УДК 007.52+004.8+172

МОЖЕТ ЛИ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ СТАТЬ ЧЛЕНОМ ОБЩЕСТВА КАК АВТОНОМНАЯ ЛИЧНОСТЬ?

А. МЕЦ[1)]

[1)]*Тартуский университет, ул. Юликооли, 18, 50090, г. Тарту, Эстония*

Отмечается, что в результате стремительного развития искусственного интеллекта и роботостроения скоро появятся искусственные человекоподобные существа. Поднимаются вопросы о наделении их моральным статусом и юридическими правами. Предполагается, что роботы и искусственный интеллект (мозг робота) могут быть автономными агентами, личностями. Эта возможность рассматривается через сравнение машинного и человеческого мышления на разных уровнях, преимущественно с опорой на научную материалистическую аргументацию. Во-первых, анализируется существенное отличие в том, как люди и машины распознают материальные объекты и ориентируются среди них. Обосновывается, что эта способность лучше развита у машин, но при этом они не могут формировать понятия, как это делают люди. Делается вывод, что машинам не хватает базовых пространственно-временных знаний. Во-вторых, изучается социальное мышление, поскольку рассматривается возможность для роботов стать автономными членами общества, и отмечается, что некоторые аспекты социального мышления у них достаточно развиты (например, интерактивная речь). Однако некоторыми авторами оспаривается тезис о том, что автономия на самом деле необходима роботам. Также обсуждаются нейронные и феноменологические основы самости, сознания и личности, чтобы выявить некоторые дальнейшие фундаментальные вопросы, связанные с возможностью управления роботами. На основании проведенного исследования утверждается, что неорганическое существо не может быть личностью в смысле полноценной социальной субъектности.

Ключевые слова: искусственный интеллект; роботы; нейронаука; когнитивная наука; сознание; личность; философия искусственного интеллекта.

Благодарность. Исследование проведено при финансовой поддержке Тартуского университета (грант PHVFI20930), Эстонского исследовательского совета (грант PRG462) и Европейского фонда регионального развития. Идеи, высказанные в статье, впервые были озвучены на конференции «Философия и экономика в эпоху цифровой трансформации», проходившей на базе Белорусского государственного экономического университета 15 декабря 2020 г. Автор выражает благодарность А. А. Головач за приглашение выступить на этой конференции и Д. Г. Добророднему за предложение опубликовать статью в настоящем журнале, а также Х. Хосейнпур, Н. А. Х. Абделазем Мухаммед, Дж. Б. Смит, Т. Луик и М. Раштипур за полезные обсуждения по философии искусственного интеллекта, которые легли в основу настоящей статьи.

Образец цитирования:
Мец А. Может ли искусственный интеллект стать членом общества как автономная личность? *Журнал Белорусского государственного университета*. *Философия. Психология*. 2022;1:32–41 (на англ.).

For citation:
Mets A. Can artificial intelligence become a member of the society as an autonomous personality? *Journal of the Belarusian State University. Philosophy and Psychology.* 2022; 1:32–41.

Автор:
Аве Мец – кандидат философских наук; научный сотрудник Института философии и семиотики.

Author:
Ave Mets, PhD (philosophy); researcher at the Institute of Philosophy and Semiotics.
avemets@ut.ee
http://orcid.org/0000-0003-3105-1313

CAN ARTIFICIAL INTELLIGENCE BECOME A MEMBER OF THE SOCIETY AS AN AUTONOMOUS PERSONALITY?

A. METS[a]

[a]*University of Tartu, 18 Ülikooli Street, Tartu 50090, Estonia*

The development of artificial intelligence and robotics is proceeding so rapidly that many philosophers and technologists believe them to soon become human-like beings, and consequently consider attribution of moral and legal rights to them. Such attribution presupposes that robots and artificial intelligence (robot's brain) can be autonomous agents, persons. This article discusses this possibility by comparisons of machine and human cognition on different levels, with a primarily materialist-scientific argumentation. Firstly, how humans and machines recognise and navigate the sheer material world of objects differs in essential ways. Although this is the cognition most developed in machines, they cannot form concepts like humans do, thus they lack basic spatio-temporal knowledge. Secondly, social cognition is considered, since this is the context for robots as autonomous members of the society, and some aspects of this are fairly developed in them (like interactive speech). Some authors' discussions hint that autonomy is not really desired from robots. The third part discusses the neural and phenomenal foundations of self, consciousness and personality, to bring out some further fundamental issues with the possibility of robot agency. It will be concluded that a non-organic being cannot be a locus of personality necessary for social subjectivity.

Keywords*:* artificial intelligence; robots; neuroscience; cognitive science; consciousness; personality; philosophy of artificial intelligence.

Acknowledgements. The work was supported by the University of Tartu (grant PHVFI20930), the Estonian Research Council grant (PRG462), the European Regional Development Fund (Centre of Excellence in Estonian Studies). The ideas expressed in this paper were first formulated for the conference «Философия и экономика в эпоху цифровой трансформации» (15 December 2020; Belarus State Economic University, Minsk). The author thanks A. A. Golovach for the invitation to present at that conference, D. G. Dabrarodni for discussion and invitation to contribute to the journal and H. Hosseinpour, N. A. H. Abdelazim Mohamed, J. B. Smith, T. Luik and M. Rashtipour for discussions on the philosophy of artificial intelligence that have informed and inspired this paper.

Introduction

I. Asimov [1] formulated three laws for robotics in 1942, the second of which reads as follows: «A robot must obey the orders given it by human beings except where such orders would conflict with the first law (i. e. cause harm to human by action or inaction – *A. M.*)», implying robots' adherence to human superiority. However, robots and artificial intelligence (AI) – the brain of robots – seem to be evolving so rapidly that philosophers and technologists-engineers are seriously discussing their status in society as its autonomous members. A widespread optimism concerning the capabilities of AI and robots, both in utopian and dystopian (or neutral) keys, is seeping from science fiction into futuristic and philosophical literature [2–7]. There is certainly some backing from real life too, such as robots performing human-like functions as a member of the board of an organisation, as a family member, as a conversation partner, and others [2], strongly impinging on the human viewer an impression of real intelligence and personality. Thus many philosophers argue that robots might have to be assigned the same moral (and legal) rights and agency, or personhood, as humans are assigned [2].

This is particularly poignant in possible cases of (moral) conflict, for instance, when having to choose between a human's rights and a robot's rights. If a robot is assigned personhood, autonomy and rights, then in such conflicts it may be preferred to the human's rights, causing genuine suffering to the latter whereas the robot's suffering would not be genuine if it has to defer to human. The argument for robot rights often draws from the fact of past extensions of rights from free men only, to women, children and slaves, to animals, to other kinds of species and environment in general [2]. An obvious counterargument is that there is clear ontological difference between living beings and machines (see also [8]), that will be discussed in detail below.

Although D. J. Gunkel [2] discusses the option that AI could deserve rights without an ability to have them, mostly it seems that this ability is assumed implicitly. The ground for this may be that robots come to resemble humans or other living beings, they seem to have certain personal traits that are similar to human ones, such as autonomy and consciousness. This contradicts Asimov's second law which precludes robot's own initiative. Contrary to this, we do not expect that a human person always obeys all the orders that are given to them. We assume that a person is autonomous to a certain extent and they have the right to be autonomous and to initiate their own actions. Can a robot be autonomous, refuse to comply with a human order? What are the features of autonomy and is it possible to build a robot with such features, having personhood? Autonomous action in society also presupposes ethical

rules, their knowledge, understanding and adherence to them. Can artificial intelligence understand morality? Can it be conscious?

To answer the above questions, I will look at some aspects of how robots and artificial intelligence work and how they function differently from humans. This will significantly be a materialist argumentation, including cognitive neurological and other natural scientific accounts to inform the posed philosophical questions. The first section discusses the cognition of the physical world – the world of space and objects – since this is the basis and environment of any action. It is concretely determinable by its external, measurable and numerically modellable traits; and it is most implemented and thus furthest advanced in machines by now, giving an idea of how an AI could navigate our world. This will inform the next section about the social world, learning about and coming to terms with it. In the third section, I will consider some arguments about the possibility of authenticity of artificial minds.

Cognition of the physical world

Some aspects that cognition includes are physical perception, processing of the percept into an «image» (meant also about non-visual percepts), acting upon it (thereby gaining additional percepts of its object), contextualisation and expression, memory and recollection. The European Commission High-Level Expert Group on AI (henceforth – Commission) includes similar aspects also in the definition of AI: perception of «the environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take to achieve the given goal» [8].

Importantly, the Commission includes under AI software and possibly also hardware which act in the physical or digital dimension, meaning that AI and robots are taken on equal terms. For an important reason I consider it necessary to differentiate between robots and AI in some contexts. If (general) AI is taken to be analogous to human, as it often is, then human is thereby, perhaps unconsciously, reduced to their intelligence, the latter (again unconsciously) equated to the brain, or more accurately – cerebral cortex – in a vat: like we feed a computer with codes and data which it uses to compute a result, the cerebral cortex would analogously be fed with percepts which it is to process by its supposedly innate codes. It is true that new-born babies have some very abstract unconscious innate ideas: distinction between visual and audible, a kind of notion of number, the notion of object that takes a certain space, moves in continual manner, and can not occupy the same place as another object, etc. [9–11]. However, human is not only their cerebral cortex and not even only their brain: besides this central nervous system, we have peripheral nervous system and the whole rest of the body besides. And human intellect or mind works not disconnectedly from the rest [11–13]. Those new-borns' ideas listed above are like hypotheses that will be put to test through experiments with the external world, learning that things are weighty and fall if unbalanced, causes and effects of things, telling things' nature from creatures' nature, etc., and the development of a person's intelligence depends on the richness of its experiences with the world as a child [9]. We have those many senses to explore and interact with the physical world: besides vision and hearing we have touch, taste and smell, proprio- and nociception (body position and pain), etc., which give us input about the world, both the environment and our own body. We also have our inner life – thoughts, feelings and emotions, influencing each other and the learning process. Importantly, we do not learn by mere passive glancing at the world but by practice, that is, bodily activity related to objects and situations, and this has a fundamental role in the development of human mind and intellect. Many of those senses and their enabled percepts and learning modes are such that an AI as a software can not be provided with their analogues but an appropriately built robot can, especially motion, relating to proprioception and handling of objects, and change of position and location and thus perspective on, and shifting of, the surrounding space and its markers.

Let us consider a simple example of cognition. When a person is familiar with one object, he is able to recognise other objects similar to it; already small children have flexibility to counter new situations [9]. For example, they see a car standing or moving, walk around it, get inside and take a lifts etc., acquiring different views on it. Thereupon they will recognise other vehicles among the objects that they encounter. An AI needs to «see» millions of photographs with cars of different colours, shapes, taken from different perspectives and distances, so that its «knowledge» about machines would not depend on a small set of these properties. And still they will not recognise a toppled vehicle's underside in a car accident on the road as what it is but perhaps as a dog instead. Yet this, along with recognising regularly situated vehicles and persons, is very important for autonomous vehicles which have to recognise the traffic conditions in order to decide its tactics of moving from their current position towards the prescribed destination.

Take another example – an artificial photograph of a non-existent person. Everything looks very truthful, but the glasses' earpiece is directed towards the middle of the ear, not to its upper end where the temple should rest on the ear. The machine is not familiar with the temples and ears as objects and how they are connected; it only recognises patterns. Those patterns were obtained from photos of persons fed to the machine. A human has the ideas educated with experience with

objects not merely visually, but bodily and conceptually; even when they themselves wear no glasses, they understand how the earpieces of glasses function, namely that they have something fundamentally physical to do with ears. Humans build concepts upon their cognition and can thereupon extrapolate their knowledge to new situations.

Those were examples of AI failing in conceptualising and contextualising objects. Yet there are important differences already on the sensory and processing level which undoubtedly conditions the level discussed. On a phenomenological level, a human person *feels* the properties of objects qualitatively. The various qualities – visual, audible, tactile, etc. – are linked to the object or situation perceived and participate in the formation of a single unified concept to which they belong. In a computer, all «knowledge» enters in discretised numerical form and is stored as data – vectors of zeros and ones [14], it does not feel, does not «understand» the quality or the object as such, and does not form concepts. As seen in the above example, the machine merely «perceives» patterns which it educes from the data vectors if it has an appropriate code for doing so [15]. Also, due to concepts formed through comprehension and experience, a person recognises untruth and errors (including in data), while a computer does not. Certain learning algorithms, such as neural networks, simulate the exclusion of single erroneous data by assigning them appropriately low weights, but if errors abound, they also would include them as part of the data on the conceptual level [15].

The two «information processing devices» themselves work quite differently. The brain is constantly at work, its billions of neurons firing at once; which of the processes becomes conscious is decided randomly, by noise that enhances certain synapses to fill the workspace («an internal system [the brain], detached from the outside world, that allows us to freely entertain our private mental images and to spread them across the mind's vast array of specialised processes» [5, p. 151]) [10; 16], all this being carried out and influenced by the various chemicals in the brain. A computer, in comparison, has very clear and unambiguous calculation processes: on the «sensory» level of cognition, signal versus thermal noise is determined by a threshold for the voltages determining 0 and 1 [5]. Human memory is plastic, reinventing itself, reconstructing and reinterpreting, even confabulating, past events and knowledge [10; 16]. Computer memory is fixed: an item once stored there can become defective, but not be reshaped into a different form [12]. The human cognitive extrapolation even works on the perceptual level, since the brain fills in the parts of the «image» which are not perceived due to the specificities of human perceptual organs (for instance, the retina has a blind spot, and is two-dimensional, so the spotless, three-dimensional image is inferred by the brain; [10; 16]). For analogous compensation, the AI would need additional sensing devices or special code deducting its measurement apparatus's inherent idiosyncrasies and errors.

There are robots who do have to learn mechanically about objects (including walls) and their positions, such as the robotic vacuum cleaners and self-driving cars. For this recognition they wear lidars and radars (aside with cameras, possibly), or they just bump into objects as obstacles to be circumvented. However, the measurement results obtained from those sensors provide numerical coordinates and dimensions of objects, not a concept of object per se. A robot geared with limbs and necessary sensors like gyroscopes and limb position measures can also cognise and manipulate its own «body» – robots have been taught to navigate obstacles, walk, flip in the air, etc. Such sensorimotor tasks are indeed one of the easiest to implement in computing machines, being realisable with clearly numerically definable conditions and the said sensors [17]. In humans, some aspects of intelligence, such as basic visual, verbal, and sensorimotor, need early stimulation to form, to build up the necessary synaptic networks in the brain [10; 12]. Later stimulation may not bring any result at all, or the result is very different than early obtaining both on the neural and phenomenal level, like with learning a second language. There is no such difference making in computers.

Navigating the social world

Objects are, despite the pitfalls described, relatively easy to learn. But this is only a physical-mechanical orienteering in the world. For an AI (robot) to be a member of the society, it must understand the social world – recognise social situations and decide how to behave in them. It must know morality and law. How could a robot apprehend what the situation is and in which way it requires a moral assessment and decision? One might think that in a social decision-making situation, utilitarianism is the easiest to apply, since it relies on calculation, and calculation is what AI does. The robot should somehow know, firstly, the moral value of the objects (including living objects) and their relations that it encounters, and, secondly, the various consequences of possible actions and the values of those consequences. Those values must inescapably be expressed in numerical terms, since that is the only language that a computational machine understands (a machine can, of course, process unstructured data like text, but it does not understand it). But, as W. Wallach and C. Allen argue [18], in real life there are so many consequences and such a variety that it is unrealistic to calculate them, even for an AI. In addition, there are often important consequences in terms of social values that can not be objectively measured, or even modelled, due to their abstraction and complexity, and lack of direct relation to

measurable properties, for instance friendship, justice, or a sense of security. Some of those could perhaps be assigned some relative, conditional numerical values, possibly on the basis of the particular society or community where the robot is to operate, which would count as a type of non-exact, pragmatic measurement. However, meaningfulness of such attribution of numbers would be highly doubtful.

A person has moral rules and values that govern their behaviour. Could such rules be programmed into AI? For example, «don't lie». To do this, the AI must know the truth and its relation to what it would testify in the format of text, speech, numbers, photographs, videos, etc. For a person, as explained above, these different types of representation of the world are combined in practice as concepts. For a machine, however, they are rows of zeros and ones that do not have to coincide and be comparable at all. To make them correlate with each other, the machine must be specially taught to relate the patterns of one format with the patterns of another format, and even then it may not deduce any detailed, causal relations between them (for instance a motion of a scythe in a video format should be one-to-one accompanied by a certain bounded whirring sound; instead, it only associates the set of this type of motions with the set of that type of sounds). Relating to text format, one and the same idea or image can be described in many different ways, that human would, but a machine may not, understand to describe the same thing. A machine would also not necessarily recognise if the words make up a coherent and truth-value-capable text at all (for instance, 'Colourless green ideas sleep furiously' would make a legitimate sentence).

And consider a situation in which one would have to lie: in Nazi Germany, people hide Jews in their homes and the Gestapo comes and asks: «are there any Jews here?» Could AI know that in such a situation the rule «don't lie» does not apply, as doesn't Asimov's second law? More mundane situations can be thought of where a true or accurate response is not required or adequate, for instance such questions as 'Do I look ok?' or 'Did it taste good?' (the latter even inapplicable to a machine «person»). Such are fine nuances and distinctions of ethics, morality, and etiquette, and generally in social-cultural interaction, which are even demanding to learn for a human person; an AI, a straightforward number cruncher, can not make such distinctions (see also [19; 20] about the blindness of the big data machines to the reality, and its neglect, behind numbers).

The robot does what it is commanded according to how it is built and programmed. If it does not, this is considered a technical defect. In principle, it is possible to program a robot to disobey certain orders, but then it is the engineer who decides which orders to obey and which not, and the robot is not autonomous. It is also possible to make disobedience independent of the will of the engineer: for instance, when the robot realises that it was given an order and that it is able to perform as ordered, then a random function is activated which decides whether to obey the order or not. Yet, human disobedience, their free will, is not random, but deliberated, dependent on previous experience [5]. A person has their own thoughts, feelings, preferences, goals, values, based on which they decide to refuse to comply with an order that is contrary to their goals and values. Is a robot capable of acquiring, for example, through machine learning, any understanding of morality and personality, or will such a possibility become too narrow, causing overfitting (over-specialisation to narrow conditions [15]) and an inability to extrapolate data?

A person's goals and values are linked to how they are «composed». They are a biological creature in need of air, water, food, sleep, etc.; they feel these needs bodily. They are a mental – intellectual and emotional – being; they need self-awareness, self-development, security, companionship, etc. (many of such needs also felt bodily). They can suffer and be ill physically and mentally. These circumstances constitute an important basis for their values and behaviour, which are aimed at, or comprise, one's (and possibly others') bodily and mental self-preservation, integrity and advancement. If a robot is to learn behaviour on the basis of its composition, which would be adequate for its needs and necessary for its autonomy, the result will differ markedly from human behaviour. It becomes «aware» of itself and its needs via built-in sensors, but the corresponding «knowledge» must be programmed in it. For example, he will know when his battery needs to be charged; if it doesn't act upon the reaching of the respective threshold in time, its battery will die and it loses the possibility to signal the need of recharging. Current computers do this and other rudimentary introspection («disk space, memory integrity, or internal conflicts» [5, p. 238]). But even if it is not charged, the robot will not suffer any such ill effects as pain or distress from this, as a person does due to lack of food and sleep or other adverse conditions. And those machine needs are still only material and mechanical. It can be built and programmed to express intellectual and emotional traits (ability to speak, grimaces), but this is a simulation and again made by engineers.

From a functionalist perspective, S. Dehaene [5] considers computers' introspection to yet miss «three critical functions» that distinguish them from humans: flexible communication, plasticity and autonomy. Flexible communication means that one program's output becomes input for the entire organism at all times (it enters the workspace). This must be understood to include both sensory as well as rational-social input, or better: percepts interpreted in physically and socially relevant and appropriate ways. Plasticity means the system adapts to the input information and its environment and itself with the help of a brain-like learning algorithm. Contemporarily, AIs have the risk to overfit to a set of training data, disabling it to adequately process

new incoming information. Autonomy means that it has «its own value system to decide which data are worthy of slow conscious examination in the global workspace. Spontaneous activity would constantly let random "thoughts" enter the workspace, where they would be retained or rejected depending on the adequacy to the organism's basic goals. Even in the absence of inputs, a serial stream of fluctuating internal states would arise» [5, p. 239]. Without human intervention, the machine should be able to set its own goals. This would allegedly lead to an artificial consciousness.

As members of human society, robots would be expected to pay heed to human values and behavioural norms, not to harm humans and the rest of the environing world. Obviously, its goals should not counteract human existence. It is in no position to extrapolate this from his own state and needs. If it has the proper «organs» for perceiving and processing information – sensors such as a camera, microphone, etc., and learning codes (for example, neural networks) – it could in principle learn patterns of behaviour. There are caveats, however. Firstly, it would again need a huge sample of regular examples to come to perceive some kind of regularity; it can not learn from small samples and extrapolate [9]. Moreover, as with the glasses' temple and ear, and the scythe and whirring, so it will be with the causes of behaviour: the machine has not even a concept of causality, not to mention the ins and outs of a specific behaviour, its connections with human nature and social norms. To a certain extent, perhaps, they could be programmed with a description of their environment's needs and be attached appropriate sensors, for example: a living person must be in an atmosphere which consists of 21 % oxygen, 78 % nitrogen, etc., supported by solid ground, excluding professional swimmers during certain motions of swimming, etc. It already transpires that the number of such rules and their exceptions exceeds the possibilities of implementation, since, in principle, there are infinitely many possible situations, not to mention the socially significant situations, pregnant with cultural meanings, whose «mechanical» configuration, but not the moral and social significance, may be accessible to a sentient machine.

Nevertheless, many engineers are optimistic about robots achieving a general intelligence and exceeding humans with their abilities, thus possibly rendering humans obsolete. Why would anyone need such robots – what were the aim of creating them? J. Agar [21] discerns three AI categories according to their purposes (A, B and C): A – advanced automation (aims: practical and technological); C – computer based central nervous system or «computer based studies of the central nervous system» (aims: fundamental, biological); B – bridge and building robots; later basic research in AI (intelligence theory). B should bridge the two sides A and C, or implement a model of the nervous system in practice, automate it. C will be scrutinised in the next section; here let us consider some social context of those purposes.

Thus far, both robots and AI are tools used for performing narrow tasks – category A, such as assembling cars, helping social workers with heavy duty (like lifting disabled people), planting or searching for mines on a battle field, sex services, etc., with the former; heavy and complex computational labour (big data based questions) like interpreting stacks of files or data for detecting features in organisms, landscapes, natural language, «human resource», etc., with the latter. Moral issues have arisen in both, for instance the infamous AI Twitter account Tay who learned racist slurs, or autonomous cars which have to take prompt decisions in difficult situations (see also [19; 20] about tools as less autonomous AI systems, which could in principle become parts of general AI's cognitive systems). But why we would need a robotically embodied general AI who may be detrimental to human life? Besides the technocratic curiosity to try out what human being is capable of in terms of divine creation, one answer proposed is to offer companionship. It would be a replacement of human companionship for people with impaired sociability, for instance, or in pursuit of an ideal companion [6; 7]. This is an age-old desire already reflected in folk songs as «making a soul out of copper», to have a perfect life partner.

The robot as a companion would adapt to its human's character by learning from them or being programmed for them, and because we choose our companions according to our character. With real people, who have their own character, aims, and life story at least partly independently of their cohabitants, this character stays and will play a part in forming their relationships, including causing frictions. The main aim of choosing an artificial companion is that it is, or can be made, perfect, hence it must not cause frictions, and hence it is not allowed to have character traits unpleasant to their human companion. If they are made more human-like, with their weaknesses, to render them more relatable [7], those would be chosen such as are more tolerable to their human companion. Thus they are not a truly autonomous general AI. Furthermore, required to perfectly align with the human, they are not a true other, since they are there for the human convenience only, for offering possibly constant emotional labour [6], thus they would still be mere tools.

Authenticity of mind and self

According to D. C. Dennett [4], differentiating AI from natural (human) intelligence leans on the Cartesian dualism of mind and body (which he denounces): since a machine is a mere matter, has no soul or mind, it can not be compared to human. This distinction, however, need not at all lean on dualism but can be drawn from pure materialism (P. L. Núñez [22] even reckons dualism to be consistent with materialism). Hu-

mans and computers consist of very different materials which can not function in exactly the same ways, in contrast to what some authors have claimed. The claim has been that it does not matter whether we build life out of carbon or of silicon [5; 23], they are both in the same group in the periodic table, closely related, and possess to some extent similar properties. However, there is a good reason why such a huge and ever expanding branch of chemistry as organic chemistry or carbon chemistry (and further, biochemistry) exists. Carbon is a very special chemical element, with quite distinct properties. Whereas most elements manifest varying valencies depending on which other substances and in which conditions they react, then carbon has a fairly constant valency of 4. Due to this and to the stability of bonds between its own atoms, it is able to form exceptionally stable catenations of a variety of lengths, and cyclic (aromatic) compounds, with widely varying structures, properties and capacities [24]. Those are the compounds that compose and run living nature, including the brain.

S. Dehaene [5] strongly believes that computers will be able to simulate the brain to the extent that artificial consciousness, a sentient robot, will be possible. His claim is founded on his own *in silico* experiments concerning unconscious and conscious perception, where even the constant unconscious activity of the brain (see also [16]) could be simulated, and the random emergence of conscious processes out of the unconscious ones [5]. This is relevant because intelligence is associated with what human is conscious or aware of, and it only manifests in conscious beings. Our conscious subjective experiences are understood to build up our personalities and life-world, feeding into our autonomy and morality; and our on-going activity depends on what and how we are conscious of. On a theoretical level, S. Dehaene defines consciousness as a process: «…consciousness reduces to what the workspace does: it makes relevant information globally accessible and flexibly broadcasts it to a variety of brain systems» [5, p. 154]. On an operational-technical level, consciousness has two signatures: firstly, «the sudden activation» of the «anatomical network of interconnected high-level areas, involving primarily the prefrontal and parietal lobes»; and secondly, the P3 wave, «a large positive voltage that peaks at the top of the scalp», due to «many more neurons [being] inhibited than… activated (during conscious perception – *S. D.*), all these positive voltages end up forming a large wave on the head» [5, p. 157, 165]. This means that the different apparatuses – functional magnetic resonance imaging, electroencephalograph and magnetoencephalography interact with the brain's magnetic and electric properties and translate them into certain kinds of images (of maps, lines, dots), conveying this particular information about the brain activity.

Is the mind or consciousness really mere (electromagnetic) information (flow)? What is information? S. Dehaene takes it to be what the apparatuses of reading brains give us signs of: electrical signatures of brain activity, read out by those apparatuses. According to information theory [14], which is a mathematical-logical theory, information is both physical, being encoded by a signal, and abstract, carrying a message. But just as no imaging method provides the full account of the brain, each only showing some aspect of it [25], so each physical means provides just some aspects of the message to be conveyed. There is no isomorphism between the physical and the abstract facets; for instance, when speech is written down, the script is like a model of the speech, conveying the abstract message, but not many other aspects, such as the variety of sounds (different pronunciations of the same letter), rhythm and pace, tone and timbre of the speaker's voice, etc., which provide other kinds of information about the physical situation. The computational message carrier conveying the brain's activity only captures some aspects of it, the voltages or frequencies. Although the computer itself functions due to voltage differences which move the electrons carrying the message, this can not ground the comparison, since this should then also ground comparisons to any other electrical device, or even the grid itself. In the brain the part of the picture concerning voltage differences inducing flux is analogous [12; 26], but the carriers of the charge – and of the message – are chemical: ions, amino acids, proteins and others – hundreds of different neurotransmitters. The various neurochemicals have many tasks in the brain and body: determining feelings and emotions, regulating daily rhythms of wake, sleep, nourishment, etc. This can not inhere in a computer. The computer can only encode an impoverished picture of all that is going on in the brain (and body) (see also [22]).

Computer-simulating the brain is like simulating any other real world system, i. e. an experiment *in silico*. Any action of the ions and proteins must be simulated by code; and those actions, and those of the different parts of the brain, are diverse, participating in varying combinations in different effects [12]. Even if consciousness is indeed successfully simulated, as S. Dehaene reports, can it be a general AI? Will it be intelligent, for instance, in the sense as the narrow artificial intelligence is intelligent now (which some thinkers do not even consider as intelligence, e. g. [9]), by doing what is its task – and it would have a general set of tasks – immeasurably more efficiently than human would do it? Would such an ability rise from its simulating human consciousness; that is, would it be the bridging (category B) AI described by J. Agar [21]? Or will it merely imitate, on computational basis, being a human being, doing things in the same way as humans do, thinking, remembering, feeling and perceiving, sleep-dreaming and exploring like humans do?

When for instance chemical reactions or weather or climate are simulated on a computer, we do not say

that computer becomes the system it simulates, it does not become those chemical substances and their interaction, or weather or climate. It is still only a computational machine which helps us assess the models we build about real world systems and make predictions by calculations and simulations on the basis of those models. Why then should we say that it becomes conscious(ness) when it simulates consciousness, or intelligent when it simulates intelligence? It is a model that captures some (informational) aspects of the modelled part of the world, but the world itself is more than the particular set of interactions it has with this set of apparatus: it is the matter it «consists of» and all its multifarious properties and interactions with the world, including those that can not be uniquely and unambiguously measured, modelled and calculated. When a computer, by the simulation, yields readings similar to the signature of consciousness or some process of the brain, this is a mathematical-numerical similarity, analogous to implementing one and the same mathematical function in descriptions of very different real world systems (whose measurement is then also expected to yield similar patterns accordingly), in which case we do not say that those systems are ontologically similar because of their numerical similarity.

In general, intelligence has many different aspects, some of which can be more easily computationally simulated or implemented (logical-mathematical, bodily-kinaesthetic), others less (visual-spatial, interpersonal, discussed in previous sections, and linguistic), and some not at all (creative, intrapersonal) [17]. Those mostly concern the conscious experiences meant in the above discussion, although many of them are learned and trained so thoroughly by humans that they need not put any conscious effort into achieving them. The last mentioned – intrapersonal intelligence – is what concerns mind and self, human as a moral and reflexive personality, most intimately. Let us consider two further points in this regard, that I would call self-consciousness and holism of personality.

The described notion of consciousness as manifested by the P3 wave is only one of many different notions of consciousness discerned and discussed in cognitive and neuroscientific literature [27; 28]. N. Block [27] calls the P3 wave-related aspect, where a stimulus reaches the global workspace, access-consciousness, and discerns it from phenomenality, in which case a stimulus may not reach the global workspace (but can nonetheless affect a person's decisions even remaining unconscious [5; 22]) and reflexive consciousness, which means ability to reflect upon some details of what was perceived. However, a distinction important from the perspective of mind and personality is consciousness of one's self, versus consciousness of something external. Although those are philosophically not clearly separable, since we are always already situated with respect to the external world and learn about ourselves in and through this situatedness environmentally and socially, usually one perceives oneself as lying under those perceptions and gathering them as one persistent subject. Thus A. Morin [28] presents various models of levels of consciousness, in which the levels constituting the self are self-awareness (focussing attention on self; processing private and public self-information) and meta-self-awareness (being aware that one is self-aware). Even if those aspects of self could be neurally describable and modellable, for instance, perhaps, via the homunculus (the parts in the brain mapping the different body parts), or mirror neurons via which we mirror the other person (as the meta-self-awareness includes awareness of other people being aware of oneself), or some other configurations, the phenomenological facet of the self is in itself intimate and inherently subjective.

D. J. Chalmers [23] writes about subjective experience as the particular sensations accessible only to the first person, such as the brightness of this blue colour or the intensity of this feeling of pain that one may be experiencing at a certain time. (I surmise this problem can be extended to purely mental internal world of a person too, such as reminiscences or ideations, as causing subjective experience.) He discerns the problem of subjectivity as a hard problem from the «easy» problems of physical counterpart of perception, studied by cognitive and neurosciences, easy due to their objective material observability and measurability (see also [22]). Yet, D. J. Chalmers still deems information in the information-theoretic sense to possibly be a fundamental concept for a corresponding theory (which would open it for information-technological exploration), contradicting his own contention that subjective consciousness is a basic, irreducible term. This would presuppose comparability of those purely subjective experiences, since information is an abstract, objective, digitisable, hence countable-measurable notion. Subjectivity must somehow be expressed as symbols carrying messages that can be read by some entity external to the issuer of those messages. This is contradiction in terms, since what was originally meant was exactly the subjective, non-externalisable, hence non-comparable aspect of consciousness (comparison premises the possibility to juxtapose the objects to be compared). This can again only yield knowledge about the same aspects of consciousness that cognitive sciences already study. Even if some informational facet of the subjective consciousness, that is, measurable either with the brain imaging technologies or psychological questionnaires, renderable by a scientific model, is identical between two persons, they are still separate organisms and separate persons (see also [12; 29] about impossibility of identity of persons). There is a fundamental gap between persons in a sense in which another person will always remain transcendent to oneself. No such gap exists between computers, who can exchange information without loss.

The notion of subjective experience has been criticised [5] as spurious, since brain images can show us mental states, which, according to this approach, is the subjective experience. However, we also know that one and the same stimulus, which presumably translates as one and the same pattern of brain activity in different people, can call forth very different reactions and feelings of pleasure or distaste or indifference. This «how something feels the way it does to me» is the sort of subjectivity that we may not be able to convey transpersonally and technologically.

Holism is meant in both the above mentioned sense that a human being is not a mere brain but includes the rest of the body, as well as as beings with sensed spatio-temporal finitude, extension and situatedness, with their life stories and memories (also T. Viik [6] underlines this). It does not necessarily mean integrity of personality, whose obvious exceptions are certain mental disorders; it must also be kept in mind that divergences occur to even the spatio-temporal and bodily holism, often accompanying certain brain damages (e. g. stroke; see [29]). Influenced by feelings and emotions (realised by the neurotransmitters), the brain evolves with the activities practiced and experiences gained [11–12], starting with the bodily and emotional parts of the brain, and working towards the more abstract parts realising meanings, associations, generalisations and reasoning [12]. Those experiences form and continuously shape both the person's life story (a diachronic aspect) as well as their repertoire of states of mind, the different aspects of intelligence and spirituality (a synchronic aspect). This essentially includes conceptual or symbolic (linguistic) cognition of oneself and one's surroundings, which is fundamentally social and normative [13; 28]. S. Greenfield [12] even suggests that the self *is* mind, as opposed to emotions, since emotions are linked to the connectedness to the external world (neurologically: domination of local brain circuits), and hence to disconnectedness from one's own mind (the large neuron assemblies; see also [29]). However, under personality we usually also include one's preferences with respect to, and reactions to, sensory stimuli, such as tastes, sounds, colours, etc. Greenfield's self may be construed as in opposition to the surroundings, to which our senses and attending to them connects us, manifesting one's sensual personality versus intellectual personality.

Particularly this intimate intellectual phenomenological aspect, the existential mind, the apprehension of temporality and fragility of human life and experience, childhood and mortality [6], is what we consider most human. Although computers can also break down, software outdates and hardware wears out, this is not comparable to human fragility. Software and hardware can, with little effort, be replaced with their equals and the machine will work just the same as before. Human mind and organism, if broken, take a lot of time to heal and may retain scars, or irredeemably succumb to fate. Human life stories are intimate parts of them (unless they are severe amnesiacs), their dispositions, creativity and moods inspire them to initiate short and long term projects involving themselves, others and their closer and farther surroundings. Computers have no stories of their «life» and development, childhood memories or hopes and expectations for «life». If an AI does confabulate something out of what they have learned from a set of informational units, this is not intimate to them. They are not selves. And those mental, spiritual, emotional selves are important aspects in which AI could not fully understand a human being and human society.

Conclusions and perspectives

The types of perception and processing of information in humans and robots differ both qualitatively and quantitatively. A robot or AI can receive data (not perceive or feel properties) and find regularities or surface patterns in them, and it is able to do so quickly and in huge quantities. But it has no comprehension of concepts, such as thingness and causality. This basic perceptual difference gives clues about more complex aspects of being human: self, personality and sociality, which include dynamic memory and apprehension, autonomous sociality and morality. An AI can be made to imitate various facets of those human being and doings, even creating artwork, but those are mere simulations of isolated phenomena, not aspects of one and the same self in the sense as a human person has various bodily, intellectual, spiritual, etc., aspects. It takes a specific kind of organic being to give rise to such diverse sets of manifestations of mind and consciousness.

The preceding discussion is necessarily of narrow scope both in the sense that many human characteristics, as well as exceptions to human functioning in a society could not be taken into account. For instance, mostly normal, average neuro-typical human persons were given as comparison to AI, whereas there are many different conditions which render human experience very much different from this, to which I could only hint cursorily. Many such cases are described in [11; 22; 29; 30]. For instance, whereas most people can walk without conscious effort and do something else all the while, like converse or enjoy a scenery, then people who have lost proprioception or some other precondition for normal motion need to fully concentrate on walking, this activity requiring their full conscious energy. Another comparison that I could not even hint at are moral digressions such as sociopathy and others, meaning that not only machines but also human beings can fail to function as social and moral beings and severely disrupt society. Even just normal people are not always socially perfect. Also the various ways in which an individual is dependent on the society and hence not fully autonomous could not be discussed.

Another branching theme I could not delve into is the various definitions of both intelligence and consciousness. For instance, an alternative understanding of intelligence encompasses a being's (an organism's) ability to adapt to external conditions and survive (and thrive). Obviously organisms may find themselves in adverse conditions called «danger» which threaten their lives or integrity, which is not yet a reason to deem them unintelligent. So with this qualification in mind, machines could also be considered intelligent for the conditions for which they have been created. However, this applies to any tool which serves its purpose well enough. Different kinds of consciousness, or a spectrum thereof, could be thought of as applying to other organisms besides human, with different kinds of nervous system (possibly lacking a central nervous system altogether) and analogously extended to machines. However, those are still organic beings, not mere electrical devices. But if consciousness presupposes organicity, or minimally at least nerve cells, there is some progress in connecting them to technology. Examples are electronic implants, electronically working prosthetic limbs reacting to the wearer's neural signals, and scientific experiments with single neurons linked to electrodes. This, however, is yet far from a human-like mind, which emerges with a hundred billion neurons. This is yet unrealisable in laboratory conditions.

Even if robots will not resemble humans in any substantial way, this does not warrant a legitimacy of bad behaviour in their presence. When or if they become sufficiently good learners, they may be able to learn behavioural patterns as normal both between living beings, as well as towards them. Treating them with respect may therefore be necessary for a purely precautionary cause, so they would learn respectful treatment of other beings, but also as a virtue ethical exercise for their human companions.

References

1. Asimov I. Runaround. *Astounding Science Fiction*. 1942;29(1):94–103.
2. Gunkel DJ. *Robot Rights*. Cambridge: London: MIT Press; 2018. 256 p.
3. Alexandre L. *La guerre des intelligences. Intelligence artificielle versus intelligence humaine*. Paris: JC Lattès; 2017. 250 p.
4. Dennett DC. *From bacteria to bach and back. The evolution of minds*. New York: W. W. Norton & Company; 2017. 496 p.
5. Dehaene S. *Consciousness and the brain: deciphering how the brain codes our thoughts*. New York: Viking Press; 2014. 346 p.
6. Viik T. Falling in love with robots: a phenomenological study of experiencing technological alterities. *Paladyn Journal of Behavioral Robotics*. 2020;11(1):52–65. DOI: 10.1515/pjbr-2020-0005.
7. Coeckelbergh M. Artificial companions: empathy and vulnerability mirroring in human-robot relations. *Studies in Ethics, Law, and Technology*. 2010;4(3):1–17. DOI: 10.2202/1941-6008.1126.
8. Kerikmäe T, Müürsepp P, Pihl HM, Hamulak O, Kocharyan H. Legal person- or agenthood of artificial intelligence technologies. *Acta Baltica Historiae et Philosophiae Scientiarum*. 2020;8(2):73–92. DOI: 10.11590/abhps.2020.2.05.
9. Dehaene S, Yann Le Cun, Girardon J. *La plus belle histoire de l'intelligence*. Paris: Robert Laffont; 2019. 288 p.
10. Vishton PM. *Understanding the secrets of human perception*. Chantilly: The Great Courses; 2011. 488 p.
11. Wang S. *Neuroscience of everyday life*. Chantilly: The Great Courses; 2010. 608 p.
12. Greenfield S. *The private life of the brain*. London: Penguin Books; 2002. 258 p.
13. Miller DL. *George Herbert Mead. Self, language, and the world*. Austin: University of Texas Press; 1973. 280 p.
14. Schumacher B. *The science of information from language to black holes*. Chantilly: The Great Courses; 2015. 362 p.
15. Mueller JP, Massaron L. *Deep learning for dummies*. Hoboken: John Wiley and Sons Inc.; 2019. 368 p.
16. Viskontas I. *Brain myths exploded. Lessons from neuroscience*. Chantilly: The Great Courses; 2017. 232 p.
17. Mueller JP, Massaron L. *Artificial intelligence for dummies*. Hoboken: John Wiley and Sons Inc., 2018. 339 p.
18. Wallach W, Allen C. *Moral machines. Teaching robots right from wrong*. New York: Oxford University Press; 2009. 275 p.
19. Zweig K. *Ein Algorithmus hat kein Taktgefühl. Wo künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können.* München: Heyne Verlag; 2019. 320 p.
20. O'Neil C. *Weapons of math destruction. How big data increases inequality and threatens democracy*. New York: Crown Publishers; 2016. 272p.
21. Agar J. What is science for? The Lighthill report on artificial intelligence reinterpreted. *British Journal of the History of Science*. 2020;53(3):289–310. DOI: 10.1017/S0007087420000230.
22. Núñez PL. *The new science of consciousness: exploring the complexity of brain, mind, and self*. Amherst: Prometheus Books; 2016. 350 p.
23. Chalmers DJ. The puzzle of conscious experience. *Scientific American Special Edition*. 2002;12(1):90–100.
24. Cotton FA, Wilkinson G, Murillo CA, Bochmann M. *Advanced inorganic chemistry*. New York: John Wiley and Sons Inc.; 1999. 1376 p.
25. Giere RN. *Scientific perspectivism*. Chicago: University of Chicago Press; 2006. 160 p.
26. Heller HC. *Secrets of sleep science: from dreams to disorders*. Chantilly: The Great Courses; 2013. 462 p.
27. Block N. Paradox and cross purposes in recent work on consciousness. *Cognition*. 2001;79(1–2):197–219. DOI: 10.1016/S0010-0277(00)00129-3.
28. Morin A. Levels of consciousness and self-awareness: a comparison and integration of various neurocognitive views. *Consciousness and Cognition*. 2006;15(2):358–371. DOI: 10.1016/j.concog.2005.09.006.
29. Bolte Taylor J. *My stroke of insight: a brain scientist's personal journey*. New York: Viking; 2008. 192 p.
30. Sacks O. *The man who mistook his wife for a hat*. London: Picador; 1985. 233 p.

Received by editorial board 28.06.2021.