Ethics in Progress (ISSN 2084-9257). Vol. 9 (2018). No. 1,
Art. #3, pp. 44-61. Doi: 10.14746/eip.2018.1.3

Why Can´t We Regard Robots As People?

André Schmiljun

ORCID: https://orcid.org/0000-0002-9415-8495
(Humboldt-Universität zu Berlin, schmiljun@insystems.de)

1. Introduction

Can a robot[1] be a person[2]? Although this question seems weird, it is not far-fetched: in spring 2018 at "I/O conference" Google relaunched a software named "Duplex", a voice system, that sounds like a natural person, being able to arrange appointments on the telephone for its clients like booking a table in a restaurant or a hair cut in a hair salon. "Duplex" deployed human-sounding vocal cues, such as 'ums' and 'ahs' — to make the conversational experience more comfortable.[3] For humans on the other end of the line, the pantomime AI was apparently very realistic.[4] Due to applications such as this, some philosophers and scientists like Petersen, Sparrow[5], Veruggio, Abney are already

1 It is important to differentiate at the beginning between the two notions Artificial Intelligence (AI) and robot. Following the Oxford English Dictionary, AI involves the project to develop computer systems that are able to perform tasks, normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages. A robot, on the other hand, is a machine performing certain human functions and movements automatically.

2 Compare (Schmiljun 2017a, 75) in order to read about the complex philosophical debate concerning the term person and how it is differentiated from notions like personality, people or personhood.

3 Source: https://techcrunch.com/2018/05/10/duplex-shows-google-failing-at-ethical-and-creative-ai-design/, July 21th 2018.

4 Of course, such an AI-based chat-box is limited to its defined application, due to it can be easily exposed when a spontaneous unexpected question comes up during conversation, for instance, if the waitress on the line suddenly changes the topic.

5 Sparrow also belongs to the optimistic group of robot scientists that believes robots will be regarded as people. In comparison to Petersen, Veruggio and Abney, he argues that robots will become people when they pass the Turing Triage Test, a test that "sets out a necessary and sufficient condition for granting moral standing to artificial intelligence. Machines will be people when we can't let them die without facing the same moral dilemma that we would when thinking about letting a human being die." (Sparrow 2014, 305-307). Further he adds:
"(…) [M]oral standing of machines cannot be divorced from the question of the proper conditions of application of the only concepts that we posses that might allow us to recognize 'machine persons'. Any conclusion that we wish to draw about whether or not machines might be persons or what

convinced that it is only a matter of time until the first robot will be developed that can be regarded as a person, probably also inspired by science fictional movies like *Star Trek*, *I Robot* (2004) or *Ex Machina* (2014).

Initially, their strategy is to show that being a person means to have required certain organizational patterns, emergent properties, like cognitive skills, (self)-consciousness, awareness or moral competences and that being a person is not a question of the particular material that happens to constitute humans (organic biology) (Veruggio & Abney 2014, 353; Petersen 2014, 284). I will demonstrate that this argument implies more philosophical problems than it ought to solve. At the start, I will briefly outline three assumptions of their argument, with the aim to deliver a current example of the first theory of person to be discussed in this paper, defining *a person as a set of (sufficient and necessary) attributes*. Afterwards, I will develop a second common theory of person that defines *a person as a class of all human beings*. I will argue in the second chapter that both theories are no useful candidates in robot ethics, as well as the first and second assumption, and that we need a new approach which I finally reveal in the last chapter.

First assumption: Petersen, Veruggio and Abney seem to favour a variety of *non-reductive physicalism*[6] that compares human beings to a sort of modular building system of micro and macro properties. Though the macro properties are nomological connected with micro properties, they cannot be derived from knowledge of micro properties (Beckermann 2001, 221). To regard someone as a person, is nothing we can derive from his or her certain physiology or biology. "Being regarded as person" is a macro property that emerges or occurs, once we have merged the right (carbon based or artificial) components or material in a system. "Considerations suggest that biological humanity is not morally crucial if robots could attain first-person self-consciousness and deliberative agency, the usual requirements for moral responsibility and the hallmark for moral personhood" (Veruggio & Abney 2014, 354).

Second assumption: Further, they seem to be convinced that having once fully understood all biological processes of the human body, we can substitute them consequently with artificial body parts. Even human mind with its cognitive abilities and intelligence can be replaced by artificial algorithms. What appears here, is obviously an idea of *materialism*. It is an idea that suggests that our body (but even our world and universe eventually, too) is a result of smaller material particles. It is an idea that

would be required for them to become persons must be drawn from this fact, rather than from claims about empirical human psychology" (cf.).
Although Sparrow's theory is more differentiated and develops an interesting argument, he also accepts the categorical identification of machines becoming persons.

6 The term „non-reductive physicalism" has been used in literature many times, although it's not really clear what it means. As Beckermann puts it, the term is, if understood literally, a *contradictio in adjecto*. It implies that between mental and physical properties there must be a relationship of dependency, but that mental states cannot be reduced to their physical states (Beckermann 2001, 204). It must be mentioned that all three authors don't explicitly mention that they are physicalists or non-reductive physicalists. I derive the first assumption recurring to their recently published articles.

is strongly influenced by *positivism and scientism*, thinking the world itself exists as a nomological connected whole. Everything that exists has a material origin. In the 18th century, the French philosopher Julien Offray de La Mettrie for example believed that human body and mind work like machines [L'homme machine 1748].

> Der Mensch ist eine Maschine (…). Also kann man nur praktisch, oder durch einen Versuch der Zergliederung der Seele, nach Art der Aufklärung über die körperlichen Organe, ich will nicht sagen mit Sicherheit die Natur des Menschen enträtseln, aber doch wenigsten den möglichst höchstens Grad von Wahrscheinlichkeit über diesen Gegenstand erreichen (L'homme machine 1748, 12).

In fact, this group of robot scientists seems to restore a modern variety of La Mettrie's point of view. As Veruggio and Abney mention: "If we could gradually replace all of our higher brain biological functions with mechanical, robotic replacements, until we had a completely robotic brain, with no interruption in first-person self-consciousness, why would the result not qualify as a moral person, even if no longer completely a biological human?" (Veruggio & Abney 2014, 353) To continue, they believe that there is no relevant or at least mentionable difference between biological and artificial organs, concerning moral status of individuals: "And, if so, why would biological arms, or legs or digestive tract be morally crucial? Surely those humans with artificial arms, legs, and so forth, are full moral persons" (Veruggio & Abney 2014, 353-354).[7]

Third assumption: Following all three authors, a person is a set of certain attributes or properties. Abney and Veruggio, for example, believe that one necessary attribute of a person is a "deliberative system capable of agency" (Verruggio & Abney 2014, 47). A person is a being that has morality. To put it concisely, someone we regard as a person, is someone who has moral agency and is able to follow rules among other beings in society. Equipped with what Abney calls a deliberative system, a person is able to structure alternative possible futures as mental representations and can choose its actions based on which representations he or she wishes to become experienced reality (Abney 2014, 47). Following him, even emotions are not important for moral agency. As he argues, psychopaths and sociopaths don't lose their personhood and are still legally and morally responsible for their crimes although they have dysfunctional or missing emotional effect. He says: "But is the ancestral emotional system needed as well? What of hypothetical creatures that could rationally deliberate, yet lack emotions? Would they have morality? In other words – could (emotionless) robots be moral persons? Yes, they could" (Abney

7 Another futuristic experiment has been developed by Kevin Warwick. He argues that we can take a human brain and put it into robot body. To prove his theory, he used a neural cortex of rat fetus applying enzyms to disconnect the neurons from each other, then creating a modular closed-loop system between a 'physical' mobile robot and a cultured neuronal network, using the multielectrode array method. As he says, this method allows bidirectional communication between the culture and the robot (Warwick 2014, 319-20). Due to his first successes with the rat neurons, he concludes that we someday have the tools and technologies to develop a human brain in a robot. Again, this argument is inspired by the idea that each and every part of us as human beings can be replaced. In his case, the only thing left of a human being would be the brain. All other parts will be substituted.

2014, 47). Abneys understanding of moral competence is close to Georg Lind's theory. In his opinion, "moral competence is the ability to resolve problems and conflicts on the basis of inner moral principles through deliberation and discussion instead of violence and deceit" (Lind 2016, 13). Developing a robot with such an ability, would be a huge advantage. In many everyday situations, moral dilemma can arise. One popular example concerns the behaviour of autonomous vehicles that are the "first robots to be integrated into our lives at scale" (Lin 2017, 1), and that will likely become a standard transport solution for us in the next years. On the streets, these cars can be involved in conflicting moments, for instance, if they have to decide whether to damage people walking on the pathway or the ones sitting in the car? Lin, therefor, rightly points out that "accidents (…) will continue to happen as matter of physics, (…) even it has the best and most perfect sensors, software, and hardware the future has to offer, there'll still be the risk of crash" (Lin 2017, 2). We can never preclude that a robot car will be surprised by an unseen child who suddenly appears from behind a parked truck.

In another paper Veruggio and Abney add one more property to "their set" of requirements for personhood. They admit that a robot must attain first-person self-consciousness, too. For sure, it is not enough to construct a robot only capable of making decisions. It is important that the robot has a kind of "self, awareness" (Veruggio & Abney 2014, 354) and is able to understand the meaning of its words, in fact it must be guaranteed that the robot doesn't simulate speaking. Wittgenstein proves with his *Sprachspiele* (language games) that language requires agents who learn the meaning of words while interacting with one another. Otherwise we might have a scenario like Jean Searle illustrated it with his "Chinese room" (1984). It seems like the robot is speaking to us, although it is not understanding what its words and sentences mean. Following Petersen, being a person is a matter of having required higher "complicated organizational patterns" that allow us to regard a being as person. These patterns are a result of evolution and can occur not only in human beings, but also in any other similar intelligent subject. Thus, Petersen is convinced that even extra-terrestrial beings could fulfil specifications of personhood and might have everything it takes to be "ethically valuable" (Petersen 2014, 285).

Above-mentioned attempts belong to one – rather classical - theory that *understands a person as a set of (sufficient and necessary) attributes*. Through history of philosophy, yet, different definitions were brought up in the light of this concept.[8] A second attempt is to *define persons as a class of all human beings*. Robert Spaemann has made this prescriptive idea prominent. He argues that each and every person belongs to a kind of community of persons ("Personengemeinschaft"), from the day they are born. Being part of it, its members receive inalienable and indispensable personal rights. He points

8 Again compare (Schmiljun 2017, 75-76) where I summarize briefly some psychological, philosophical concepts of "person". A recommendable extent introduction into the notion has been published by Sturma (1997).

out: "Personenrechte werden nicht verliehen und nicht zuerkannt, sondern von jedem mit gleichem Recht in Anspruch genommen" (Spaemann 2006, 263). Further, the status of a person depends on biological affiliation to humankind. Spaemann tries to merge semantical meaning of person with the one of human beings. If a human being exists, it automatically can be regarded as person, without exception in regard of beginning and end of life. According to Spaemann, being a person is not a quality of humans. Humans are persons. "Das Sein der Person ist das Leben eines Menschen" (Spaemann 2006, 264).

In the following chapter I will demonstrate that both theories reveal definitional difficulties which doesn't make them suitable options for robot ethics. The same applies to a position of *non-reductive physicalism* and *materialism* which Veruggio, Abney and Petersen use as assumptions for their argument.

2. Why Can´t Robots Be People?[9]

> I believe current attempts to understand the mind by analogy with man-made computers that can perform superbly some the same external tasks as conscious beings will be recognized as a gigantic waste of time
>
> (Nagel 1986, 16).

One challenge concerning the first theory of person is that its varied definitions never seemed to be completed, with indefinite lists of properties, which fail to give precise parameters of what a person should be. While Locke for example believes that a person is a thinking, comprehending being that has reason, reflection and can consider itself as itself (Locke 1894, 27, 9), Sturma notes that persons live their life self-determinedly, making moral decisions and following individual plans, ideas and beliefs (Sturma 1997, 348). The same uncertainty can be detected among Abney, Veruggio and Petersen who claim different qualities they think to be decisive to constitute personhood. Of course, this definitional vagueness does not really harm scientists' belief that building Artificial People will be someday possible. It is not unlike – they could reply – that robots with "strong AI" in the sense of Wallach or Searle could fulfil some of the (Wallach & Allen 2009, 57) expected qualities like a "deliberative system capable of agency" (Verruggio & Abney 2014, 47).

Philosophical difficulties only appear if we transfer their notion to beings who lack these qualities due to physical or mental diseases, deformities or stage of life. To illustrate

9 A prominent opponent of the theory of APs is Selmer Bringsjord who already argues in his book, published in 1992, that Artificial Intelligence might bring up many things but not a robotic person. Although robots will become maybe smarter and autonomous or probably even pass difficult versions of Turing Test, the "Person building project" will fail. As he puts it: "Roughly, *very* roughly AI will down the road give us robots (or androids) whose behaviour is dazzling, but will not give us robotic *persons*" (Bringsjord 1992, 2). Nowadays, Ewa Nowak points out that the notion "person" is too metaphysical and should rather be substituted by less conflicting terms like "agents" (Nowak 2017, 117).

my point, I will provide a short *Gedankenexperiment*: Let's assume being a person meant to be capable of deliberation and acting as fully-blown moral agent. Further, let's assume robot engineers and scientist were able to build a machine with all proper qualities, so that we could regard these new creatures as persons, thus we are confronted with some questions: what about human beings who lose their ability to deliberate rationally, for example after an accident? Will they lose their personhood, too? And what about babies? Evidence suggests that children develop agency gradually through years and are not equipped with a mature sense of morality comparable to adults when they are born. To conclude, such suggested definition – *treating a person as set of properties* - would exclude a significant group of human beings, but artificial beings, too, from personhood, the moment they lack one of these mentioned properties.

Nearly the same result appears when we take the second theory of persons into account. Spaemann's strong definition excludes any other being from personhood, even if its sharing resembling qualities like (self)-consciousness, awareness or moral competences, language. As long as this being is not human, it is impossible to be part of "Personengemeinschaft". Again, we can imagine that the same robot engineers and scientists from our first "Gedankenexperiment" had become even more advanced and had created an artificial being which exactly resembles human beings in their cognitive, psychological and behaviouristic appearance. We could even consider Petersen's scenario of "Person-o-Matic machine" that is able to produce any conceivable person "out of plastic and metal" to just about any specification. To continue, the machine can also create a person out of biomolecules, by synthesizing sequenced human-like DNA from amino acids which might be placed in homegrown cellular container, "allowing the result to gestate in an artificial uterus" (Petersen 2014, 284-285). Whereas Petersen admits that there would not be any perceivable difference between carbon-based and organic-based person, according to Spaemann, these new creatures are *per definitionem* excluded from personhood. Artificial beings, built by "Person-o-Matic machine", will become in our scenario moral agents without being regarded as persons. We would receive a two-class system of two types of species claiming the same legal rights for them. Spaemann explains his restrictive idea as follows:

> Personenrechte sind Menschrechte. Und wenn sich andere natürliche Arten im Universum finden sollten, die lebendig sind, eine empfindende Innerlichkeit besitzen und deren erwachsene Exemplare häufig über Rationalität und Selbstbewusstsein verfügen, dann müßten wir nicht nur diese, sondern alle Exemplare dieser Art ebenfalls als Personen anerkennen, also zum Beispiel möglicherweise alle Delphine (Spaemann 2006, 264).

Self-consciousness and rationality don't qualify beings to be a person. Spaemann contradicts any theoretical approach that predicates personhood on a set of conditions, because it leads to arbitrariness in regard to other species that act on a comparable human-intelligence level. If we accepted self-consciousness and rationality as *sufficient attributes for personhood,* then dolphins might be persons, too. Neither first, nor latter

theory of person provide a helpful definition in order to avoid philosophical conflicts in ethics.

Finally, to favor a position of *non-reductive physicalism* (first assumption) and *materialism* (second assumption) according to possible Artificial People, requires some metaphysical consistency. On one hand, *non-reductive physicalism* claims that we are able to understand **all** microstructures of an object or subject (whatever they are) on the grounds of scientific empirical methods and that everything existing in the world is physical (and a part of a monistic system) bound by "trans-ordinal laws" (Broad 1925, 77-78)[10].

Secondly, although we fully might know all about microstructures of a system, the appearance of its macro properties like for example mental states cannot be derived from bottom-level of the system (Beckermann 2001, 221) which demands a "property dualism" (Kallestrup 2006, 459). It follows, modern robot scientists are driven by an optimistic perspective asserting that we someday gain all necessary information about human biology and may understand all physical laws so that, in a next step, we are prepared to build a machine which appears to develop same organizational patterns "(…) that [let] robot [become] a person" (Petersen 2014, 284).

But if we are not able to explain emergent properties from their microstructures, what makes us sure that we have the right constituents that will get us the desired results like a robot regarding itself as a person or a robot having mental states? And how can mental states cause physical actions and how will they be realized? Such upward explanations are limited to a certain point. Let's say, we had all knowledge to merge components to a human-like robot, and further add, this robot was intelligent, had self-consciousness and wanted to be regarded as person, *non-reductive physicalism* does not tell us how to reduce subjectivity, emotions and thoughts or the status of being a person to a physicalistic theory.

It even allows us to conclude that macro properties are epiphenomenal and can be multiply realized, a position brought up by Jaegwon Kim[11]. According to him, *non-reductive physicalism* is faced with two major problems: — the downward causation and causal exclusion arguments. Kim shows that mental states become effective for actions and must not only be realized by physical states. He argues that both upward and same-

10 There already many varieties of emergent theory in literature. It is C. D. Broad who provides a very clear explanation of emergent properties in his outstanding work *The Mind and Its Place in Nature* (1925): "Put in abstract terms the emergent theory asserts that there are certain wholes, composed (say) of constituents A, B, and C in a relation R to each other; that all wholes composed of constituents of the same kind as A, B, and C in relations of the same kind as R have certain characteristic properties; that A, B, and C are capable of occurring in other kinds of complex where the relation is not of the same kind as R; and that the characteristic properties of the whole R(A, B, C) cannot, even in theory, be deduced from the most complete knowledge of the properties of A, B, and C in isolation or in other wholes which are not of the form R(A, B, C). The mechanistic theory rejects the last clause of this assertion" (Broad 1925, 61).

11 Thanks to Jaegwon Kim, a differentiated perspective towards emergence theory was brought up into discussion including his popular theory of supervenience.

level causation entails downward causation (Kim 1999, 22-23)[12]. For example, if I burn my hand and remove it suddenly, I do it, because I have pain, precisely, I am in a mental state of pain. On the other hand, I can describe the same situation recurring to a biological story without mentioning mental states. I remove my hand because some signals have been sent to my brain, delivering an information of alarm which again release signals to my arm muscle and let me remove my hand[13]. The question now is which explanation is providing initial cause of my withdrawal?

Kim's second argument of causal exclusion replies to this question by arguing for the possibility of mental causation, implying the dilemma for a *non-reductive physicalist* to choose between either overdetermination or an epiphenomenalism (Schlosser 2006, 139). Causal exclusion involves the idea that no normal behavior or event can have more than one sufficient cause. A behavior cannot have as its cause, both a physical and a mental event, without leading to a kind of overdetermination. A *non-reductive physicalist* now can accept this fact and can give up the position of causal closure or he considers an epiphenomenalism, thus the view that mental states are caused by physical events in the brain, but have no effect upon any physical events (Beckermann 2001, 47).

Taking this strong objection into account, *non-reductive physicalism* is a serious philosophical problem for robot science. If we assume that "being a person" with all its related "what is it like to be"- characteristics[14] and its "complicated organizational patterns" – is a higher-level property of a system (not depending on a certain nature, human carbon based or artificial or even extra-terrestrial DNA), then this theory still needs to explain how property dualism can be integrated into a monistic concept of

12 Kim explains downward and upward causation more in detail which - at this point - is helpful to consider the definitional difference between both terms: "The very idea of downward causation involves vertical directionality – an 'upward' direction and a 'downward' direction. This in turn suggests an ordered hierarchy of domains that gives meaning to talk of something being located at a 'higher' or 'lower' or 'the same position in relation to another item on this hierarchy. As is familiar to everyone, positions on such a hierarchy are usually called 'levels', or sometimes 'orders' (Kim 1999, 19). Same level causation, eventually, involves causal relations between two properties. For example, one thought causes another thought (Kim 1999, 22).

13 The correct logical argument behind this example is given here: "Suppose that a property M, at a certain level L, causes another property M+, at level L + 1. Assume that M+ emerges, or results, from a property M∗ at level L (M∗ therefore is on the same level as M). Now we immediately see a tension in this situation when we ask: 'What is responsible for this occurrence of M+? What explains M+'s instantiation on this occasion?' For in this picture there initially are two competing answers: First, M+ is there because, ex hypothesis, M caused it; second, M+ is there because its emergence base M∗ has been realized. Given its emergence base M∗, M+ must of necessity be instantiated, no matter what conditions preceded it; M∗ alone suffices to guarantee M+'s occurrence on this occasion, and without M∗, or an appropriate alternative base, M+ could not have occurred. This apparently puts M's claim to have caused M+ in jeopardy. I believe that the only coherent description of the situation that respects M's causal claim is this: M causes M+ by causing its base condition M∗. But M's causation of M∗ is an instance of same- level causation. This shows that upward causation entails same-level causation; that is, upward causation is possible only if same-level causation is possible" (Kim 1999, 23).

14 I quote Nagel's very prominent question here as a kind of short version of the problem he discusses in his book "What is it like to be a flat" where he outlines the difficulties to reduces the first-person view of a being into physical notions (Nagel 1974, 438). Elsewhere, Nagel asserts: "Eventually, I believe, current attempts to understand the mind by analogy of with man-made computers that can superbly some of the same external tasks as conscious beings will be recognized as a giant waste of time" (Nagel 1986, 16).

world, without risking a case of epiphenomenalism or overdetermination of mental and physical events. Maybe a strategy to solve this vicious circle is to follow Kim's own *theory of supervenience*[15]. He argues that higher-level properties must be reducible to lower-level properties and we can even question the difference between these two levels (Kim 1999, 47). If he was right, then mental and physical states are simply two different properties of the same entity. Thus, any change of our mental states includes a change in our physical states. But this approach again brings with it a host of new questions – some of them comparable to the identity theory of mind and brain (Stärk 2013, 32) – which cannot be outlined and answered here, due to it would extend the scope of this paper.

To return to the problem of *non-reductive physicalism*, it appears that – inspired by this theory – somehow strange ideas come up like for example Petersen's scenario of a "Person-on-Matic", a futuristic machine producing Artificial People on artificial DNA. Although creative and detailed as it is, Petersen leaves out the important question of realisation. He develops a metaphysical argument about the possibility of a precise and full reconstrunction of human beings, including also the whole process of evolution.

> Perhaps, after you feed [Person-on-Matic] with complete information about your DNA makeup and DNA makeup of a willing partner, the Person-on-Matic uses a combination of this information to construct from scratch a viable zygote that matures in an artificial uterus, much later producing an infant, exactly as might have been produced by the more traditional route. (…) After all, it results in people just like the people we now create by traditional means (Petersen 2014, 286).

What robot scientists like Petersen encourage here is their belief in *materialism.* Every elementary particle and atom in the universe without exception is made of material (Gabriel 2013, 16). Given this premise, they think our scientific methods will allow us to detect and analyse every microstructure and substitute it with an artificial one. Nagel argues that this view is "based ultimately on a form of idealism" (Nagel 1986, 26). Materialists believe that the only access of understanding the reality is provided by restrictive objectivity and that we can cross out subjectivity in our research. But as Nagel puts it, "we will not know exactly how scrambled eggs taste to a cockroach even if we develop a detailed phenomenology of the cockroach sense of taste" (Nagel 1986, 25). One result derived from above mentioned premise is that even mental states have to be material. Thus, emotions, thoughts and words are material entities. But do they have the same ontological status as cars, atoms or plants (Schmiljun 2017b). Above, if

15 Kim introduces this interesting theory of dualist monism named "supervenience" in his in 1993 published book "Supervenience and Mind". Kim differs two versions of supervenience: a weak and strong supervenience. "Weak supervenience (…) requires that *within* any possible world there are not be two things agreeing in B but diverging in A, and this condition is met in each of these cases" (Kim 1993, 60). To illustrate this point: For any being or object x1 and x2 (in the real world) the following applies: if x1 and x2 have the same physical properties, then they share the same mental properties. Strong supervenience is defined like that: For any being or object x1 and x2 in any possible world w1 and w2, the following applies: If x1 in w1 has the same physical properties as x2 in world w2, then x1 in w1 has the same mental properties as x2 in w2 (Beckermann 2001, 208).

really everything was material, then even the verity of theory of materialism must be a configuration of elementary particles and atoms – an objection developed by Gabriel. But not every thought is automatically true, only because it is manifested in our brain. Otherwise probably all of our thoughts, constituted in our brain, might be categorically true. Materialists need to show an explanation how such a materialistic fundament of verity or realisation can be understood, in order to justify their theory. Is verity a result of elementary particles? (Schmiljun 2017b, 133-134).

As illustrated, the argument of Artificial People is based on three conflicting assumptions (first: the idea that it is possible to define persons either as a set of properties or class of human beings and apply it to robots, second: non-reductive physicalism, third: materialism). Thus, without taking these theoretical doubts into consideration, rather weird descriptions of a possible future follow. Petersen asserts: "We should treat APs well, whether organic or inorganic, not because they could be mistaken for people, but because they *are* people (Petersen 2014, 294) In the next chapter I will draw a new approach to the debate on the ground of Christine Korsgaard that avoids metaphysical ballast.

3. Why Should Robots Rather Be Regarded As Ends in Themselves?

On the ground of Kant's transcendental philosophy, Korsgaard develops her argument why we should treat animals not as "mere means" and "instruments" for human purpose, but rather as ends in themselves.[16] I assert that we can apply her argument to robots, too, once these robots are capable of pursuing their own goods. Before I will explain my point, I try to reconstruct Korsgaard's essential ideas.

The first significant assumption of her argument is that animals have agency, understood as the competence to act autonomously and efficaciously, just like any other human being[17]. As Korsgaard puts it:

> When an animal acts, he is determined by his form, by his instincts, to produce a change in the world, guided by his conception or representation of the world. But an animal's form is what gives him his identity, what makes him the animal he is (…) Action is self-determination, and, to that extent, it is autonomous that the question of its efficacy can come up. If one thing causes another, there is no room for success or failure. But if an animal determines herself to be the cause of something, and yet does not bring that thing about, then she has failed. Autonomy and efficacy are the properties of agents – all agents, not just human agents (Korsgaard 2009, 106-107).

16 Kant asserts: „Beings the existence of which rests not on our will but on nature, if they are beings without reason, have only a relative worth, as means, and are therefore called things, whereas rational beings are called persons because their nature already marks them out as an end in itself, that is, as something that may not be used merely (...)" (*Groundwork for the Metaphysics of Moral* 4: 428).

17 Her definition matches also with the explanation, given by Oxford Dictionary, where it said that agency is an action or intervention producing a particular effect. Nevertheless, we have to differ this definition to moral agency which somehow seems to be more complex.

Equipped with agency, Korsgaard further concludes – contradicting Kant – in her second assumptions that animals are *ends in themselves*. Kant believes that we are allowed to kill animals. He advises to kill the animal quickly and without pain, and it shouldn't be for the sake of sport (Korsgaard 2018, 99). The term "ends in itself", in the sense of Kant, addresses something that is not a mean to any purpose and provides itself a motive to pursue it without any further reason. Kant states that only human beings, persons, in the first place, can be considered as ends in themselves due to their ability to make rational choices, "in general every rational being, exists as end in itself, not merely as means to the discretionary use of this or that will, but in all its actions, those directed toward itself as well as those directed toward other rational beings, it must always at the same time be considered as an end" (Kant, Groundwork 4:428). Rational choices are the result of an assessment, based on reasons and deliberation. They deliver decisions about what should be done (Korsgaard 2012, 9). Above, rational choices have the character of a law, in the sense of an objective or principle (universal law), a value and good absolutely "to which every rational being must then be responsive" (Korsgaard 2012, 8). Whereas Kant presupposes rationality as a criterion for an end in itself[18], Korsgaard suggests a second interpretation of ends in themselves. According to her, the term "end in itself" can have an *active and a passive sense*. Kant refers only to the *active sense* that something is an end in itself if it is a law-maker (capable of rational, and therefore moral choice). But human beings, as she correctly argues, are not "merely rational beings" (Korsgaard 2012, 11). She criticises that the idea that rational choice involves a presupposition that we are ends in ourselves is not the same as the idea that rational choice involves a presupposition that rational beings are ends in themselves. Of course, human beings can be non-rational, too (Korsgaard 2012, 11). Not every decision and not every choice can be traced back to rationality[19]. Therefore, she concludes a *passive sense* that something is an end in itself if it pursues its own good and interests (even if it is not capable of rational or moral choice). As she puts it, many things that we take to be good for us are not good for us merely insofar as we are autonomous rational beings. "Food, sex, comfort, freedom from pain and fear" affect us insofar as we are animate beings (Korsgaard 2012, 13).

Given that, Korsgaard believes animals can be ends in themselves at least in the second sense, as they are beings with interests and have (first assumption) autonomy. Animals are able to like or dislike things, be happy or suffer. They have a certain point of view that let things become either "good for" them or "bad for" them. For example, if I stop feeding my cat it will be a matter of time and she will starve. To cease feeding my cat is something that is "bad for" her. On the other hand, if I organize a scratching post,

18 Kant believes that the original nature of human beings is rational. Therefore, he concludes: "(…) rational nature exists as end in itself. The human being necessarily represents his own existence in this way" (Kant, Groundwork 4:429).

19 It rather seems like our brain operates on two different methods of making decisions. One method grounds on rational choice, the other method grounds on intuitivism and emotions. Kahnemann developed this theory in his outstanding book named Schnelles Denken, langsames Denken (2011).

I know this will be "good for" her, due to it helping her to take care of her claws. But acting on such reasons, I accept my cat as an autonomous being with certain interests. I value her as a being that is able to pursue her own good[20]. As a result of this, Korsgaard is convinced that human beings owe duties to other human beings as well as to animals, in contrast to Kant who says: "As far as reason alone can judge, a human being has duties only to human being (himself and others), since his duty to any subject is moral constraint by that subject's will" (Kant, MM 6, 442). To understand Korsgaard's disagreement with Kant, it is helpful to consider her understanding of rational and instinctive beings. To her, *rationality* is a normative power "grounded in a certain form of self-consciousness" (Korsgaard 2018, 39) and can be seen as a distinctive characteristic of human beings[21]. Animals on the other hand are driven mostly by an *instinctive* behaviour. But Korsgaard doesn't interpret the term as a mere reaction or movement that is wholly automatic and simply caused, not intentional. Her idea of being instinctive is connected with the ability to learn from experiences, and to solve problems by taking thought. She argues that being instinctive is compatible with being intelligent. Thus, the difference between rationality and intelligence lies in the fact that the former looks outward at the world, questions about the connections and relations like causal relations, special relations, temporal and social relations. Rationality means to look inward, "at the workings of our own minds" (Korsgaard 2018, 40). It means to be able to ask and reply normative and evaluative questions about the connections and relations to be found in the world. With this digression it becomes clear why Korsgaard highlights the value on animals as ends in themselves. As she puts it elsewhere, an animal has a kind of identity, an animal is not just only a substance. It has consciousness that gives the animal a point of view and let it become an agent that is "a subject, (…) and someone" (Korsgaard 2009, 129). To continue: "If she [the animal] is also fairly intelligent, you can interact with her, play with her, get annoyed at her, or adore her. Even if she is not very intelligent you can sympathize with her and enter into her concerns, or be hostile to her as the enemy" (Korsgaard 2009, 129).

20 Korsgaard suggests that her argument could even cover plants considered as "living organisms, functioning so as to survive and reproduce" (Korsgaard 2012, 13). For example, weeds will grow better with rain, which let us assume that rain is something which is good for them. Her idea suggests that there is a dependency between the sense of "good for" and the evaluative attitudes of a being for whom things can be good or bad. She defines "evaluative attitudes" as psychological states like desires, pains, pleasures, fear, loves, hates, ambitions, projects and principles which are experienced by sensate beings (Korsgaard 2012, 13). A weakness of her argument - at least in the quoted text - is the point that Korsgaard doesn't seem to differentiate between the autonomy of animals and other living organisms like plants. But of course, plants don't have any brain which constitutes consciousness. A tree's sensation of environment is obviously different to a sensation of a bat. Based on Aristotle, in "Fellow Creatures" Korsgaard provides a definition of animals in comparison to living things like plants. A living thing, Aristotle claimed, is mostly characterized by reproduction and nutrition, and are self-maintaining entities. An animal, in turn, is a living thing capable of perception and voluntary motions. "Animals maintain themselves in part by forming representations or conceptions of their environment in accordance with those representations" (Korsgaard 2004, 82-83).

21 She agrees with the philosophical tradition which can be traced backed to John Locke, David Hume and Francis Hutcheson: rationality is what make human beings become people (Korsgaard 2018, 42).

Dretske argues similar in his philosophical investigation on animal minds. Instead of the term "instinctive", Dretske rather says that animals are minimal rational (Dretske 2005, 213-215). Animals are rational but not in the sense of human beings. They are rational due to they follow reasons that concern their behaviour. The difference between them and us is only that they don't evaluate their reasons for their actions if they are good or bad. But like us, animals are equipped with "creature consciousness", in fact, with the ability to be "conscious of things—of objects (the bug in my soup), events (the commotion in the hall), properties (the color of his tie), and facts (that someone is following me)" (Dretske 2001, 23).

Let me summarize Korsgaard's premises and conclusions as follows: According to her, animals have agency, understood as the competence to act autonomously and efficaciously, and are beings with interest, for whom things can be good or bad. Further, she differentiates between two versions or senses of "end in itself":

> Ends in themselves are beings insofar as he or she pursues an end only if he or she thinks it is good absolutely, capable of willing his or her principles as universal law (*active sense*).

> Ends in themselves are beings with interests, for whom things can be good or bad (*passive sense*).

Based on this, she concludes that animals can be understood as ends in themselves in the second sense as beings with interests, while human beings respectively people cover both senses.

As mentioned, I argue that we can apply Korsgaard's argument to robots, too. But this requires at least another – but very crucial – assumption, namely, that robots were capable of autonomy, following their interests and good like animals someday. To put it in other words, my assumption involves a presupposition of agency. In this case, we can follow that robots might be considered as ends in themselves[22]. As ends in themselves

22 I admit that my argument is not as unproblematic as it might seem. This premise as well requires some philosophical assumptions, apart from the question whether robots will ever reach a level of intelligence und autonomy comparable to animals. I will avoid the notion "consciousness" here at this point, due to there is a very controversial debate whether animals are conscious or not. Some scientists like for example Searle (2005, 132), Davidson (2005, 117-118) or Dennett (2005, 402) are convinced that some animals have consciousness and even rationality. My premise assumes that robots will be able to be aware of their aims, interests, guided by their conception or representation of the world. Above, this argument involves the assumption that robots are capable of perception and voluntary motions and can act on instincts and emotions like animals. It might be objected that I probably make the same mistake as Abney, Veruggio or Petersen, assuming that designing a robot with intelligence of an animal is not unrealistic. The difference between them and this approach is that I don't think we can build a "human sapiens 2.0" or "Artificial People". Eric Dietrich for example believes that "moral environment of modern Earth wrought by humans, together with what current science tells us of morality (…) *morally* requires us to build our own replacement and the exist stage left" (Dietrich 2018, 531). I don't share this euphoric forecast of possible Artificial People, due it involves some doubtful and problematic metaphysical assumption, as shown. But to sum, the weakness of my assumption is that it doesn't precisely define sufficient and necessary conditions, to be fulfilled by a robot, so that the argument runs without "animals". Without a concreter restriction, I believe this argument would work for any other imaginable creature that somehow meets the nature of animals, too.

we would need to regard robots' autonomy and their certain points of view. We cannot simply treat them as means for our purposes, just like Korsgaard recommends not to treat animals as mere means for our purposes. It will be a matter of fact that these robots are beings with interests, for whom things can be good or bad. Surely, this implies that I need to respect them as non-human and in-organic moral agents, even if they are not capable of rationality like us and can't ground their actions on deliberation. In her recent paper, Nowak already discusses such an idea, saying that an

> artificial intelligent device [has] no biological instincts and natural ends, however, analogously to the animal, they [may someday] produce representations of the world and are provided with some laws and ends whose repeated applications, combined with learning process, may give them identity, and even some selfhood as an individual agent (Nowak 2017, 171).

I agree with her that the term *agent* is less metaphysical or spiritual then terms such as "person" or "subject"[23]. Besides, Scheutz and Schermerhorn demonstrate in their work that moral agents seldomly make only rational decisions. Time, knowledge and many other resource limitations provide rather "affective evaluations", meaning fast, low-cost mechanisms for "estimating the value of an object, event, or situation for an agent", instead of complex and more computationally intensive *cognitive evaluations*" (Scheutz & Schermerhorn 2009, 74). Both scientists, of course, follow current psychological studies published by Kahneman, Wakker, and Sarin (1997) or Blanely (1986). Thus, Scheutz and Schermerhorn focus in their different experiments on the possibility to integrate affect mechanisms in social robots, in order to let them cope better with "intrinsic resource limitations of the real world" and help them to become "sensitive to human affects", when there are demanded to interact with humans [Scheutz and Schermerhorn 2009, 74]. Together with Kramer, Brick, Anderson and Dingler, they have developed a platform named DIARC – a "distributed integrated affect cognition and reflection" architecture – that should help robots to interact with humans naturally [Scheutz, Schermerhorn Kramer, Brick, Anderson and Dingler 2006, 1]. DIARC integrates cognitive abilities like natural language understanding and complex action planning and sequencing "with lower level activities (such as multi-modal perceptual processing, feature detection and tracking, and navigation and behavior coordination" (Scheutz and Schermerhorn 2009, 74-75). Having social robots acting on the ground of DIARC, it will foster their acceptance in our society. Additionally, it will help humans to regard robots as moral agents and ends in themselves.

23 Following this interpretation, we certainly avoid mistakes to exaggerate the status of robots in our society like in the case of the social robot "Sophia" that has been produced by Hanson Robotics. Sophia has become famous through several interviews. She can move her head and neck and can gesture with her hands. She can be seen as the next evolution of a chatbot that is able to understand and mimic conversation (Hambling 2018, 198). Her example shows what happens if the abilities of robots are too much overestimated. It may lead to wrong assumptions and confusing consequences. In her case, it resulted into the doubtful political decision of the government of Saudi-Arabia to acknowledge her with the citizenship in October 2017.

4. Conclusion

Responding to the question "Will robots in the coming age (…) be persons?", Selmer Bringsjord puts it clear in a short reply: "Robots will get very flashy, but (…) they'll never be people" (Bringsjord 1992, 6). As shown, the modern strategy of robot scientists like Abney, Veruggio and Petersen has to tackle three major metaphysical assumptions which are rather problematic, then a profound solution for robot ethics. Firstly, we have to refrain from the idea that we can develop a precise definition of persons that can be applied to robots or used as a mean to differentiate between artificial and human beings. Secondly and thirdly, non-reductive physicalism and materialism provoke strong philosophical problems like overdetermination, epiphenomenalism or ontological coherence and coherence of variety. Thus, Korsgaard's argument, applied to robots, given the fact that robots might have autonomy and could follow their interests someday, is maybe a reasonable attempt to encounter robots on a more neutral level. The advantage of this attempt lies obviously in the fact that robots don't need to fulfil a complicated list of psychological, cognitive or behaviouristic conditions to be regarded as human-like Artificial People. If we regard them as ends in themselves and beings with interests, we need to take care for them the same way we need to take care for other living beings.

References

Abney K., Veruggio G. 2014. "Roboethics: The Applied Ethics for a New Science." In Lin P., Leith A. & Bekey G. A. (Eds.), *Robot Ethics. The Ethical and Social Implications of Robotics*. Cambridge: MIT University Press (347-64).

Abney K. 2014. "Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed." In Lin P., Leith A. & Bekey G. A. (Eds.), *Robot Ethics. The Ethical and Social Implications of Robotics*. Cambridge: MIT University Press (35-52).

Beckermann, A. 2001. *Analytische Einführung in die Philosophie des Geistes.* Berlin: Walter de Gruyter.

Bringsjord S. 1992. *What Robots Can and Can't Be?* Luxemburg: Springer Science+Business Media S.A.

Broad C. D. 1925. *The Mind and Its Place in Nature*. New York: Harcourt, Brace & Company, Inc.

Davidson D. 2005. "Rationale Lebewesen." In Wild M. & Perler D. (Eds.), *Der Geist der Tiere*. Frankfurt am Main: Suhrkamp (117-31).

Dennett D. C. 2005. „Das Bewusstsein der Tiere: Was ist wichtig und warum?" In Wild M. & Perler D. (Eds.), *Der Geist der Tiere*. Frankfurt am Main: Suhrkamp (389-407).

Dietrich E. 2018. "Homo Sapiens 2.0. Building the Better Robots of Our Nature." In Anderson M. & Anderson S. L. (Eds.), *Machine Ethics*. Cambridge: University Press (531-37).

Dretske F. 2005. „Minimale Rationalität.“ In Wild M. & Perler D. (Eds.), *Der Geist der Tiere*. Frankfurt am Main: Suhrkamp (213-222).

Dretske F. 2001. „Animal Minds.“ *Philosophic Exchange* 31(1):21-33.

Gabriel M. *Warum es die Welt nicht gibt?* Berlin: Ullstein.

Hambling D. 2018. *We: Robots*. *The Robots That Already Rule the World.* London: Quarto Publishing.

Kahneman D. 2011. *Schnelles Denken, Langsames Denken*. München: Pantheon.

Kallestrup J. 2006. “The Causal Exclusion Argument.” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 131(2):459-85.

Kant I. 1785. *Groundwork for the Metaphysics of Moral*. Ed. and trans. by A. W. Wood. New Haven and London: Yale University Press.

Kim J. 1993. *Supervenience and Mind*. Cambridge: Cambridge University Press.

Kim J. 1999. “Making Sense of Emergence.” *Philosophical Studies* 95(1/2):3-36.

Korsgaard C. M. 2012. “A Kantian Case for Animal Rights”. In Michel M., Kühne D., & Hänni J. (Eds.), *Animal Laws – Tier und Recht. Developments and Perspectives in the 21st Century*. Zürich/ St. Gallen: DIKE (1-28).

Korsgaard C. M. 2009. *Self-Constitution. Agency. Identity and Integrity*. Oxford and New York: Oxford University Press.

Korsgaard C. M. 2004. “Fellow Creatures: Kantian Ethics and Our Duties to Animals.” *Tanner Lectures on Human Values* 24:77-110.

Korsgaard C. M. 2018. *Fellow Creatures: Our Obligations To the Other Animals*. Oxford: University Press.

Lind G. 2016. *How to Teach Morality?* Berlin: Logos Verlag.

Lin P., Leith A., & Bekey G. A. 2014. *Robot Ethics. The Ethical and Social Implications of Robotics.* Cambridge: MIT University Press.

Lin P., Jenkis R., & Abney K. 2017 (Eds.). *Robot Ethics 2.0. From Autonomous Cars to Artificial Intelligence*. Oxford: Oxford University Press.

Locke J. 1894. *An Essay Concerning Human Understanding.* Ed. by A. Campbell Fraser. 2 vols. Oxford: Clarendon Press.

La Mettrie J. O 1774. *Der Mensch eine Maschine* (org. L’homme Machine). Berlin: Holzinger.

Nagel T. 1974. “What Is It Like to Be a Bat?” *The Philosophical Review* 83(4):435-50.

Nagel T. 1986. *The View from Nowhere*. Oxford – New York – Toronto: Oxford University Press.

Nowak E. 2017. “Can Human and Artificial Agents Share an Autonomy, Categorical Imperative-based Ethics and ‘Moral’ Selfhood?” *Filozofia Publiczna I Edukacja Demokratyczna* 6(2):169-208. E-access: https://pressto.amu.edu.pl/index.php/fped/article/view/13198/12903 , https://doi.org/10.14746/fped.2017.6.2.20

Petersen S. 2014. "Designing People to Serve", in Lin P., Leith A., & Bekey G. A. (Eds.), *Robot Ethics. The Ethical and Social Implications of Robotics*. Cambridge: MIT University Press (283-98).

Scheutz M. & Schermerorn P. 2009. „Affective Goal and Task Selection for Social Robots." In *Handbook of Research and Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence*. Hershey: IGI Global (74-87).

Schermerhorn P., Kramer J., Brick T., Anderson D., Dingler A., & Scheutz M. 2006. „Diarc: A Testbed for Natural Human-Robot Interactions." *Proceedings of AAAI 2006 Robot Workshop*.

Searle J. R. 1984. *Minds, Brains and Science*. Cambridge, MA: Harvard University Press.

Searle J. R. 2005. „Der Geist der Tiere." In Wild M. & Perler D. (Eds.), *Der Geist der Tiere*. Frankfurt am Main: Suhrkamp (132-52).

Schlosser M. E. 2006. "Causal Exclusion and Overdetermination." In Di Nucci E. & McHugh J. (Eds.), *Content, Consciousness and Perception*. Cambridge: Cambridge Scholars Press (139-55).

Schmiljun A. 2017a. "Robot Morality. Bertram F. Malle's Concept of Moral Competence." *Ethics in Progress* 8(2):69-79.

Schmiljun A. 2017b. „Symbolische Formen und Sinnfelder. Probleme und Unterschiede eines gemeinsamen Projekts." In Hamada Y., Favuzzi P., Klattenhoff T. & Nordsieck V. (Eds.), *Symbol und Leben. Grundlinien einer Philosophie der Kultur und Gesellschaft.* Berlin: Logos (129-44).

Spaemann R. 2006. *Personen. Versuche über den Unterschied zwischen ‚etwas' und 'jemand'.* Stuttgart: Klett-Cotta.

Stärk J.-P. 2013. *Das Leib-Seele-Problem, die Hirnforschung und die exzentrische Positionalität*. Hamburg: Diplomica.

Sturma D. 1997. *Philosophie der Person. Die Selbstverhältnisse von Subjektivität und Moralität.* Paderborn – München – Wien – Zürich: Mentis.

Wallach A. & Allen C. 2009. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.

André Schmiljun
(Berlin)

Why Can't We Regard Robots As People?

Abstract: With the development of autonomous robots, one day probably capable of speaking, thinking and learning, self-reflecting, sharing emotions, in fact, with the raise of robots becoming artificial moral agents (AMAs) robot scientists like Abney, Veruggio and Petersen are already optimistic that sooner or later we need to call those robots "people" or rather "Artificial People" (AP). The paper rejects this forecast, due to its argument based on three metaphysical conflicting assumptions. Firstly, it is the idea that it is possible to precisely define persons and apply the definition to robots or use it to differentiate human beings from robots. Further, the argument of APs favors a position of non-reductive physicalism (second assumption) and materialism (third assumption), finally producing weird convictions about future robotics. Therefore, I will suggest to follow Christine Korsgaard's defence of animals as ends in themselves with moral standing. I will show that her argument can be transmitted to robots, too, at least to robots which are capable of pursuing their own good (even if they are not rational). Korsgaard's interpretation of Kant delivers an option that allows us to leave out complicated metaphysical notions like "person" or "subject" in the debate, without denying robots' status as agents.

Keywords: artificial people; moral agency; non-reductive physicalism; materialism; ends in themselves; animals.

Ethics in Progress (ISSN 2084-9257). Vol. 9 (2018). No. 1, Art. #3, pp. 44-61.

Doi: 10.14746/eip.2018.1.3