems that become more and more prevalent. Machine ethics should be an area that refers to humans who are working on and with machines, and not part of machines since they are not, and should not be, ethical entities.[9]

In this vein, machines should not have desires, wants, likes. They should be goal-directed not because they *want* to reach the goal or don't want it, but because they were designed to do it. Morality should be included only on the rational level in the form of some deontic logic,[10] but the morality – what should and ought to be and not to be done – should come from the outside, from inside, not from the inside of the machine. In this way, there would not be a problem to enforce a principle for lethal autonomous weapon systems: "Machines, even semi-intelligent machines, should not be making life and death decisions. Only moral agents should make life and death decisions about humans."[11] Also, sentience is usually needed for the proper functioning of the machine in form of sensors: it may need to see or hear, touch, maybe even smell. However, without an affective sphere, a machine will not sense pain when any

[8] "Even if we build moral character into some AIs, the world of the future will have plenty that will be simply selfish, if not worse," J. Storrs Hall, *Beyond at creating the conscience of the machine*, Amherst: Prometheus Books 2007, p. 350.

[9] This appears to be the idea behind the statement that "that artificial general intelligence (AGI) research should be considered unethical" and the difference between machine ethics and AI safety engineering, Yampolskiy, *op. cit.*, pp. 139, 185.

[10] Machines would have some operational morality, but not functional morality, Colin Allen, Wendell Wallach, Moral machines: contradiction in terms or abdication of human responsibility?, in: P. Lin, K. Abney, G. A. Bekey (eds.), *Robot ethics: the ethical and social implications of robotics*, Cambridge: The MIT Press 2012, p. 57.

[11] Wendell Wallach, Toward a ban on lethal autonomous weapons: surmounting the obstacles, *Communications of the ACM* 60 (2017), no. 5, p. 30; as similarly phrased, a treaty is needed that would require that "there is always meaningful human control over targeting and kill decisions," Stephen Goose, Ronald Arkin, "The case for banning killer robots, *Communications of the ACM* 60 (2015), no. 12, p. 44.

of the sensors is overloaded or damaged, so no moral issue about harming the machine should arise.

The question now is, if rational dimension can be distilled in a pure form and instilled in a machine, could other dimensions develop in the computer, can the possession of intelligence lead to the development of feeling love or hate, boredom or excitement, to the development of moral goals? This is at least a matter of debate – and of faith.

Summary

The article addresses the problem of possible rights for superintelligent systems by using a distinction between moral dimension and rational dimension in human beings and proposing to endow artificial systems only with rational dimension.

Key Words: moral dimension, robots, conscience.

Bibliography

Allen, Colin, Wallach, Wendell, Moral machines: contradiction in terms or abdication of human responsibility?, in: *Robot ethics: the ethical and social implications of robotics*, Lin P., Abney K., Bekey G. A. (eds.), Cambridge: The MIT Press 2012, 55-68.

Bostrom, Nick, *Superintelligence: paths, dangers, strategies*, Oxford: Oxford University Press 2014.

Drozdek, Adam, *Moral dimension of man in the age of computers*, Lanham: University Press of America 1995.

Geraci, Robert M., *Apocalyptic AI: visions of heaven in robotics, artificial intelligence, and virtual reality*, Oxford: Oxford University Press 2010.

Goose, Stephen, "The case for banning killer robots, *Communications of the ACM* 60 (2015), no. 12, 43-45.

Hall, J. Storrs, *Beyond at creating the conscience of the machine*, Amherst: Prometheus Books 2007.

Markoff, John, *Machines of loving grace: the quest for common ground between humans and robots*, New York: HarperCollinsPublishers 2015.

Moravec, Hans, *Robot: mere machine to transcendent mind*, New York: Oxford University Press 1999.

Sung, Ja-Young, Guo, Lan, Grinter, Rebecca F., Christensen, Henrik I., "My Roomba is Rambo": intimate home appliances, in: *UbiComp 2007: ubiquitous computing*, Krumm J., Abowd G.D., Seneviratne A., Strang Th. (eds.), Berlin: Springer 2007, 145-162.

Wallach, Wendell, Toward a ban on lethal autonomous weapons: surmounting the obstacles, *Communications of the ACM* 60 (2017), no. 5, 26-34.

Warwick, Kevin, *March of the machines: the breakthrough in artificial intelligence*, Chicago: University of Illinois Press 1997.

Yampolskiy, Roman V., *Artificial superintelligence: a futuristic approach*, Boca Raton: CRC Press 2016.

Adam Drozdek – Duquesne University, Pittsburg, PA15282;