Béla Pokol

The layers of being and the questions of robot ethics

Abstract

The paper seeks to analyze the new ethical dilemmas that arise in the social contexts of the robot world. It is based on the theoretical foundation of the ontology of Nicolai Hartmann, which finds the place of ever-increasing artificial intelligence in reality among the layers of being. From this starting point, it examines the summative studies of the robotics analysis already developed in English and looks at their correction that needs to be made in the theory of four-layered human existence in comparison with the analyzes so far.

Keywords: layers of being, robot ethics, artificial intelligence, Nicolai Hartmann

Human existence and the life of human communities are based on the cumulative regularities of the layers of being that are built upon each other through evolution, according to the theses of Nicolai Hartmann's ontology (Hartmann 1962). The accelerated development and increasing use of artificial intelligence (AI) in recent years in this structure directly affects the top layer of the four (physical, biological, spiritual and intellectual) layers of being, increasing its strength to the detriment of the lower ones. And with the later development of artificial intelligence, eventually breaking away from human control and gaining independence, it can be perceived as an evolutionarily created new layer of being. Unlike the three previous evolutionary leaps, however, it would not require all the lower layers of being. Taking into account the robots that are the physical incarnations of AI today, AI only needs the physical layer of being. (Pokol 2017). Against this theoretical backdrop, the analyses in this study seek to explore the emerging moral and related legal dilemmas within the mechanisms of contemporary societies that are increasingly permeated by artificial intelligence, while at the same time considering the extent to which the analytical framework changes when the multi- layered nature of human lives, and thus society, is constantly kept in mind.

1. The preliminary questions of robot ethics

In his study of the etic problems of the robot world, *Keith Abney* identifies three areas to group the problems: 1) the field of requirements and prohibitions for robot makers and programmers (such as medical ethics);

2) secondly, the field of requirements to be programmed into robots, first formulated by Assimov under the heading "Three Laws of Robotics"; and 3) finally, perspectivistically for the future, the question of the moral demands and "human rights" that robots might have at that time in possession of self-awareness emerges (Abney 2011: 35). A common dilemma for all three areas is the choice between the main starting points of moral theory already elaborated in the various moral philosophical schools of the comprehensive moral philosophical communities. One such school can be identified as the *deontological starting point* (the rule is the rule, and these must be followed), for which Kant's moral philosophy is best known, and the polarising opposition school, which considers consideration of the *consequences of action as the basis for moral decision-making*. Finally, thirdly, mention can be made of the *virtue ethics* school, which focuses not on the requirements to be considered in every situation in defining morality (like the two previous schools, albeit of the opposite direction), but on the enduring dispositions of the human personality, more simply, on socialised moral values. Here the person does not ask what the moral rule is in a situation, because in the increasingly complex modern world there are often no clear rules, but how a brave, just, faithful, true man decides (Abney 2011: 37).

Of the three schools, the school of deontology is only possible for robots that are used in the narrowest domain and follow the exact rules without being able to weigh the rules, because all situations can only be calculated and controlled in such a narrow domain, but even here unforeseen situations can arise and steer the robot decision in the wrong direction. For example, it could in principle be fed into the decision algorithm of a combat robot to "never kill a child!". However, in the case of child soldiers in African wars, this would mean a predetermined liquidation of the combat robot (Abney 2011: 42). In the case of general- purpose robots, the deontological approach is completely inapplicable. However, the consequentialist school of moral philosophy, which is also tied to the consideration of individual situations, also seems better only because of its life-like nature. Here the guiding premise is to "increase, not decrease, the happiness of as many people as possible with the chosen decision!" and this is impracticable because it would require the processing of a huge amount of information, most of which could not be done in a timely manner even with the greatest capacity of computer data storage. Keith Abney's position, therefore, is that with respect to the second area of robotic morality (i.e. moral decision premises programmed into the robotic algorithm), there is a mixture of deontology and virtue ethics that gives the best perspective, and a mixture of these can create the best built-in robotic moral version.

According to this, the more abstract moral norms (moral virtues) form the decision framework, and the built-in goals and decision contexts always specify the determinants of the decision chosen by the robot in the given situations:

> "The hybrid approach of hypothetical rather than categorical imperatives (within a deliberately restricted, not universal, frame) coming from virtue ethics appear the best bet for near-term robotic morals (in sense two). [...] The emphasis on being able to perform excellently in a particular role, and the corresponding specifity of the hypothetical imperatives of virtue ethics to the programming goals, restricted contexts, and learning capabilities of non-Kantian autonomous robots, makes virtue ethics a natural choice as the best approach to robot ethics." (Abeny 2011: 51)

The connections between layers of being and morality are touched on indirectly by Abney, where he opposes emotivism, which identifies morality with moral emotions, and the cognitive perception of morality, which opposes it. It shows that if morality is tied to emotions because of the emotivists' viewpoint, then primates with emotions cannot be excluded from morality either, which is absurd:

> "Such views, in additon to being unable to explain why nonhuman animals lack morality, also have struggled to explain the apparent congnitive meaningfulness of ethical claims and especially ethical disagreement. (They also naturally have severe difficulties accounting for the ethics of emotionless robots" (Abney 2001: 46).[1]

In contrast, he sees the position of evolutionary psychology, which emphasises the new decision-making mechanism of human evolution as an explanation for morality, meaning that humans have increasingly also developed a system of cognitive decision-making that reshapes current decisions in such a way that the instinctual-emotional first step of thinking is always followed by a second cognitive consideration, thus correcting the first:

> "Evolutionary psychologyts claim there are not one but two of decison-making systems within most humans. The first is an instictual, emotionally laden system that serves as the default for much human activity, particularly when stressed or under pressure. Many other animals share this noncognitive

[1] It is worth pointing out that thise view held by Hegel in legal philosophy long before the advent of scientific psychology in 1820. See Georg Wilhelm Friedrich Hegel, "Grundlinie der Philosophie des Rechts. (Werke Band 7.) Suhkamp, Frankfurt am Main. 1979. 301. p.

> decision-making system, in which (quite literally) we »know not what we do« – or quite why we do it. […] But this »ghost in the machine« does not exhaust human agency; Libet and others found we also have a »veto« ability that can, after its subconscious initiation, still alter our action, in accord with a decision by a second, conscious cognitice system" (Abney 2011: 46).

Abney recalls the mutually shaping effect of the two overlapping layers and almost recalls Nicolai Hartmann:

> "In humans, this deliberative system overlays the ancentral instictual, emotional (and faster) decision-making system and so reason is quite often trumped by our instictual drives" (Abney 2011: 46).

Having concluded that the upper (cognitive-rational) layer of the two-layered human decision-making mechanism is responsible for moral decision-making, Abney poses the question of whether moral decision-making is in principle possible without a reconfigured, overridden lower layer? After all, the answer to this question also depends on whether a moral decision is possible for robots without an emotional layer. In this question, he then decides exactly the opposite way as Hartmann did earlier. It is very possible – he says–- that a rational decision-making mechanism is sufficient for moral decision-making even without an emotional layer of being:

> "Hence, deliberative system cabable of agency necessary for the existence of morality, and so for moral personhood. But is the ancestral emotional system needed as well? [...] In other words – could (emotionless) robots be moral person? [...] The key to moral responsibility and personhood is the possession of moral agency, which requires the capacity for rational deliberation – but not capacity for functional emotional states, therefore, robots may well qualify" (Abney 2011: 47).

Based on Hartmann, there are two problems with these analyses. On the one hand, given the three layers of being above the physical layer of being, it can be seen as flawed that Abney combines biological stimuli with the determinants of the emotional layer. Already here there is building on each other and transformation, and an instinct of the raw instinctual world is supplemented by the emotions of the higher layer of mental existence. For example, the ferocity of a biological sex drive is informed by feelings of togetherness, not to mention the intellectual-symbolic overrides that still build on it, and

the sex-altering aspects of the sublimated love relationships they produce.[2] That is, it is not a double but a triple decision-making mechanism that must be analytically separated in human decision-making, and in addition to the most basic instinctual reactions and determinants, their emotionally reshaped manifestations are still under more rational considerations at the intellectual level. However, a decision and the instinct that directly determines it, respectively its emotional transformation and its intellectual overwriting, are embedded in the interdependent laws of all three upper layers of being. Thus, human morality in societies of all human civilisation requires, in order to survive as a race, that men and women live together in some form of permanent community in order to have children and be educated. A larger community is necessary for the successful struggle and survival of the struggle with the forces of nature and other groups of people, and within these larger communities they must interact in more or less harmonious relationships to organise common activities. Moral virtues (norms and values) are therefore tailored to and maintained by the laws of the specific physical, biological, spiritual-emotional and intellectual layers of being of humans and their communities, and it is only because of the narrowing of moral theories in recent decades that conscious moral choices have become the focus of moral philosophy. Hegel in the early 1800s or Rudolf von Jhering in the 1870s and then Nicolai Hartmann in the 1920s still saw clearly that each person in his socialisation only takes on the accumulated moral norms and values, virtues of many generations, from which the broader communities are maintained, without which individuals could not be fit to live.

From this follows another problem with Abney's analysis, and that is that moral choice seems to consist only in following norms according to an intellectual-rational calculation, but does not require the lower psycho-emotional layer of being. Moreover, as we have seen, the laws of the biological layer of being and the instinct that imparts this to every human being are important for the moral decision. But also in view of this, it can be said that moral norms, moral virtues exist only in human communities (and thus are socialized in the people of the next generations), because only in this way is possible a lasting and harmonious human existence in human communities defined by all four human layers of being. Thus, if an artificially intelligent being can exist with the spiritual layer of being alone, and at most needs only a physical-mechanical body to have self-consciousness and to exercise conscious activity, or to be able to reproduce itself permanently in time, then

[2] See Luhmann's work, which analyzes this process historically: Niklas Luhmann, "Liebe als Passion: Zur Codierung von Intimität", Suhrkamp, Frankfurt am Main 1994.

the moral norms of human existence based on the biological-psyhic layers of life have no function. The moral norms would mean only external things for such being. Thus, if such a robotic being can constantly rebuild its program and even its hardware with the Deep Learning algorithms – as it does for the most part today – then the erosion of moral norms, which are external and functionless for it, is almost inevitable. That is, although it is possible to program instructions that mimic emotions into robots, and they can still account for the decision-making aspects (prohibitions, decision priorities) required by moral norms in today's robots, which are still essentially under human control, but when they reach some level of self-learning capability, it may be uncertain whether the inference of those norms will remain. In the distant future (but in the case of exponential progress in even twenty to thirty years), it would be wrong to assume, in the case of robots in the robotic world, freed from human control and self-aware, the survival of the norms of the human world in the robotic world.

2. Operational morality, functional morality and full moral personality

To better analyze the moral dilemmas and problems of the robot world, the three-way division used by *Colin Allen* and *Wendell Wallach* in their joint study seems useful. Based on different degrees of decision autonomy, they denote the degree of *operational morality* for robots that can only perform the actions determined by the programmers who created their algorithm and possibly by their specific users, and fully fed into them. On the other hand are those that have reached the level of *functional morality*, and this means that they choose the specific action in each situation based on the information provided by their sensors among the action frames fed into their algorithm. Finally, the most autonomous level of morality is seen in robots that reach the level of *full moral personality* with the cessation of human influence, although this type cannot be considered probable now and in the near future, but later their creation can be assumed:

> "System with very limited autonomy and sensitivity have only 'operational morality', meaning that their moral significance is entirely in the hands of designer and users. As machines become more sophisticated, a kind of 'functional morality' is possible, where the machines themselves have the capacity for assigning and responding to moral challenges. The creators of functional morality in machines face many constraints due to the limits of present technology. This framwork can be compared to the categories of artificial ethical agents described by James Moor (2006: 18) which range from agents

> whose actions have ethical impact (implicit ethical agents) to agents that are explicit ethical reasoners (explicit ethical agents.) As does Moor, we emphasize the near-term development of explicit or functional moral agents. However, we do recognize that, at least in theory, artificial agents might eventually attain genuine moral agency with responsibilities and rights, comparable to those of humans" (Allen–Walach 2011: 57–58).

Without going into the possible criticism of whether it is worthwhile to use the degree of morality for robots that have already been fully defined by programmers under the name of operational morality, the robots of functional morality are really interesting in today's stage of development in the robotic world. Self-driving cars, self-propelled combat robots, and to a lesser extent robotic nurses in elder care and health care facilities that have already achieved this autonomy, have such robots and self-driving cars slowly rolling between us or transporting us (mostly only in Japan and the United States today), and the moral decision problems they raise give practical significance to their analysis. The authors go through the possibilities of choice among the trends in moral theory already seen above, and they see the virtue ethics direction as suitable for creating the functional morality of robots. According to their analysis, the moral values (virtues) fed in this way can give the decision framework, which is clarified by training through neural learning mechanisms, and in this way the more general viewpoints of virtues become practical moral decision factors:

> "The virtue-based conception of morality can be traced to Aristotle. Virtues constitute a hybrid between top-down and bottom-up approaches, in that the virtues themselves can be explicitly described (at least to some reasonable approximation), but their acquisiton as moral character traits seems essentially to be a bottom-up process. Placing this approach in a computational framework, neural network models provided by connectionism seem aspecially well suited for training (ro)bots to distinquish right for wrong" (Allen–Wallach 2011: 59–60).

Structurally, this is broadly analogous to the decisions that people are used to making in their daily lives, which are based on abstract moral reasoning and adapted to particular situations, and which are driven less consciously than with mere moral sense. But with the important difference that because of the lack of consciousness and self-awareness of today's advanced robots, the hybrid determinants subtly tuned by programmers (framing virtues plus their training-concretized memory without consciousness) give the more or

less accepted moral norms of today's human societies for the appropriate or approximate decisions. As for the third version, whether it is really possible to theoretically accept robots according to the degree of full moral personality according to human morality, can only be judged skeptically in the case of a robotic world theoretically beyond human control and having achieved full autonomy, according to the above explanation.

3. The devaluation of the physical-biological environment as a moral problem?

The four-layered human existence and the growing weight of the upper, intellectual layer of being and the devaluation of the lower layers have characterized the human evolution so far, but the increasing adoption by robots of the various kinds of work and environmental perceptions will lead in the future to a major change in human socialization, of which the direction of paying attention to reality and turning the details of the real world into experience for him will be largely changed. In a study, *David Zoller* analyzes the increasingly widespread takeover of the work of humans by robots in terms of how this process deteriorates the perception of everyday reality in human consciousness and how the skills and observation capabilities that still exist today are disappearing. The fact that this can already be observed by anyone is enough to recall the telephone numbers already stored in cell phones and thus largely erased from consciousness, or the spatial orientation information that is disappearing from our consciousness due to GPS, and the partial death of this ability. (A recent brain research also found that in case cab drivers in London could claim that the tiny part of the brain in which a group of brain neurons specialized for this purpose to store the vast amount of information about the streets of London disappeared with the spread of GPS, and this brain sector shifted to another function instead).

Zoller brings this problem closer to moral issues by basing moral decision making on the perception of the whole of reality and, in this case, on the formation of human identity and on the detailed perceptual knowledge acquired by an adult from childhood. If future generations socialize themselves since childhood to be surrounded by robots and to have their immediate environment perceived by robots without performing perceptual activities and tasks instead, they will not only be disenfranchised but also lack detailed knowledge of today's adults. As responsible beings, they also cannot grow up to make moral decisions, in other words, they become childish:

> "My own argument is premised on the way that skill opens up corners of reality, so to speak, that are inaccessible to the unskilled. [...] The maturity or adulthood we earn by adjusting ourselves to the 'real world', of course, has a certain moral and personal appeal: a world of lazy psychological infants is, we might think a worse world on a variety of spectra" (Zoller 2017: 81, 86).

The fact that these realms of reality go beyond our perception, and there comes instead the mechanical information processing of these robots, allows us to adapt in ways that are now unconscious because of this change, and this also shakes our moral identity, Zoller says:

> "Given that automating a skilled activity means agreeing that we will exit some niche of perceptual reality, and maybe exit it forever […]. The more suddenly, broadly, and pervasively we hand our perceptual facility over to the robots, the more likely we will make mistakes and simply 'lose data' that were surprisingly integral to our moral and social lives" (Zoller 2017: 86).

While it must be acknowledged that Zoller, in contrast to previous analyses that focus on robots taking over jobs as a unilateral human facilitation – apart from the already discussed socially negative consequences of unemployment (see, e.g., Ford 2014) – went deeper by looking more closely at the change in human perceptual capacity, it must be criticized for unconsciously placing too much emphasis on the layers of being in the physical-biological environment. Looking at Hartmann's layers of being, this change can be read completely differently. The changes outlined by Zoller do not mean the loss of the perception of the whole reality and the ability to do so, but only the ability to perceive the physical-biological layers of being and to pass them on to robots and software bots. In this way, man's liberated perceptual abilities and brain sectors can be more reconstructed to process information about his spiritual-emotional layer of being and his intellectual layer of being, respectively. His moral decisions will therefore be made in the future with less physical and biological environmental information – these will be shut down by robots in mechanical processes – and these decisions can instead be based more on the information from the spiritual-emotional and rational-intellectual layers of being. The diminishing importance of the two lower layers of being, and instead the greater expansion of the two upper layers of meaning for human existence, can of course significantly reshape the foundations of our moral decisions and the incentives that play a role in them. For example, the implantation of dozens of body sensors and their connection to information bases collected in the clouds, as well as automatic

diagnosis by robots of health software and automatic activation of specific doses of drugs implanted in the body, may make the alarms provided by pain genes in our cells largely obsolete (see Kelly 2016: 34–56). Prenatal genetic engineering therefore makes it possible to minimize this, and the conditions of painless human life may redefine the moral obligations and incentives involved today. Overall, therefore, we do not share Zoller's concerns about moral infantilization.

4. Moral dilemmas and responsibilities in hybrid and networked systems

In a study, the authors *Wulf Loh* and *Janina Loh* examined the issues of moral and legal responsibility that arise in currently developed self-driving cars (Loh–Loh 2017: 35–48). They assume that today's self-driving cars are only at the stage of operational morality, so they do not even achieve functional moral autonomy vis-à-vis their manufacturers and programmers. The authors have taken this position based on a moral decision-making structure developed by *Stephen Darwall*, which is divided into four aspects and aims to separate the aspects of autonomy necessary for moral decision-making. The aspect of autonomy required for the level of overall moral personality is called *personal autonomy*, i.e., the ability to possess and choose between personal values, goals, and ultimate aims in life. *Moral autonomy* is the other aspect, and this means that their values and goals include moral principles and ethical beliefs, and along with these they always consider alternatives when making decisions. These two do not exist in today's robots, and only humans are capable of such autonomy, but the aspect of *rational autonomy* is already available to robots at the level of functional morality. This means that the robot can weigh reasons of different weights when making a decision. Their algorithm can already enable this by incorporating pure abstract decision frameworks – leaving some freedom – in which the weighting between possible decision directions is done in light of specific data constantly recorded by their sensors, and they decide based on that data. Finally, the fourth aspect of autonomy is *decision autonomy,* and this means the robot's ability to make decisions not only by external data – continuously concretizing the built-in framework determinants – but also its internal decision priorities without changing them.

Based on the authors' examples – two types of robots already in use (*Kismer* and *Cog*) – it seems possible to achieve this degree of autonomy based on their self-learning mechanisms, which are integrated into the robot's algorithm and are no longer externally controlled:

> "Cog the first robot that can interact with its surroundings due to its embodiment, might pass as an example of a weak functional responsible agent, since its ability to communicate as well as judgments has been improved over that of Kismer. Even more importantly, Cog's overall autonomy has evolved, since it includes an "unsupervised learning algoritm" (Loh–Loh 2017: 40).

Since the current algorithm for self-driving cars does not yet include such an unsupervised self-learning mechanism, they are only at the level of operational morality, and this moral and legal responsibility lies entirely with their developers (designers, manufacturers, and programmers) and car dealers or owners, and respectively between the occupants of the car.

But even with this level of technology, self-driving cars already surpass humans, leaving them – and especially their programmers – with moral dilemmas not seen in the case of humans in extraordinary and unexpected driving situations. For example, if within the braking distance directly in front of the car a group of children jumps into the road to retrieve a rolled ball, the driver cannot stop or even brake at that average speed, leaving him or her with no moral or legal responsibility in the dire event. But self-driving automation, which can react much faster, may still have to make a decision if it can't stop but crashes into a pillar – potentially seriously injuring the car's occupants – or drives and kills children to avoid doing so. But technical capabilities far beyond humans could create a dozen similar new aspects of moral decision-making for self-driving cars in the future. The authors fo study therefore suggest that a separate ID card will soon be created for owners of self-driving cars, in which the final setting of the car software program, the dilemmas left open by manufacturers, must be decided at the time of purchase, so that moral and legal responsibility for the following can be assumed:

> "Since these dilemma situations do not allow for on-the-fly-decisions, the driver will have to take them beforehand. This means that the driver will have to fill out a moral profile of some sort, maybe in the form of a questionnaire, maybe in the sense of a setup program much as with today's electronic devices. For convenience, it seems plausible that these moral settings can be saved to a sort of electronic identification device, like an electronic key or the driver's smartphone, assuming that issues of data security can be solved" (Loh–Loh 2017: 46).

The development of networked robots and the gradual becoming of "smart objects" (smartphones, smart TVs, etc.) around us have only recently begun,

and as they expand, the Internet of Things (IoT) will become more and more involved in our lives in the future.

Human-robot hybrid systems are thus expanding to include additional aspects, and this creates another set of moral and legal dilemmas. *Adam Henschke* analyzes these in his new study (Henschke 2017: 229–43). Smart things are widely available through multifunctional smartphones, smart televisions, robotic vacuum cleaners, and semi-self-driving automated cars with a variety of sensors, but even in everyday life in much of the world, these other smart things have been developed that are already beyond the research lab stages and have already reached the homes of high-tech users with small-scale production. These, however, as we have already experienced with smartphones, etc., will proliferate in a few years and their mass use raises new moral and legal dilemmas. One example is the smart refrigerator, which contains food with RFID (radio frequency identification) and thus digitally identified quantity, shelf life, etc., and the smart refrigerator constantly reads this data, detects the depletion of each food quantity, and since it is connected to the web-based sales mechanisms of nearby supermarkets on the Internet, it can order food and other household items to be automatically delivered. In Japan's aging society, an increasingly large amount of elderly people can be cared for through the use of care robots, and in fully digitized smart homes, such a robot can also care for helpless elderly people, taking over ordered food deliveries in this way. By observing and communicating with the helpless elderly person entrusted to its care, it can call the family doctor or, if necessary, the hospital by phone if its built-in algorithms make a more serious health problem likely.

This example shows how, in a decade or two, robots will be needed in more and more parts of the world, to solve more and more of the work through omnipotent robots and smartphones that can be used in comprehensive information systems to fulfil their functions. However, this growing indispensability of the Internet of Things also creates new dangers and moral dilemmas compared to simple robots. Adam Henschke points out in his writing that the novelty of the Internet of Things compared to single robots is that the latter mainly raises the problem of physical security and the risks have to be assessed in this dimension. (E.g. a robot hoover recently inflicted serious injuries on an unexpected occupant, but one or two fatal accidents of self-driving Tesla cars can also be cited for this). In contrast, security problems and dangers in the Internet of Things occur in two different dimensions. Here, in addition to physical security, information security issues also play a role, since the aforementioned elderly care robot, which

is connected to the software of hospitals, doctors and other places on the internet, can provide hackers or others with information about data recorded by its built-in camera and other sensors. They can share the continuously collected health data about an elderly caregiver not only to the software of the hospital in charge, but also to those who make malicious intentions and plans. In the same way, our smart TVs with a range of applications can not only fulfil their convenience but, with their built-in cameras and microphones, transmit the entire life of the home to software and information databases that we do not see.

This vulnerability can also lead to a physical vulnerability, such as when a hacked automatic door lock is opened remotely for an intruder by external instructions from smart devices. Or, as has already happened in an elegant beach hotel, the electronic smart locks were blocked by a criminal group from outside the flats, and the hotel guests of the wealthy elite were prisoners until the required ransom was paid. However, Henschke also mentions the possibility of the electronic lock of a billionaire's car being blocked by criminals after he got out and his trapped children in it being released on a sunny day only if he transferred hundreds of thousands (Henschke 2017: 234). Immediately after the incident, the said elegant hotel replaced the electric locks that could be swept from the outside and reinstalled the good old traditional locks. After such an incident, the said billionaire will probably also restrict the internet functions of his car for a while. All this, however, forces choices in moral and legal dilemmas and elections that are broadly worth pondering. In the world of our objects, which is becoming more prevalent in the Internet of Things, the old simple things are already being dropped, and we will not be able to replace the objects we wave into the cloud database at will. Just as we would not give up the internet today despite all the negatíve aspects that vulnerability brings.

One such dilemma of the Internet of Things embedded in networked and comprehensive cloud databases is which of the conflicting requirements of the two types of security – physical security and information security – should be given priority? For example, making the smart home of an elderly person who is barely able to move around fully remotely monitorable by medical centres through cameras and microphones may be important to some extent, but it may also mean exposing the most intimate manifestations of life beyond what is necessary. If the emphasis is on information autonomy and limited observation and transparency, the information that is still needed in rare cases may not be passed on to the care centre, and the elderly care recipient may die. Henschke points out that there are often typical priorities,

and for example, in a smart TV, information security has a higher priority, and for this purpose, we can easily address the constraints here. However, with thousands of self- driving car applications tied to cloud software, we pay more attention to physical security requirements and only secondarily to information security requirements (Henschke 2017: 239).

5. Self-learning, machine learning and responsibility

As mentioned above, the main problem of the future will be the dilemma of self-driving cars that are detached from humans and can no longer be blocked from the outside in certain unexpected situations, when the algorithm of this self-driving car, built on neural self- learning, has already decided autonomously. As this has been the main direction of artificial intelligence development in recent years, it is almost certain that this will not be circumvented in this area either. Therefore, today it is worth taking a closer look at the dilemmas of moral and legal responsibility of robots with a high degree of neural self-learning and their makers, owners and users. This question is addressed in their joint study by *Trevor N. White* and *Seth D. Baum* (White–Baum 2017: 66–79) and by *Shannon Vallor* and *George A. Bekey* (Valor–Bekey 2017: 338–53) analysed from different angles.

Trevor and Baum's study not only considers designers, builders and users, but also takes into account the 'punishment' of the robot itself in the case of advanced robots, which already have a punishment and reward system built into their programming, and repeated punishments and rewards reinforce in their programming the decision directions (positive or negative) regarding the selection of future robot responses. This also integrates punishment/reward into the learning algorithm. When the situation arises in the future, the robot's decisions are encouraged in the right direction, and the robot does not need to have consciousness and self- awareness to do this. This way of reinforcing self-learning through repetition is also acceptable according to the authors:

> "Non-conscious robots could conceivably be punished with some sort of reduced reward or utility as per whatever reward/utility function they might have. Specifically, they could be reprogrammed, deactivated, or destroyed or put into what is known as a 'Box': digital solitary confinement restricting an AI's ability to communicate or function. To make this possible, however, such robots ought to be based (at least in part) on reinforcement learning or similar computing paradigms (expcept ones based on neural network algorithms)" (Trevor–Baum 2017: 71).

The neural learning system, however, is judged by the authors to be such that designers and programmers already lose control over the robot's reaction to a given situation and should therefore be banned from the outset as a potential source of danger, possibly disaster:

> "Designers could be similarly liable for building robots using opaque algorithms, such as neural networks and related deep-learning methods, in which it is difficult to predict whether the robot will cause harm." (ibid.)

In the case of algorithms that allow such opaque robot behaviour, it is no longer enough to prescribe liability after the fact, but the prescriptive prohibition is the appropriate thing to do:

> "Hence, instead of liability, a precautionary approach could be used. This would set a default policy of disallowing any activity with any remote chance of causing catastrophe. In effect, people would be held liable not for causing catastrophe but taking actions that could cause catastrophe" (Trevor–Baum 2017: 74).

If one agrees in principle with the authors on the dangerous character of neural deep learning software mechanisms, it only needs to be reiterated that this is a ban on the main way to develop artificial intelligence and therefore it should be considered unlikely in the light of the powers behind industry, military, etc. Therefore, it seems advisable to look for other paths that try to find other solution without banning neural deep learning.

It should of course be emphasised that neural network learning, which mimics the functioning of the central nervous system, can be controlled by involving external human control before a reality-building effect can be triggered. However, this is increasingly falling short for a number of reasons, and this is analysed by Vallor and Bekey in the study cited earlier. One reason is that the advantage of using artificial intelligence instead of humans, the incredibly fast responsiveness would be lost if retrospective human control were introduced. Moreover, ninety-nine percent of the time, the responses are correct, many times higher than human performance. Moreover, the quality of much slower human control may be questionable, as the robot's decision may be more correct than the superior human decision. The latter happened with IBM Watson's drug diagnosis algorithm, and the unusual cure highlighted by artificial intelligence from millions of oncology studies and diagnoses, and later synthesised by them, proved more correct than the oncology decision it overrides:

> "Watson's diagnoses and treatment plans are still vetoed by licensed oncologists. Still, how reliably can a human expert distinguish between a novel, unexpected treatment recommendation by Watson that might save a patient's life – something that has reportedly already happened in Japan – and the oncological equivalent of 'Toronto'?" (Vallor–Bekey 2017: 343).[3]

The dilemma of losing speed and thus eliminating the robot's advantage is also and not even the weakest competitor would have been lost. So it has become a symbol of wrong decisions made by artificial intelligence, which is rare but causes tragedy in many cases illustrated by the robotic soldiers and decision-making software used in war situations. Here, the question constantly arises whether the robot soldier entering the most dangerous area and building can use the destructive weapons in its possession to destroy those inside without an external human decision, or instead the order of destruction may only be given by remote human confirmation. In the same way, the dilemma arises as to whether a reconnaissance aircraft could be destroyed immediately by the robotic aircraft, or this could only be done with human intervention from the remote command room based on the information it transmits? The constraint of speed requires the robot itself to make and execute a decision, as the loss of time from an external human decision could lead to robot destruction if it broke into a dangerous location. But shooting down friendly fighting machines that have already happened several times, or killing children and women misidentified as enemies in the invaded area, argues against this (Vallor–Bekey 2017: 349).

Detailed neural network learning algorithms, the latest trend in artificial intelligence, already provide self-learning software with multiple depths for the simple computational starting position by continuously feeding in billions and billions of data, thus gradually making the starting position the most advanced. In this technique, between the inputs of the self-learning software and the task-specialised outputs, thousands of intermediate neural layers are found between the massive data, independently finding patterns and regularities and highlighting them for use. By combining billions of pieces of data, they can highlight and use the smallest regularities that are imperceptible to humans when making decisions:

[3] "Toronto Mistake" was one of Watson's fundamental mistakes in a nationwide television quiz when he beat everyone with his answers to the toughest questions. As a final mistake, he made Toronto one of the US cities and not even the weakest competitor would have been lost. So it has become a symbol of wrong decisions made by artificial intelligence, which is rare but causes tragedy in many cases.

> "Between the input and output node layers are »hidde«, layers of nodes that function to process the input data, for example, by extracting features that are especially relevant to the desired outputs. Connections between the nodes have numerical 'weights' that can be modified with the help of a learning algorithm; the algorithm allows the network to be 'trained' with each new input pattern until the network is optimised. [...] The interest in neural network has grown in recent years with the additon of more hidden layers giving depth to such network, as well as feeedback or recurrent layers. The adjusment of the connections strengths in these more complex networks belongs to a loosely defined group of techniques known as deep learning" (Vallor–Bekey 2017: 341).

The effects of decision patterns highlighted by these detailed learning algorithms, while often achieving surprisingly good results in practice, may not be understood by designers and programmers, and their decisions may consistently cause surprises, including varying degrees of unpleasant surprises. Who should bear the legal and moral responsibility for this?

6. Identity in the world of artificial intelligence

James DiGiovanna raises thought-provoking questions when he asks in his study how the identity of people with brain implants may change. He discusses this issue together with the question of the identity of robots, which in the future may appear as fully artificial beings and already have self-awareness (DiGiovanna 2017: 307–21). Let us consider the two problem areas separately.

The possibility of memory augmented by brain implants has been developed in recent years in mouse experiments and has been shown to be effective. All hope to mitigate and cure the effects of rapidly spreading Alzheimer's disease in ageing societies (see Kaku 2014: 132–33). DiGiovanna is exploring the possibility of other developments in the coming years, in addition to diseases spreading en masse, to increase brain capacity. And if a technical solution is found to the problems that remain in this field today, it is almost certain that this will become commonplace, first among the elite, then in society as a whole, to increase the greatest value of human intelligence. This means, however, that the permanent identity of each individual, which is the basis for contacts in communities, may be more or less annulled and it may become uncertain how much we can expect our partners to survive their qualities that we have known and loved so far:

> "The ability to rewrite mental content such as ethical values, the capacity for emphaty, and general personality traits undermines personhood. [...] A para-person that could experiment with worldviews, completely adopting and deleting values systems, preferences, and bases for judgement, would be largely lacking in what is commonly understood as the most basic element of personal identity" (DiGiovanna, 2017: 311).

This was the basis of our choice in the case of our friends, wife and girlfriend, but in the same way, our closer human relationship with some of our work colleagues is based on the love of their traits, while the relationship with others is only coldly collegial. Therefore, our lives in society and in various small communities within our society are based on our permanent identities, and this can change fundamentally after the addition of heart, hearing and other physical enhancements when the brain is changed with brain implant.

With the gradual changes in one's life, one's consciousness is always being rebuilt in detail, and this leads to small gradual changes in one's identity, which in the modern world are intensified by the information expansions of the last century. In comparison, however, in the future we will be able to plant a whole range of information – the contents of books and studies, smaller libraries – with brain implants in our heads and together with it handle new basic logical and value-processing mechanisms that we did not have in our lives before, and we did not have the skills to do so. Now, this will fundamentally affect the contact between the individual and his communities. In any case, the bases of contact based on the present permanent identities could be eliminated by this change. After such a new content of consciousness – especially if the values of the contact partners have been supplemented and reclassified – I cannot know to what extent my boyfriend, my girlfriend, my wife, my colleague, etc. are the same. Whether the qualities we have loved in them so far are still alive, or in the same way the experiences we have had together so far, which provided the same response in our close relationship even without words, are still relevant to him. This can only be exacerbated by the possibility that the brain implants in the brain, which complement our knowledge, norms and logical abilities acquired with the help of our biological brain, are constantly being updated anew and anew from the outside, as we already know today. Moreover, they can constantly connect to the information bases of their software stored in the clouds. To what extent will our friend, equipped with such, remain familiar, on whom we can rely, because "yet we know him!"?!

This question of identity also extends to legal and moral problems. To what extent can I respect someone for their past behaviour or simply despise

them because after a brain refreshment they can either be a "moral athlete" or just a cold advantage-seeker. Or does legal responsibility for yesterday's action make sense for someone who has since thought and acted differently? The other side of this is whether, if we can change the consciousness of a sociopath by brain implant and their consciousness is partially erased and a new socially friendly consciousness is introduced, is a system of punishment still necessary? And this raises the question of whether, in addition to voluntary brain implant, its forced installation is acceptable? Or, in part, can it be made compulsory by the state for all children to be screened and tested in childhood, as is currently the case with compulsory vaccinations? DiGiovanna calls para-persons the future humans with such augmented brains – avoiding the name cyborg, which has already been invented for them in science fiction – and given the current state of laboratory research, this future does not mean a distant future at all, and the probability of its realisation is high. Dealing with the legal and moral dilemmas and adapting today's solutions to the situation of the time therefore require extensive consideration.

Beyond para-persons, in the case of fully artificial and, unlike today, self-aware robotic beings, the likelihood of which cannot be ruled out, even if it is not as great as the former, the question of identity can be addressed by raising new aspects. DiGiovanna places the content of identity at the centre in order to expose the dilemma of identity in relation to robotic beings. Some details of humans and their consciousness are constantly changing, but their enduring characteristics and value preferences change only slightly, even over many years, so that those who always live in their environment can more or less be ascribed an identity that embraces change. It is the slowness of change that enables me, even in today's fast-moving world, not to be disappointed by my previous experiences with the motivations and characteristics of those who come into contact with me. But that is exactly what disappears for robots, which are thousands and millions of times faster than humans at processing information and learning themselves in a very short time:

> "Slow change of character and appearance is part of what makes personal identity [...] But with an artificial person, sudden and radical change in both the physical and mental becomes possible" (DiGIovanna 2017: 311, 307).

Permanent value preferences in information processing and cooperation based on them are already problematic with robots due to the mass and speed of their information acquisition as well as their constant self-learning and self-changing. DiGovanna's proposal may also mean that the possibility of "self-awareness" and "ego-awareness" need to be reconsidered even for a future powerful MI

robot. These presuppose the permanent identity of a person, but this relies on the slowness of our changes in consciousness and thus the permanence of our information processing. When an artificial being is freed from human control and switched to independent information processing and from that to self-learning and self-transformation, it can learn thousands of times a day, every hour and even every minute, and can transform itself in its ever-shorter new cycles, then what we call a stable self-consciousness, ego- consciousness, in contemporary humans almost disappears. With this emphasis, DiGiovanna also adds a new question to the much discussed question, i.e. how the question of self-consciousness and ego-consciousness of the future advanced robot consciousness will stand. And how can one imagine moral value without permanent self-consciousness and ego- consciousness?

For this reason alone, the train of thought calculated by the mechanical extension of the current human image that such a robotic being will probably also be "super-ethical" in the case of superintelligence must be considered wrong (see Petersen 2017). In this context, however, it is also necessary to address more comprehensively the explanations and analyses that, in the case of the development of robots with their own consciousness – by human analogy – provide for the recognition of their moral needs and the granting of human rights in their writings. For these analyses conceive of future robots simply as a new kind of human companion and an extension of human existence. Once the robots' programmes have incorporated emotions into their algorithms, these analyses demand that society pay attention to robots' emotions and grant them human rights as well:

> "It probably needs to be legislated how much pain and danger a robot can be exposed to. [...] It could easily be that this would lead to further ethical debates about other rights of robots. Can robots own property? What happens if someone is accidentally injured? Can they be sued or punished? Who is responsible for them if they are sued? Can a robot own another robot? Such questions then give rise to another question: should robots be given an ethical purpose?" (Kaku 2014: 251).

Our previous explanations answered several questions from these, based on the robotics studies conducted in the intervening period, but the basic problem behind them should also be highlighted, as whole studies and volumes have emerged from similar assumptions, e.g. a new volume in this area edited by *Jason P. Doherty*: "AI Civil Rights: Addressing Artificial Intelligence and Robot Rights."

Now, the basic problem with this line of thought is that it ignores the fact that rights and ethical requirements can only arise in robots when ego-

consciousness and self-consciousness are created. But it also means that if this really happens in the future, they will simultaneously be freed from human control by the thousandfold development and built up as a separate new layer of being above the previous four layers of being of human societies. From that time on, however, they would be indifferent to the whole biological sphere and the human societies connected with it and would not need "judicial protection". That is, a robotic world that reached this level would not be part of human society as a "new comrade" in dominion over the world, but as human existence emerged from the primate world and rose above the animal-biological layer of being and became more and more autonomous, so now the artificial machine intellect, detached from biological preconditions, rises above human society. In contrast to the previous construction of ever newer layers of being on the lower layers of being, the new layer of being of artificial intelligence would only need the lowest physical layer of being, and for it the biological and psychological-emotional layers of being would not be necessary. These robotic beings would not need rights and ethical demands, but they will dominate the whole reality, including human societies, as we humans dominate the four-layered earthly world today. In this way, those analyses are rahter right that discuss whether, if the robot world really reaches this level, what will happen to humanity?!

7. The moral credo of "Unabomber", the "mad mathematician"

In the mid-1990s, after many years of a series of bombings and an FBI chase, a secret perpetrator called "*Unabomber*" gave the reason for his actions in a one-and-a-half-hundred- page pamphlet that he spoke out against the inhumanity of the development of technological society since the Industrial Revolution. His peculiar language was recognised by his brother and by notifying the FBI, the long-suffering bomber was captured. It turned out to be *Theodor John Kaczynski*, a mathematician from Harvard. At one point in his university career, he became the enemy of a society dominated by technology and began his series of explosions, targeting the developers and major users of that technology. Several died and more were wounded in the process, and he planned to retaliate even more if they had not been arrested.

Now that the exponentially evolving impact of technology development over the last thirty years has become truly indisputable, and the scale and impact of its further acceleration has already been the subject of several comprehensive analyses, it is worth refocusing on arguments made by Unabomber, the "mad mathematician". This is what *Jai Galliott* does in his new study. He places the resistance fighter, who has since been busy developing his theses in his prison cell, among theorists and movements of

antitechnology, and an attempt is made to highlight his main theses in light of the current state of the robot world (Galliott 2017: 369–85).

Kaczynski has only drawn the practical conclusions of the earlier theses of *Jacques Ellul*'s 1964 volume *The Technological Society*, which in their own way were also a continuation of *Osvald Spengler*'s 1922 work analysing the decline of Western civilisation. Both authors explained the decline in terms of technological development (Spengler, 1995). The purely pessimistic and resigned tone in Splengler and Ellul then became a moral resistance in the case of Kaczynski, and after seeing that there was no way to reform this development, he believed that only revolutionary violence remained to prevent the destruction of humanity. Decades after his pamphlet, it is now worth considering how the current state of the robotic world and the more radical changes that are already largely visible could mean the endangerment of humanity, or at least a significant deterioration in its condition.

As a starting point for their approach to technological society, it is worth highlighting that both Splengler and Ellul and Kaczynski view human existence as embedded in the physical-biological environment. From this they conclude that human existence is destroyed when, as a result of the industrial revolution, human life becomes more and more technologically mediated and in this way more and more distant from the physical-biological environment:

> "Ellul wrote that the machine trends not only to create a new human environment, but also to modify man's very essence and that the millieu in which he lives is no longer his. He must adopt himself, as though the world were new, to a universe for which he was not created. Kaczynski shares this sentiment" (Galliott 2017: 373).

On the other hand, if we take into account Hartmann's thesis, which keeps in mind the four interdependent layers of being of human life (physical, biological, mental and intellectual) and which assumes in the course of evolution the ever stronger transforming effect of the upper layers on the lower layers, the above thesis is exaggerated and without reason and it must be classified as too pessimistic.

Kaczynski and his predecessors see it as a decay of human life when the top intellectual layer of the four strata of being becomes ever more dominant over the lower ones. However, this has been the case, albeit more slowly, over the last two or three thousand years and one can single out the use of metals and especially iron, from which the transformation of the human environment was fundamentally altered. The industrial revolution has only accelerated this, and especially since the 1950s it has become tumultuous to base the various

activities of human communities on intelligence and the technology associated with it. That is, human life is by no means based only on the physical-biological layers of being. So when their share and decisive power in human life diminish and this environment is widely mediated and transformed technologically, it does not mean that human society is destroyed. In all this, only the weight of importance of the four layers of being in reality shifts, making human life more based on the intellectual layer and radically increasing the dominance of this spiritual layer of being over the lower one. This assessment of ours could only be suspended if at some point in the evolution of the robotic world this world were to truly emerge from human control and artificial intelligence were to rise as a new layer of being above the human societies that had hitherto been at the peak of evolution. The elevation of Kaczynksi as a prophet would then only be prevented by the fact that under such circumstances and its dangers, the failure to be elevated as a hero would be the least of the problems. However, to the best of our knowledge and belief, this can only be considered an unlikely option today, and rather the growth of human societies characterised by artificial intelligence without a new autonomous layer of being can be considered a realistic vision for the future.

Bibliography

Abney, Keith 2011. "Robotics, Ethical Theory and Metaethics: A Guide for the Perplexed. In: Patrick Lin – Keith Abney –George A. Bekey (eds.): *Robotethics.* The MIT Press, Cambridge Massachusets. London. pp. 35–54.

Allen, Collin – Wendell Wallach 2011. "Moral Machines: Contradiction in Term or Abdication of Human Responsibility?". In: Patrick Lin – Keith Abney – George A. Bekey (eds.): *Robotethics.* The MIT Press, Cambridge Massachusets. London. pp. 55–68.

DiGiovanna, James 2017. "Artificial Identity". In: Patrick Lin – Ryan Jenkins – Keith Abney (eds.): *Robot Ethics 2.0.* Oxford University Press. New York. pp. 307–21.

Doherty, Jason P. (ed.) 2016. AI Civil Rights: Addressing Artificial Intelligence and Robot Rights. Kindle Edition. Ford, Martin, The Rise of Robots: Technology and the Threat of a Jobless Future. Basic Books.

Galliott, Jai 2017. "The Unabomber on Robots". In: Patrick Lin – Ryan Jenkins – Keith Abney (eds.), *Robot Ethics 2.0.* Oxford University Press. New York. pp. 369–85.

Hartmann, Nicolai 1962. *Das Problem des geistigen Seins. Zur Grundlegung der Geschichtsphilosophie und der Geisteswissenschaften.* Walter de Gruyter Verlag. Berlin.

Hegel, Georg Wilhelm Friedrich 1979. *Grundlinie der Philosophie des Rechts.* (Werke Band 7.) Suhkamp, Frankfurt am Main. S. 301.).

Henschke, Adam 2017. "The Internet of Things and Dual Layers of Ethical Concern" In: Patrick Lin – Ryan Jenkins – Keith Abney (eds.): *Robot Ethics 2.0.* Oxford

University Press. New York. pp. 229–43.
https://doi.org/10.1093/oso/9780190652951.003.0015
Kaku, Michio 2014. *The future of the mind. The Scietific quest to understand, enhance and empower the mind.* Doubleday. New York.
Kelly, Kevin 2014. *The Inevitable: Understanding the 12 Technological Forces That Shape Our Future.* Penguin Books.
Klinewicz, Michal 2017. "Challenges to Engineering Moral Reasoners" In: Patrick Lin – Ryan Jenkins – Keith Abney (eds.): *Robot Ethics 2.0.* Oxford University Press, New York. pp. 244–57.
Loh, Wulf – Janina Loh 2017. Autonomy and Responsibility in Hybrid System. In: Patrick Lin – Ryan Jenkins, Keith Abney (eds.): *Robot Ethics 2.0.* Oxford University Press, New York. pp. 35–50.
https://doi.org/10.1093/oso/9780190652951.003.0003
Luhmann, Niklas 1994. *Liebe als Passion: Zur Codierung von Intimität, Suhrkamp.* Frankfurt am Main.
Pokol, Béla 2013. *Theoretische Soziologie und Rechtstheorie.* Kritik und Korrigierung der Theorie von Niklas Luhmann. Passau. Schenk Verlag.
Pokol Béla 2018. *Künstliche Intelligenz: Die Entstehung einer neuen Seinschcht?* (Ki – im Spiegel von Nicolai Hartmanns Ontologie.) Pázmány Law Working Papers No.2018/12. https://doi.org/10.2139/ssrn.3225107
Splengler, Osvald 2007. *Der Ubntergang des Abendlandes.* Albatros Verlag/Patmos Verlag.
Talbot, Brian – Ryan Jenkins – Duncan Purves 2017. "When Robots Should Do the Wrong Thing" In: Patrick Lin – Ryan Jenkins – Keith Abney (eds.): *Robot Ethics 2.0.* Oxford University Press. New York. pp. 258–73.
Vallor, Shannon – George A. Bekey 2017. "Artificial Intelligfnce and the Ethics of Self-Learning Robots". In: Patrick Lin – Ryan Jenkins – Keith Abney (eds.): *Robot Ethics 2.0.* Oxford University Press. New York. pp. 338–53.
https://doi.org/10.1093/oso/9780190652951.003.0022
White, Trevor N. – Seth D. Baum 2017. "Liability for Present and Future Robotics Technology". In: Patrick Lin – Ryan Jenkins – Keith Abney (eds.): *Robot Ethics 2.0.* Oxford University Press. New York. pp. 66–79.
https://doi.org/10.1093/oso/9780190652951.003.0005
Zoller, David 2017. "Skilled Perception, Authenticity, and the Case Against Automation". In: Patrick Lin – Ryan Jenkins – Keith Abney (eds.): *Robot Ethics 2.0.* Oxford University Press. New York. pp. 80–92.
https://doi.org/10.1093/oso/9780190652951.003.0006

Béla Pokol
Professor Emeritus
University Eötvös Lóránd
E-mail: belapokol1@gmail.com
https://orcid.org/0000-0002-1170-1764