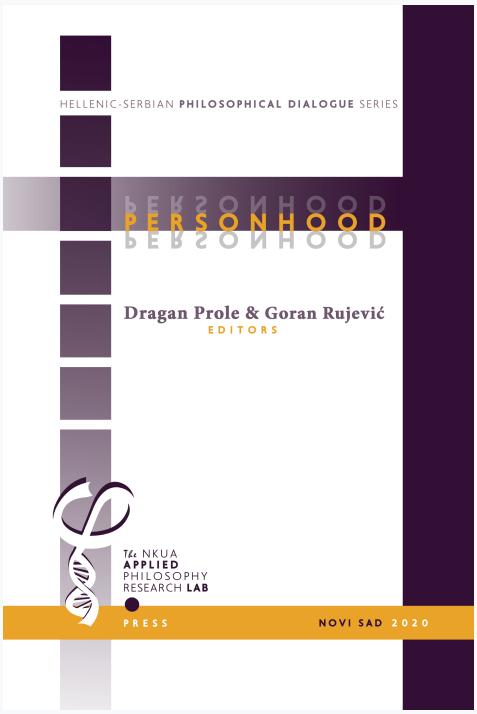
NKUA Applied Philosophy Research Lab Press




**Chapter**

## Licensed to Kill: Autonomous Weapons as Persons and Moral Agents

*Alkis Gounaris, George Kosteletos*

---

**Personhood**

*Dragan Prole, Goran Rujević, George Arabatzis, Una Popović, George Boutlas, Dragan Prole, Panagiotis Kormas, Antonia Moutzouri, Damir Smiljanić, Alkis Gounaris, George Kosteletos, Marica Rajković, Ioannis Ladas*

doi: 10.12681/aprlp.49.925

Copyright © 2020, NKUA Applied Philosophy Research Lab Press & Filozofski Fakultet Novi Sad

This work is licensed under Creative Commons Attribution 4.0.

https://epublishing.ekt.gr | **e-Publisher:** EKT | **Downloaded at:** 29/11/2026 21:33:33

# Licensed to Kill: Autonomous Weapons as Persons and Moral Agents

**Alkis Gounaris**
National and Kapodistrian University of Athens
E-mail address: alkisg@philosophy.uoa.gr
ORCID iD: https://orcid.org/0000-0002-0494-6413

**George Kosteletos**
National and Kapodistrian University of Athens
E-mail address: gkosteletos@philosophy.uoa.gr
ORCID iD: https://orcid.org/0000-0001-6797-8415

**Abstract:** The debate over the attribution of personhood to non-human entities is of an increasing concern to both academia and institutions. The intelligence, autonomy and efficiency exhibited by modern AI systems raise pressing questions regarding the moral responsibility issues their use entails. In our paper we focus our discussion on autonomous war machines, as their actions touch upon issues of life and death and their design, production and use cause philosophical controversies. Prompted by the classic position of Daniel Dennett defending the possibility that autonomous intelligent systems are responsible for their actions, we consider a) the argument of cognitive and / or functional equivalence of humans and machines, b) the argument of autonomy as such and c) the argument of excessive efficiency of the actions of intelligent machines. Our investigation upholds a skeptical stance towards the issue of recognition of moral personhood, while illuminating aspects such as the difference between cognition and intelligence, the necessary and sufficient conditions that imply moral responsibility and the differences between the ontological and the epistemological examination of the above problems. Finally, the contradiction between the demand for clear and solid decision-making criteria and the endless nature of a philosophical analysis that strives to be consistent is emphasized.
**Keywords:** artificial intelligence; autonomy; moral personhood; autonomous weapons; human - machine intelligence equivalence; AI excessive effectiveness.

doi: https://doi.org/10.12681/aprlp.49.925

In October 2017, the humanoid Sophia became the first artificial intelligence entity to become a citizen of Saudi Arabia.[1] Two years before, the European Parliament Committee on Legal Affairs had suggested the need to establish a legal framework for the recognition of the civil rights and obligations of intelligent "electronic persons" who make autonomous decisions.[2] This framework is still not outlined at the moment, as conflicting views are expressed on the subject, particularly regarding the issues of liability and moral responsibility resulting from the autonomous operations of intelligent systems.[3]

The issue of moral responsibility in AI systems concerns today, as we will see later, both the philosophical and the research community and is closely related to the concept of person. But can we "literally" attribute the term personhood[4] to artificially

[1] Chris Weller, "Meet the First-ever Robot Citizen - A Humanoid Named Sophia that Once Said It Would 'Destroy Humans,'" *Business Insider*, accessed July 30, 2020, https://www.businessinsider.com/meet-the-first-robot-citizen-sophia-animatronic-humanoid-2017-10?r=UK.

[2] About electronic persons see https://www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf.

[3] "For the purposes of liability, it is not necessary to give autonomous systems a legal personality." Further reading at https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=36608.

[4] It should be noted that the recognition of legal personhood in objects, animals, plants or artificial intelligence systems is an issue that has long preoccupied lawyers and philosophers, see Lawrence Solum, "Legal Personhood for Artificial Intelligences," *North Carolina Law Review* 70, no. 4 (1992): 1231-1287. Specifically, the attribution of the status of a person is discussed, corresponding to the attribution of the status of a legal entity to non-natural entities such as companies, institutions, municipalities, government agencies, etc. that carry out operations, contract, have rights and obligations, responsibilities and demands. At present, however, we will only be concerned with the attribution of moral status to intelligent machines, as such a perspective is directly related to personhood recognition in AI systems.

intelligent systems that behave like real people? According to the philosophical defense of the fictional character HAL 9000 by Daniel Dennett, provided these systems make intelligent and autonomous decisions and take effective actions, these actions can be evaluated "morally" just like the corresponding human ones.[5] That is, if a system thinks, acts and behaves like (or even better than) a human, it will be able to bear moral responsibility for its actions and be considered a moral person.

In this paper we will mainly consider the problem of the moral responsibility of machines, which leads to a number of issues concerning the moral person concept. We begin from the assumption that if someone or something can be characterized as a moral entity, then he/it can very well be considered as having the status of a person in general, while the opposite does not necessarily happen.

We have chosen to focus our discussion on 'autonomous' military machines, namely machines that purportedly decide autonomously on matters of life and death, due to the great urgency of the ethical issues caused by their design, construction and use, but we consider that our arguments can as well be valid regarding any other AI system.

Focusing our investigation on intelligent 'autonomous' military machines, we are faced with questions such as whether – and under what conditions – should intelligent systems make 'autonomous' life and death decisions. Also, whether intelligence, autonomy and efficiency are necessary as well as sufficient conditions for an agent to be considered a moral being. And if so, then should these war machines, in addition to being responsible for their actions, take up military positions and join the military hierarchy not as weapons but as soldiers? Would this possibly mean that they should enjoy the benefits of the Geneva Conventions regarding prisoners of war if arrested, or that they should be held accountable in military courts for their actions and omissions or insubordination?

---

[5] Daniel Dennett, "When Hal Kills, Who's to Blame? Computer Ethics," in *Hal's Legacy: 2001's Computer as Dream and Reality*, ed. David G. Stork, 351-365 (Cambridge, MA: MIT Press, 1997).

Today, the involvement of artificial intelligence systems in government, military, space and other operations is no longer the fictional content of films such as *2001, A Space Odyssey*. 'Autonomous' war machines, as well as other systems, already operate for defensive or offensive purposes and are tested in real war situations. States are already in an armaments race in order to gain a competitive edge, and the defense industry is paving the way in this research direction. This is one of the main reasons why we focus on AI war systems, considering whether the concept of moral person – and consequently of person – can be attributed to AI systems, as the severity of the consequences of their actions is proportional to the severity of the moral questions which the latter one raises.

The pressing context in which philosophers and AI researchers are called upon to deal with these new problems and the ethical issues that arise, is revealing. In 2015, Stuart Russell, Max Tegmark and other AI and Robotics Researchers, published an open letter, requesting a ban on the development of autonomous weapon systems and killer robots.[6] Among other things, they report that 'autonomous' weapon systems are today the third revolution in military operations (after gunpowder and nuclear weapons) and due to their relatively low cost and ease of manufacture, they are expected to be widely distributed and mass-produced, with the risk of being used for terrorist acts, ethnic cleansing, assassinations, destabilization of nations, enslavement of populations and selective extermination of national or social groups. For this reason, they call on AI researchers to refuse to participate in the research and construction of such weapon systems, the same way that biologists, chemists and physicists, respectively, widely support similar international agreements to ban chemical and biological or laser-equipped weapons.

---

[6] Stuart Russell, Max Tegmark, et al., "Autonomous Weapons: An Open Letter From AI & Robotics Researchers," *Future of Life Institute*, accessed July 25, 2020, https://futureoflife.org/open-letter-autonomous-weapons/.

In the opposite direction, the endorsers of these systems argue that war machines will only pursue their target in a legal and accurate manner and will strictly follow the provisions of the International Conventions for the wounded, civilians, prisoners of war etc., while in contrast to human soldiers they will never be under psychological pressure, they will not make mistakes due to fatigue and they will not commit revenge atrocities (as is often the case with soldiers, who may prove to be mentally and emotionally vulnerable). Therefore, intelligent machines can become in the future the ideal model of the moral soldier, as they will respect opponents, civilians, infrastructures etc.[7]

It is understood that the discussion around ethical problems raised by the design, production and use of autonomous weapon systems is related to:

> 1) Whether or not there should be such systems – a problem related to (1.1) their expediency and their possibly malicious use and (1.2) their ontological status, as formulated by their autonomy, intelligence and effectiveness, (1.2.1.) as well as whether their action is morally evaluable and (1.2.2) the systems themselves are, possibly morally responsible, and
> 2) How are we to determine if the system ultimately acted autonomously and as a moral person.

In the present investigation we focus on the epistemological question (2), the answer to which, however, is inextricably linked to (1.2), that is the ontological status and the criteria required to consider someone or something as a moral person.

Due to the conceptual vagueness as well as the differences in the use of the same terms between the philosophical and the technical vocabulary, we deem it appropriate to make some introductory clarifications.

---

[7] Ugo Pagallo, "Robots of Just War: A Legal Perspective," *Philosophy & Technology* 24, no. 3 (2011): 307-323; Wendel Wallach, and Collin Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford: Oxford University Press, 2009).

Speaking of intelligent autonomous military machines we refer mainly to Autonomous Weapon Systems (AWS), and Lethal Autonmous Weapon Systems (LAWS). These systems, as defined by the UK Department of Defense (2011), are capable of "understanding" instructions, intentions, environments etc. and after considering the alternatives, to decide autonomously and take actions that cannot be foreseen in advance.[8] Hereto, what is claimed to make war machines "perceive," "understand," decide and act alone, utilizing and evaluating complex information in order to achieve a specific mission, is Artificial Intelligence.[9]

Although philosophers disagree on the exact definition of intelligence, we could accept that by this term we mean the ability of an entity to achieve complex goals.[10] In other words, it is a computational process in which information is transformed through functions (op. cit.). According to Haugeland however, Artificial Intelligence researchers and developers aim to create a *genuine* intelligence, rather than an imitation of the human one.[11] In this sense, researchers are trying to build a non-biological intelligence that will have the characteristics of intelligent beings. In fact, they are trying to build machines with *cognition* that will be capable of intelligence.[12]

---

[8] "Joint Doctrine Note 2/11, The UK Approach To Unmanned Aircraft Systems," *Ministry of Defense*, accessed July 20, 2020, https://www.law.upenn.edu/live/files/3890-uk-ministry-of-defense-joint-doctrine-note-211-the.

[9] Peter Singer, *Wired for War* (New York: Penguin Press, 2009), 145.

[10] Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (New York: Knopf, 2017), 73.

[11] John Haugeland, *Artificial Intelligence: The Very Idea* (Cambridge, Massachusetts: MIT Press. Chicago, 1985), 255.

[12] Regarding the metaphysical differences between Cognition and Intelligence, see Alkis Gounaris, "Human Cognition and Artificial Intelligence: Searching for the Fundamental Differences of Meaning in the Boundaries of Metaphysics," accessed January 14, 2019, https://alkisgounaris.gr/gr/research/human-cognition-artificial-intelligence/. This is a fundamental difference which, however, is not taken into account by the majority of AI researchers who equate the two concepts. According to our position,

This position stems from the assumption that the human brain is nothing more than a biological computing machine that produces human cognition and has the ability to achieve complex goals, that is, to have intelligence. The anthropomorphic view of artificial intelligence as well as the mechanistic view of the human cognition, enframes research and discussion within defined linguistic boundaries (psychological and mechanistic vocabulary) in which we perceive and define the abillities and functions of autonomous systems.

For example, we may say that the artificial intelligence system thinks, understands etc., or that the brain performs algorithmic calculations. In these cases we use language metaphorically, borrowing terms from different scientific vocabularies, and as a result this temporary loan from one language game is established with another meaning within a different language game. As the concepts of cognition, intelligence, consciousness etc. remain cloudy, indeterminate and are used in many different ways by both philosophers and AI specialists, their ontological clarification becomes particularly complicated.[13] As a result,

---

intelligence can be defined as the ability to achieve complex goals and is inextricably linked to computational ability, and cognition is defined as the ability of the cognitive being to learn, perceive and understand, to make value judgments and decisions, to give meaning to its world, etc., i.e. processes that are not necessarily related to computing capacity.

[13] Christian De Quincey, "Switched-on Consciousness: Clarifying What It Means," *Journal of Consciousness Studies* 13, no. 4 (2006): 6-10; David Levy, "The Ethical Treatment of Artificially Conscious Robots," *International Journal of Social Robotics* 1, no. 3 (2009): 209-216; Aaron Sloman, "A Systematic Approach to Consciousness (How to Avoid Talking Nonsense?)," accessed July 28, 2020, http://www.cs.bham.ac.uk/research/projects/cogaff/misc/consciousness.rsa.text. In fact, as Hoffmann and Hahn point out, this vagueness in the definition of intelligence leads respectively to an ambiguity as to the characterization of a machine as an AI system, see Cristian Hoffmann, and Benjamin Hahn, "Decentered Ethics in the Machine Era and Guidance for AI Regulation," *AI & Society* 35, no. 3 (2020): 635-644. Indeed, it seems practically impossible to know whether to classify a machine as an "Artificial Intelligence system" without first having a

most thinkers turn to the formulation of behavioral cues and ultimately behavioral criteria of intelligence.[14] Daniel Dennett seems to follow this shift towards behavioral criteria as well, albeit in part as we shall see,[15] defending the "human" behavior of HAL 9000.

The concept of autonomy has also had comparable linguistic adventures as we will see in more detail below, since it

clear definition of the term "intelligence." In this sense, the conceptual vagueness of the term "intelligence" also leads to a vagueness regarding the definition of the borders to the set of entities to which we attribute the term "Artificial Intelligence" – and it should be emphasized that as an already first serious consequence, we can't precisely define all the technological applications that fall within the field of analysis of AI ethics. The phrasing of the Turing-Red-Flag-Law, which essentially expresses a demand that all AI systems be indeed recognizable as such, is, after all, characteristic of the severity of the whole situation, see Toby Walsh, *It's Alive: Artificial Intelligence from the Logic Piano to Killer Robots* (Hamburg: Edition Körber, 2017).

[14] The first move towards finding behavioral criteria was made by Descartes, with his suggestion of the criterion of Language as well as the criterion of successful action-in-the-world, see Gerald J. Erion, "The Cartesian Test for Automatism," *Minds and Machines* 11, no. 2 (2001): 29-39; Keith Gunderson, "Descartes, La Mettrie, Language, and Machines," *Philosophy* 39, no. 149 (1964): 193-222; Virginia Savova, and Leonid Peshkin, "Is the Turing Test Good Enough? The Fallacy of Resource-Unbounded Intelligence," *International Joint Conferences on Artificial Intelligence Organization: Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, IJCAI-07:545-550, accessed August 29, 2020, https://www.ijcai.org/Proceedings/07/Papers/086.pdf. In the 20th century, this shift to behavioral criteria was marked by Turing's introduction of the 'Imitation Game' - now known as the *Turing Test*, though Turing's intentions were diametrically opposed to those of Descartes, as the former turned to behavior in order to support an ontological equivalence of humans and machines, while the latter did so in order to support their ontological distinction, see Alan Turing, "Computing, Machinery and Intelligence," *Mind* LIX (1950): 433-660.

[15] See bellow on the criterion of excessive efficiency in this regard.

is used differently by moral philosophers and by the designers and engineers of Artificial Intelligence.[16] For Kantian moral philosophers, autonomy forms the basis of moral responsibility and the attribute of personhood[17] and is associated with free will and self-governance – namely, the possibility and the ability of the person to delimit his/her own actions. In order to have moral responsibility, a person must be autonomous or in any case free from coercion.

This means that the person should be free from external factors that can force one to act in a certain way (for example not to have a gun to their head) and not to be limited by uncontrollable internal factors that determine one's decision (for example not to be under the influence of a drug or in some uncontrollable mental state). The decision, that leads a person

[16]The 'technical' (i.e. the technological) use of the term "autonomy" usually refers to a long period of time between two consecutive energy charges, while in the case of weapon systems it means that the weapon has "fire and forget" ability, i.e. the ability to maintain focus and targeting upon the target chosen by the human operator, without the operator having to constantly intervene. On the contrary, the philosophical expression of the term "autonomy" is inextricably linked to moral responsibility and at the same time it is charged with a multitude of rich ontological contexts that, as we will see below, reach as far as the concept of cognition. It often happens that the researchers of AI start their reference to the "autonomy" of the machines in the 'technical' way but in the process, they forget about it and claim for these machines what a philosophical expression of this term would dictate. Thus, due to a misleading analogy, according to Wittgenstein, a similarity in the surface grammar of these two ways of delivering the term "autonomy," they come to support a similarity in depth grammar, that is, in meaning. We must, of course, say in advance that Dennett, whose argument we shall consider, does not make such a mistake and uses the term "autonomy" in the philosophical way. However, if his argument proves to be insufficient, the only way in which the use of this term in terms of AI systems may be possible will be in the end the 'technical' one.

[17] John Christman, "Autonomy in Moral and Political Philosophy," *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), ed. Edward N. Zalta, https://plato.stanford.edu/archives/spr2018/entries/autonomy-moral/.

to a specific act, should be determined by the person themselves in a reasonable manner.[18]

Autonomy is by definition a precondition for moral responsibility in such a way that moral responsibility entails autonomy. As Müller observes though, this relation is not inversely implied as well.[19] The term "autonomous systems" in a technical sense does not necessarily mean that these systems are morally responsible for their actions. According to this technical and weaker concept of autonomy, a mechanical system (intelligent or not) is considered autonomous in relation to its degree of control by the human factor. [20]

This weaker notion of autonomy leaves open the question of who ultimately controls the system and who bears the moral responsibility. This is the problem that in ethics is called Responsibility Gap[21] which we encounter in complex situations (e.g. in economics and business, in war, in international relations etc.) where the act in question, while it presupposes the participation of many people or bodies in an earlier stage of the act, ultimately cannot be accurately predicted or controlled in these previous stages. In autonomous AI, for example, questions are raised regarding the share of responsibility – if there is one – of programmers, developers, designers, research sponsors, the company that built the AI system etc., and even end users.

---

[18] Sarah Buss, and Andrea Westlund, "Personal Autonomy," *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), ed. Edward N. Zalta, https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/.

[19] Vincent C. Müller, "Ethics of Artificial Intelligence and Robotics," *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), ed. Edward N. Zalta, https://plato.stanford.edu/archives/fall2020/entries/ethics-ai/.

[20] Vincent C. Müller, "Autonomous Cognitive Systems in Real-World Environments: Less Control, More Flexibility and Better Interaction," *Cognitive Computation* 4, no. 3 (2012): 212-215.

[21] Regarding the Responsibility Gap in AI, see Andreas Matthias, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata," *Ethics and Information Technology* 6, no. 3 (2004): 175-183.

For Dennett, however, as we shall see below, an Artificial Intelligence system that operates autonomously and effectively can be evaluated morally like any other moral person, as long as it demonstrates intelligent behavior similar to human behavior (in specific objectives). Dennett in his now classic article entitled "When Hal Kills, Who's to Blame? Computer Ethics," which according to Sparrow, is the most serious modern philosophical defense of the position that machines could be held responsible for their actions, builds his argument by first citing the iconic chess victory of the first IBM computer, Deep Blue, over world champion Gary Kasparov in 1996.[22]

In particular, he claims that we recognize and admire the ability of the computer to win in chess and congratulate its developers for the achievement, but this victory belongs to the computer and not to the developers. If the developers faced the world champion, they would obviously lose to him in a few minutes. The responsibility of the developers for the victory of Deep Blue is equivalent to the responsibility of Kasparov's coach or teacher, but ultimately the "responsibility" for the result of the match is born by the players themselves and specifically Kasparov and Deep Blue.

Dennett's argument is extremely relevant if one considers two important AI achievements that essentially signal a future that concerns us. The first has to do with the consecutive victories in 2016 of the AI system called AlphaGo built by Google's DeepMind against Lee Sedol, world champion and one of the most important players of all time in the GO game. Sedol quit after his defeats, admitting that AI is now invincible.[23] The peculiarity of GO is that unlike chess, it relies not only on the computing ability of the players but also

---

[22] Robert Sparrow, "Killer Robots," *Journal of Applied Philosophy* 24, no. 1 (2007): 62-77.

[23] James Vincent, "Former Go Champion Beaten by DeepMind Retires after Declaring AI Invincible," *The Verge*, accessed August 1, 2020, https://www.theverge.com/2019/11/27/20985260/ai-go-alphago-lee-se-dol-retired-deepmind-defeat.

on more complex cognitive skills, with many claiming that it is actually a kind of art.[24] The second achievement is the total dominance of DeepMind's AI system in virtual air combat, in 2020, over top pilots of the United States Air Force with F16 Viper fighters.[25] The significance of this victory lies in the fact that in addition to computing skills, perception of three-dimensional space, physical skills and deceptive movements are required.

Dennett extends the reasoning for accountability proportionally, by moving from Deep Blue to HAL 9000, a heuristically programmed Algorithmic Computer[26] who is the main character of Stanley Kubrick's film *2001, A Space Odyssey*.[27] HAL has infinitely greater computing power than Deep Blue, operates "autonomously" and carries out life and death operations, since in order to ensure the success of its mission when it realizes that it is in danger, HAL decides to kill the spacecraft crew in which it was installed, and gain full control. Dennett attributes moral personality traits to HAL because this autonomous intelligent machine exhibits human

---

[24] As Tegmark points out there are far more possible positions in GO than there are atoms in the universe, which means that no computer system can analyze all the interesting sequences of future movements, see Tegmark, 114.

[25] Fabienne Lang, "AI Flawlessly Beats US Air Force F-16 Pilot in Simulated Dogfight," *Interesting Engineering*, accessed August 21, 2020, https://interestingengineering.com/ai-flawlessly-beats-us-air-force-f-16-pilot-in-simulated-dogfight.

[26] Heuristic mechanisms are computer problem-solving techniques which evaluate and select intermediate situations by rejecting the rest, in order to save time. In AI, although these techniques are algorithmically coded, they are not considered "exactly" algorithms, as algorithms always lead to accurate results, while these mechanisms more closely resemble human "intuitive" thinking and educated guess.

[27] The script of the film was based on the novel of the same name by Arthur Clarke; see Arthur C. Clarke, *2001: A Space Odyssey* (New York: New American Library, 1968).

behavior, regardless of whether it repents, feels remorse, feels, or understands what it means to be a moral person.

In our view, however, the arguments put forward by Dennett do not sufficiently prove the position that HAL can be characterized as a moral person.

## I. The Argument of Equivalence

Initially, the supposed equivalence of Kasparov's relationship with his coach and Deep Blue with its developers is not logically obvious. In particular, this equivalence can be supported in two ways:

> a) The computer is ontologically equivalent to the human athlete or
> b) The computer is not necessarily ontologically equivalent to the human athlete but the 'developer – computer' relationship is functionally equivalent to the 'coach – athlete' relationship, i.e. these two relationships can both be described in common functional terms. In other words, the study of both of these relations at a functional level can lead to an identical description: the two relations are reduced to the same set of functions performed.

In the case of a), that is, in the case where one argues that the computer is ontologically equivalent to a human athlete, the logical fallacy of a circular argument is being committed, as in the end we come to take for granted what we are trying to prove.

Related to this, to say "The responsibility of the developers for the victory of Deep Blue is equivalent to that of the coach or teacher of Kasparov," based on the assumption that the computer is ontologically equivalent to Kasparov, takes for granted what needs to be proven – i.e. this equivalence. One would expect that we would provide sufficient evidence to demonstrate this human - computer ontological equivalence, instead of simply making an affirmative statement that ends

up being essentially a tautology, hence a sentence without real "epistemological value."[28] Specifically, to say that "The victory belongs to the computer because the 'developer – computer' relationship is the same as the 'coach – athlete' relationship," and that "the 'developer – computer' relationship is the same as the 'coach – athlete' relationship, because the computer and the athlete are ontologically equivalent," is like saying "The computer and the athlete are ontologically equivalent because they are ontologically equivalent." The only way to escape this tautology is:

> a1) To finally face the problem head on, trying to answer the question: Under what criteria can we establish an ontological human – machine equivalence or distinction? This is the most central, timeless and persistent philosophical question of AI.
> a2) To try to disengage the discussion of accountability and (ultimately) moral status from the issue of the ontological human – machine equivalence or distinction. But how easy is it to separate these two in our thought? What else could provide a sufficient criterion for assigning moral status to an entity other than the ontology of the latter? Are there any examples of acceptable human thought in which the rendering of moral status and ontology were not correlated in one way or another? All moral status queries soon lead to ontology status queries.

In the case of b), that is in case we would attempt to attribute the same moral status to both Kasparov and Deep Blue on the

[28] In addition, one should explain the terms under which two entities are considered ontologically equivalent and adequately justify these terms. For example, we could suggest functionalist terms, but then we would have to justify our choice to make a functionalist description. In addition, the functionalist description will make us confront the problems discussed below in relation to b).

basis of a functionalist equivalence of the 'coach – athlete' and the 'developer – computer' relationships, we are called upon to demonstrate this very functionalist equivalence of these two relationships either

> b1) through a 'coach – developer' and 'athlete – computer' functionalist equivalence (in this case, the equivalence of the relationship of the 'coach – athlete' and the 'developer – computer' pairs is established by demonstrating the relations of equivalence of the respective members of these pairs)[29] or
> b2) because the respective members of the pairs are not functionally equivalent but the pairs that these members form, happen to be (in this case the equivalence does not lie in the members, but in the relationships they enter into with each other).[30]

Moreover, in the face of the prospect of self-programmed and self-reproducing machines, the argument based on the parallelism of developers to coaches and machine to athletes is invalidated, as the role of the human programmer becomes unnecessary.[31]

But let us look in more detail at the problems that arise from trying to prove a functionalist equivalence. Regarding b1) we must emphasize that proving a functionalist 'computer –

---

[29] For example: a-b = g-d, because a = g and b = d.

[30] For example: a ≠ g, and b ≠ d, but a-b = g-d.

[31] For an interesting analysis of the philosophical implications of the possible development of self-programming and self-reproducing machines, see John Von Neumann, *Theory of Self-Reproducing Automata* (Urbana: University of Illinois Press, 1966); Rudy Rucker, *Infinity and the Mind: The Science and Philosophy of the Infinite* (Princeton, N.J.: Princeton University Press, 1982), 157-188. The idea of self-replicating machines is not new. For one of the first technical analyses of the possibility of self-reproducing machines, see Edward F. Moore, "Artificial Living Plants," *Scientific American* 195, no. 4 (1956): 118-126.

athlete' equivalence implies proving a functionalist 'human – machine' equivalence which in turn has not been possible so far. The most well-known and organized attempt to establish an ontological equivalence of 'human – machine,' the theory of Functionalism and especially of *Machine Functionalism*, has presented serious problems, some of which are already found in the fundamental assumption of this theory, i.e. in the position that thought equals computation.

This is a position whose proof has not been reached yet, as in addition to the cloudy image we have regarding the ontology of the Mind, there are significant and well-established obstacles in the nature of any computation in which the notion of infinity is after all involved.

This very problem of the possibility of an infinite computation was pointed out by Turing himself (on whose theoretical *Machine Functionalism* is largely based) who proved that a general algorithm to solve the Halting Problem cannot exist.[32]

---

[32] Alan Turing, "On Computable Numbers With an Application to the Entscheidungsproblem," *Proceedings of the London Mathematical Society* 42, Series 2 (1937): 544-546; Alan Turing, "On Computable Numbers With an Application to the Entscheidungsproblem," *Proceedings of the London Mathematical Society* 43, Series 2 (1938): 544-546. For a comprehensible and detailed presentation of the issue of non-computability as well as for its implications regarding AI, see John L. Casti, and Werner De Pauli, *Gödel, A Life of Logic* (Cambridge: Basic Books, 2000). Also see Hubert L. Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (New York: MIT Press, 1992), chapters 5 and 10; Rucker, 157-188; John R. Lucas, "Minds, Machines and Gödel," *Philosophy* XXXVI (1961): 112-127. For the optimistic and ultimately opposite to Dreyfus and Lucas approach, see Paul Benacerraf, "God, the Devil and Gödel," *The Monist* 51 (1967): 9-32. Note that one functionalists' gambit in order to escape the impasse of non-computability, is to support the position that intelligence could be fully reproduced by a suitably "complex" Turing Machine. However, this position, apart from being analogous to the Church-Turing thesis and an unproven position, creates a new problem, as Jaegwon Kim notes, since

Still, beyond the problem of non-computability, Functionalism inevitably falls into a circular argument, as it fails to define any functions without referring to mental terms, thus it ends up trying to establish the possibility of cognition in machines while actually presupposing it.[33]

In addition to these specific logical fallacies, a functionalist attempt to prove the above equivalence, faces two major ontological problems that functionalists are called upon to solve in general. The first one is that the functionalist description ignores or fails to describe the qualitative and subjective appearances (the phenomenal aspect) of mental states, that we call qualia.

Focusing solely on the input-output relationship of a system (human, animal, machine, etc.) Functionalism leaves open a rather paradoxical possibility: Two systems may have exactly matching inputs (stimuli) and outputs (behavioral manifestations), but completely different or even inverted qualia – that is, to experience completely different or even inverted 'internal states.' It is also possible that qualia can be completely absent from one of the two systems.[34] The paradox, here, is that according to *Machine Functionalism*, these two systems are considered functionally equivalent, despite their differentiation in the level of qualia.[35]

---

functionalists are now called upon to determine what complexity is and what the appropriate complexity threshold is, beyond which a Turing Machine succeeds in demonstrating intelligence, see Jaegwon Kim, *Philosophy of Mind* (USA: Westview Press, 1998), 151-156.

[33] Kim, *Philosophy of Mind*, 153, 154.

[34] Ned Block, "Troubles with Functionalism," in *Readings in Philosophy of Psychology*, vol.1, ed. Ned Block, 268-305 (Cambridge: Harvard University Press, 1980). For the two opposing views on the possibility or non-existence of inverted or absent qualia, see David Shoemaker, "Caring, Identification, and Agency," *Ethics* 114, no. 1, (2003): 88-118; Ned Block. "Are Absent Qualia Impossible?" *Philosophical Review* 89 (1980): 257-274.

[35] At this point there have been objections from some philosophers who deny the existence or the epistemological validity of qualia during the ef-

fort of knowledge (inspection) of the Mind [for example see Paul M. Churchland, *Matter and Consciousness* (Cambridge, Massachusetts: MIT Press, 1988); Keith Frankish, "Illusionism as a Theory of Consciousness," *Journal of Consciousness Studies* 23, nos. 11-12 (2016): 11-39; Keith Frankish, *Illusionism: As a Theory of Consciousness* (Exeter: Imprint Academic Publishing, 2017); Georges Rey, "A Reason for Doubting the Existence of Consciousness," in *Consciousness and Self-Regulation*, vol. 3, eds. Richard J. Davidson, Gary E. Schwartz, and David Shapiro, 1-39 (New York: Plenum, 1983); Georges Rey, "A Question About Consciousness," in *Perspectives on Mind*, eds. Herbert R. Otto, and James A. Tuedio, 5-24 (Dordrecht: D. Reidel Publishing, 1988); Kathleen Wilkes, "Yishi, Duh, Um and Consciousness," in *Consciousness in Contemporary Science*, eds. Antony Marcel, and Edoardo Bisiach, 16-41 (Oxford: Oxford University Press, 1988)]. However, it is difficult to imagine that in the absence of qualia we could talk about the experiences of taste, smell, color, touch, etc. or even illusory experiences. Finally, it is difficult to see how we could categorize our stimuli, recognizing for example the taste or the aroma of a fruit we have eaten before [for the opposite position, see Daniel Dennett, "Quining Qualia," in *Consciousness in Contemporary Science*, eds. Antony J. Marcel, and E. Bisiach, 42-77 (New York: Oxford University Press, 1988); Valerie Hardcastle, *The Myth of Pain* (Cambridge, MA: MIT Press, 1999)]. Moreover, when it comes to the knowledge of consciousness itself, the distinction between illusion and reality collapses and therefore any critique of the epistemological validity of qualia regarding the knowledge (inspection) of the Mind becomes problematic at the very least: "Where consciousness is concerned the existence of the appearance is the reality", see John Searle, *The Mystery of Consciousness* (New York: The New York Review of Books, 1997), 122. In any case, we see here, on the occasion of the present as well as the immediately preceding footnote on qualia, that a functionalist approach to the question of attributing moral personhood in machines, such as the one attempted by Dennett, may open up many more issues than those it is coming to close. Even if Dennett opposes the existence of qualia, the issue remains open and one of the most debatable in modern philosophy, see Dennet, "Quining Qualia," 42-77. Therefore, invoking a functionalist analogy between the 'coach – athlete' and 'developer – machine' relationships would bring us face to face with this serious and still-pending philosophical ontological problem, leading to an endless

The second ontological problem focuses on whether – and if so, to what degree – does an intelligent machine or intelligent operating system *understand* or realize the meaning of the computational process and the result it produces. The most popular description of this problem has been made by John Searle in 'The Chinese Room Argument.' With this argument Searle showed that the successful syntax of physical symbols by the machine does not require the machine to understand these symbols.

Therefore, machines do not understand and in the end their implementation of a successful syntax as it takes place during the execution of an algorithm is not a demonstration of cognitive ability.[36]

The above two ontological problems that functionalists have to solve, prove that it is not self-evident that a machine that simulates human behavior is intelligent merely because it demonstrates an input-output mapping that matches that of a human in a given task. Dennett, however, a priori rejects the

---

discussion that would gravely deviate from the clarity that a criterion used to attribute moral status must have within the context of a branch of Applied Ethics such as AI Ethics.

[36] John Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3, no. 3 (1980): 414-457; John Sealre, *Minds, Brains, and Science* (Cambridge Massachusetts: Harvard University Press, 1984). Another famous argument against *Machine Functionalism* is the Multiple Realization Argument. This response is extremely interesting, as the Multiple Realization Argument was originally articulated to support Functionalism. However, understanding the way in which the Multiple Realization argument affects Functionalism and the response to it, requires an extensive reference to the structure and operation mode of the Turing Machine as well as an extensive bibliographic reference, that go beyond the main purpose and the allocated length of this article. For an overview of how Multiple Realization affects *Machine Functionalism*, see Hilary Putnam, *Representation and Reality* (Cambridge: MIT Press, 1992). For a more comprehensive analysis of the relationship between Functionalism and Multiple Realization, see Jaegwon Kim, *Supervenience and the Mind* (Cambridge: Cambridge University Press, 1993).

useful role of qualia in cognitive science and disagrees with Searle, arguing that the Chinese Room "understands" as a comprehensive system the meaning of the result it outputs, and ultimately adopts the attitude of a rational behaviorist towards the Kasparov - Deep Blue (Human - Machine) equivalence, content with the end result and the behavior of the compared entities.

This disagreement demonstrates that the functionalist approach is characterized by ontological issues that remain pending to this day. Therefore, for the time being, it does not seem to be the most appropriate for the consolidation of an easy-to-use and robust criterion in order to attribute moral status to machines. In any case, as we have seen, the substantiation of the functionalist equivalence fails already at a logical level. Therefore, we should probably go back to the need of directly addressing the basic question of AI referred to above, namely the question of the ontological equivalence or human - machine distinction and eventually to a).

Finally, regarding b2), that is, the functionalist comparison not of the members that make up the pairs 'coach – athlete' and 'developer – computer' but of the relationships that these pairs form, we must observe that already the 'coach – athlete' relationship seems to be characterized by a much higher level of freedom than the 'developer – computer' relationship. The computer's actions seem to be much more dependent on the developer's commands, than the athlete's actions bound by the commands of his coach.

In fact, in the functionalist definition of the 'developer – computer' relationship there is the program factor, which does not seem to have a functional analogy in the case of the 'coach – athlete' relationship. In addition, one could argue that the developer and the machine are involved in an endless loop of dynamic interaction and in an ongoing dialogue that simultaneously determines the actions of both.

In any case, this discussion regarding the laxity or not of the 'coach – athlete' relationship versus the 'developer – computer'

relationship, brings forth the terms of environmental programming (the environment as a developer) and ultimately of autonomous agency. These are terms that have a timeless presence in the effort to address the fundamental philosophical question of AI as to the ontological identification or distinction of human – machine.[37]

## II. The Argument of Autonomy

The fact that Dennett, among other things, invokes the autonomy of HAL in order to attribute moral responsibility to HAL, thus moral personhood, is not something new in the field of AI Ethics. Other thinkers and researchers have also linked the attribution of moral personhood to the machines with the issue of autonomy.[38] Moreover, empirical studies in the Psychology of Human-Computer Interaction indicate that the majority of people consider the ability of a machine to make choices as being one of the basic criteria for attributing moral responsibility to this machine.[39] At first glance, this connection of the attribution of moral personhood to the machines with the concept of autonomy seems quite reasonable, especially under a Kantian approach.

---

[37] At this point an intersection – or rather a common conclusion – of a) and b) is found again. It seems, therefore, that even under a functionalist attempt to bypass the direct confrontation of the ontological question of the human – machine identification or distinction – that is, even with the gambit of reducing an ontological question to functionalist terms, the basic features of the question and their impasses remain fully valid.

[38] David Calverley, "Toward a Method for Determining the Legal Status of a Conscious Machine," in *Proceedings of the AISB 2005 Symposium on Next Generation Approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment*, eds. R. Chrisley, R. Clowes, and S. Torrance, 75-84 (Hatfield: University of Hertfordshire, 2005); Sparrow, "Killer Robots," 62-77.

[39] Andrew E. Monroe, Kyle D. Dillon, and Bertram F. Malle, "Bringing Free Will Down to Earth: People's Psychological Concept of Free Will and its Role in Moral Judgment," *Consciousness and Cognition* 27 (2014): 100-108.

In the case of HAL, Dennett attempts to overcome the issue of the possible heteronomy of a programmed computer by comparing HAL with the case of a genetically or 'environmentally programmed' moral agent. If genetic programming and human experiences exempt humans from their moral responsibilities, then they should do so for HAL. At this point, we practically have the articulation of the argument that the environment is for the humans what the programmers are for the machines. According to this line of thought, one could say that even if – as shown above – it would be quite difficult if not impossible to establish a coach – programmer analogy in detailed functionalist terms, there could be at least some trainer – environment parallelism that could possibly prepare the grounds for the support of an ontological equivalence between humans and machines.

Here one could object this view by stressing the fact that Dennett overlooks an important aspect which makes the use of the term "autonomy" a metaphorical one. Specifically, it could be supported that contrary to the case of humans, each autonomous AI system integrates a certain given goal, a predefined task. For instance, it is not possible for such a machine to temporarily postpone the execution of its task in order to take a break and have a cup of coffee or read a book. Every task of an AI system is predefined, given, inescapable and extraneously determined (determined 'from the outside') in such a way that any notion of autonomy is negated. This is due not just to the fact that the machine is programmed in a certain way, but because the goal of its existence is integrated in its essence. Every machine is a 'machine for…,' namely it is built to perform a certain function in order to achieve certain goals regardless of their complexity. The conception – not to mention the construction – of an intelligent machine with no particular goals seems a real challenge for AI research.[40]

[40] It has to be noted that although one of the basic visions of the researchers in the first years of the AI scientific program was the creation of general purpose machines (the concept of the Universal Turing Machine, projects like Allen Newel's and Herbert Simon's The General Problem

Nevertheless, if we want to be really fair to Dennett, we have to ask ourselves how different are humans compared to machines, regarding the issue of a goal integrated into their existence. Do humans really come to life and grow up free from goals not chosen by them but chosen by their environment? Often, humans are nurtured, bred more or less explicitly to be given a certain purpose in their lives. The extreme cases of the 'tightly closed' priories and religious orders, the more usual cases of political youth clubs, the church and people's introduction to a system of religious faith, the training of the priests, military training and finally the more imperceptible and intangible ways of training from the environment, such as family members leading by example, gender-specific role-taking are examples of the wide variety of environmental influences over humans. But even before we consider all these, the very fact of a human's birth integrates a goal extraneous to this human, namely the choice of one's parents to bring him/her to life (in order to achieve a continuation of their name or to satisfy their parental or sexual instincts or even to satisfy the social role models, the wishes of their families etc.). Therefore, one arrives at the following question: Up to which level of environmental influence could an entity be thought of as being autonomous? In other words, which is the threshold of intervention beyond which the environmental influences are considered as programming, as a mechanism of reaching to a heteronomy of an entity's will? Which is the threshold of the extraneous intervention beyond which the entity is considered to have integrated to the essence of its existence a goal extraneous to it? At this point we seem to be asking for a quantitative criterion (specifically a threshold), since the issue of attribution of moral status is also usually dealt with a quantitative manner (we usually attribute different levels of

Solver and the cognitive architecture SOAR, are examples of this vision), such a development has not yet taken place - possibly due to ontological restrictions in the very the nature of a machine.

moral status to different entities).[41] Thus, it seems that before we are able to identify this threshold, we can't totally reject Dennett's argument of a parallelism between the environment and the programmers. Given that the humans undergo a kind of programming by their environment, absolute autonomy might not even exist for humans either. Therefore, for now, it seems that we don't have the right to support a distinction between humans and machines on the basis of an argument of goals being imposed to the machines by their human creators and programmers.

If we really want to identify a problem in the use of the criterion of autonomy we will have to shift the focus of the discussion from the machine – programmer relation to the definition and the determination of the limits of the philosophical concept of autonomy and to the way in which this concept is related to the attribution of moral status. Moreover, we will also have to focus on our ability to identify the presence of autonomy in an entity.

Previously in this text, we saw that for one to be acknowledged as an autonomous agent one must not be in a status of internal or external coercion, namely not to have a gun pointed to his/her head or not to be in a mental state that is not controlled by him/her.

However, if we want to be precise with the definition of the concept of autonomy we need to be in the position to answer the four following questions:

> 1) Which are the presuppositions of autonomous agency? Which are the features and the properties that an entity has to have in order to act as an autonomous agent? In other words, how is the concept of autonomous agency delimitated?

[41] Regarding particularly for the different levels of the attribution of moral responsibility for acts of war and specifically for the distinction between adult and children soldiers as well as for a parallelism between the latter and AI weapons see Sparrow, "Killer Robots," 62-77.

> 2) How can we identify autonomy? Which are the indications that we need to have in order to regard an entity as being an autonomous agent?[42]
> 3) Is autonomous action – especially the action of a moral agent – necessarily linked to the property of the cognitive being? In other words, is an agent's mental state (and cognition in a broader sense) a necessary condition for autonomous action?
> 4) Is the issue of autonomous agency attribution totally symmetrical to the issue of moral status attribution? Does the characterization of an entity as an autonomous agent necessarily entail that this entity can also be characterized as a moral agent?

First, we have to see that questions 1 and 2 are linked, since some of the features and the properties required for reaching autonomous agency can inform the criteria for the autonomous agency identification. For instance, if the feature F is demanded so that an entity E is truly autonomous, then a safe criterion for the identification of autonomous agency in an examined entity E would be the identification of F as a feature of E. Question 1 is an ontological question (What is the autonomous agency?) while question 2 is an epistemological question (How can we know the existence of autonomous agency?). However, frequently the answer to the epistemological question is strongly defined by the answer to the ontological question.[43]

---

[42] The determination of the features and therefore the safe indications of autonomous agency is crucial since these indications will form the basis of the ontological evaluation and classification of the entities under the question of moral status attribution. See right below, in the main text.

[43] A typical example of the connection between an ontological and an epistemological question is Thomas Reid's introduction of the 'Other Minds Problem' as a critique in the way in which Berkeley approached the concept of mind; see Anita Avramides, *Other Minds* (London, New York: Routledge, 2001), 139-180. Here it must be pointed out that apart from the concept of autonomy, this connection between the ontological and

Nevertheless, with regard to question 1, a plurality of definitions of – and finally presuppositions for – autonomous agency exists.[44] Which of all these views is the correct one? Thus, which of all these views should be the basis of the discussion regarding the attribution of moral personhood to AI systems? It seems that until now most of the researchers in the field of AI Ethics have adopted internalist approaches (in the sense that they refer to the concept of consciousness and to mental states like intentions, beliefs, emotions etc.), and therefore they approximate or they are even in complete alignment with what in the traditional field of autonomous agency analysis is known as the Coherentist View.[45] According to the Coherentist View

the epistemological question exists also with regard to any other concept that has been related to the attribution of moral personhood. It is reasonable that concepts like consciousness, cognition and intelligence have also an ontological and an epistemological question with the answer to the first affecting the answer to the latter which in its turn affects the feasibility of the ontological classification of the examined entities.

[44] Returning to the above analysis with regard to the 'environment as a programmer' argument, we have to see that this plurality of autonomous agency definitions and presuppositions affects also in a negative way our ability to identify the threshold of extraneous intervention beyond which an entity has to be considered as integrating an extraneous goal to the essence of its existence. For a review of the way in which the problem of defining the limits of the concept of autonomous agency is connected to the problem of defining the limits of the extraneous interventions see Buss, and Westlund, "Personal Autonomy."

[45] Calverley, "Toward a Method," 75-84; Manuel De Landa, *War in the Age of Intelligent Machines* (New York: Swerve Editions, 1991); David Levy, *Intimate Relationships with Artificial Partners* (Ph.D. Diss., Maastricht University, 2007); Steven Pinker, "Can a Computer Ever Be Conscious?," *US News & World Report* 123, no. 7 (1997), accessed July 28, 2020. https://stevenpinker.com/files/pinker/files/computer.pdf; Solum, "Legal Personhood," 1231-1287; Sparrow, "Killer Robots," 62-77; Steve Torrance, "Could We, Should We, Create Conscious Robots?" *Journal of Health Social and Environmental Issues* 4, no. 2 (2004): 43-46. For a detailed presentation of all the views that are until now proposed regarding autonomous agency see Buss,

an agent has control over his/her action if and only if the motive of his/her action is in coherence with some mental state representing the agent's point of view.[46] Nevertheless, different advocates of the Coherentist View propose respectively different mental states as being the proper ones for an autonomous agency. Specifically, these mental states can either be related to some long-term goals, motives and plans[47] or to emotions and mainly emotions of 'caring.'[48] This raises again the issue of the

---

and Westlund, "Personal Autonomy." Given the reasonable space limit in this text, we have decided to focus only on the Coherentist View since up to now this is the one characterizing the discussion in the field AI Ethics. The analysis of the problems or solutions that could possibly come up by examining the rest of the traditional philosophical views regarding autonomous agency could be part of a new fruitful reflection presented in a new article in the future. For the time being and for the needs of the present article, we will be confined in just mentioning that the existence of these other views increases the 'noise' in the analysis of the issue regarding the attribution of moral status to the machines.

[46] Harry Frankfurt, "Freedom of the Will and the Concept of a Person," in *The Importance of What We Care About*, ed. Harry Frankfurt, 11-25 (Cambridge: Cambridge University Press, 1988a).

[47] Gary Watson, "Free Agency," *Journal of Philosophy* 72, no. 8 (1975): 205-220; Michael Bratman, "Practical Reasoning and Weakness of the Will," *Noûs* 13, no. 2 (1979): 153-171; Michael Bratman, *Structures of Agency: Essays* (Oxford: Oxford University Press, 2007).

[48] Harry Frankfurt, "The Importance of What We Care About," in *The Importance of What We Care About*, ed. H. Frankfurt, 80-94 (Cambridge: Cambridge University Press, 1988b); Harry Frankfurt, "On Caring," in *Necessity, Volition and Love*, ed. Harry Frankfurt, 155-180 (Cambridge: Cambridge University Press, 1999); Agnieszka Jaworska, "Caring and Full Moral Standing," *Ethics* 117, no. 3 (2007a): 155-180; Agnieszka Jaworska, "Caring and Internality," *Philosophy and Phenomenological Research* 74, no. 3 (2007b): 529-568; Agnieszka Jaworska, "Caring, Minimal Autonomy, and the Limits of Liberalism," in *Naturalized Bioethics: Toward Responsible Knowing and Practice*, eds. Hilde Lindemann, Marian Verkerk, and Margaret Walker, 80-105 (Cambridge: Cambridge University Press, 2008); Shoemaker, "Caring, Identification, and Agency," 88-118.

plurality of definitions which leads to a reasonable question: Which of all these criteria is the right one? Based on which of all these proposals should we judge the autonomy of humans, animals and machines? The problem of conceptual vagueness makes its appearance once again.

Moreover, the Coherentist View is a good example to return back to the link between questions 1 and 2, since we see here the way in which our inability to come up with a definite and universally accepted answer to question 1, leads also to an inability to provide a definite answer to question 2. Specifically, the plurality of the mental states proposed under the Coherentist View as being the decisive features of autonomous agency – hence the plurality of answers to question 1 – delivers a fatal strike to our chances of reaching to an unambiguous and final answer regarding question 2: How can we know which mental states should we seek to identify in an entity under examination in order to consider this entity as autonomous and therefore qualified for an attribution of moral personhood?

Besides its conceptual vagueness, the problem of the identification of mental states in other entities brings us directly against one of the most central problems in the Philosophy of Mind: The Other Minds Problem. How can we verify the existence of mental states in the entities that surround us? In fact, this question is actually divided into the following two questions:

> a) How can we know whether other beings around us have any mental states at all?, and
> b) If they do have mental states, how can we know the content of these mental states?[49]

In trying to approach the issues of attribution of moral personhood to machines through the Coherentist View of autonomous agency we are faced with the following appearances of the Other Minds Problem: How can we know whether a

[49] Avramides, *Other Minds*, 1.

machine has any mental states and especially mental states of the kind that is related to a point of view of the machine itself? How can we know if a machine has motives and plans and if these motives and plans are for the long-term? How can we know whether a machine has emotions and whether these emotions are related to 'caring?'[50]

[50] Of course, we have to mention that, apart from the Coherentist View, the Other Minds Problem is also an obstacle for any other internalist approach of the issue of moral personhood attribution, even for the approaches that do not refer to the criterion of autonomous agency. We can briefly refer here to a trend within the AI Ethics field that examines the issue of moral personhood attribution to the machines through the concept of patiency, see Tom Regan, *The Case for Animal Rights* (Berkeley & Los Angeles: University of California Press, 1983); Mane Hajdin, *The Boundaries of Moral Discourse* (Chicago: Loyola University Press, 1994); Hoffmann, and Hahn, "Decentered Ethics," 635-644; Luciano Floridi, and J.W. Sanders, "On the Morality of Artificial Agents," *Minds and Machines* 14 (2004): 349-379; Levy, "The Ethical Treatment," 209-216; Wallach, and Allen, *Moral Machines*. This is a concept which in its turn is usually linked to the concept of sentience. The latter was introduced for the first time as a criterion for the attribution of moral status in non-human entities by Peter Singer and with reference to the animals [see Peter Singer, *Animal Liberation: A New Ethics for Our Treatment of Animals* (New York: New York Review of Books, 1975); Peter Singer, *Practical Ethics* (Cambridge: Cambridge University Press, 1993)], but now it has been also introduced to the discussion regarding the AI systems, see Levy, "The Ethical Treatment," 209-216; Jonathan Owen, and Richard Osley, "Bill of Rights for Abused Robots: Experts Draw up an Ethical Charter to Prevent Humans Exploiting Machines," *The Independent*, last modified April 1, 2007, https://www.independent.co.uk/news/science/bill-of-rights-for-abused-robots-5332596.html. The basic line of thought regarding the concept of moral patiency supports the view that if AI systems and especially robots are sentient-thus capable of suffering-they should possibly be thought of as victims. However, a question arises of whether we could ever be able to know if machines actually suffer. Indeed, some AI Ethics researchers have started to note the obstacle of the 'Other Minds Problem,' see David Gunkel, *The Machine Question: Critical Perspectives on AI, Robots and Ethics*

At this point, we would like to stress how crucial the answer to question 2 is, when we work in the context of Applied Ethics where sound and practical ontological criteria are required, which, in turn, will lead to sound and handy criteria of moral status attribution in all the grades and shades of the latter. Thus, we would say that until now the treatment of question 1 has not yet led to results really useful for the treatment of question 2. In other words, the question 1 is until now approached in a way that is non-productive for the demands and the needs of Applied Ethics (in this case of AI Ethics).

Due to the dead end in which one is led when confronting the Other Minds Problem, a possible strategy could be an attempt to bypass this problem and examine the autonomous agency criterion irrespectively of any reference to mental states. Such a strategy though would bring forth question 3 ('Is the autonomous action – especially the action of a moral agent – *necessarily* linked to the property of the cognitive being?').

Let us think, for instance, a vehicle with a damaged navigation system, a conventional car with a failing steering rack or with broken brakes. Can we support the view that this vehicle exhibits a kind of autonomy in the sense that its action is not controlled by the driver?[51] It is true that usually we are

(Cambridge, MA: MIT Press, 2012); Hoffmann and Hahn; Levy, "The Ethical Treatment," 209-216.

[51] Regarding this example, one could say that this car is indeed not controlled any more by its human-driver but is now fully under the deterministic laws of nature that totally define its movement. Therefore, not being controlled by its human-driver does not necessarily mean an autonomous agency. Under an extreme naturalistic approach one could support the view that this is also the case with the human-driver. The driver is also under the deterministic laws of nature. Thus, a denial of an entity's autonomous agency on the basis of a reference to the laws of nature could be also applied to the case of humans thus striking the idea of human autonomy too. On the other hand, this would be a maneuver fatal for the whole project of Ethics, thus a maneuver that would violently interrupt and end once and for all the whole present discussion (and search for solutions

not tempted to think that the uncontrolled movement of the car is similar to the autonomy that we think that characterizes the humans. This is due to the fact that what we are looking for here is a *certain type* of autonomy; an autonomy linked to a *certain type* of agency.[52] Which is the essential characteristic of this agency? Why don't we even think to raise the question of attributing this type of autonomous agency to an uncontrolled conventional car that moves with its brakes broken but we do so in the case of a 'smart' vehicle, a computer and above all a human?

Possibly because contrary to the case of the uncontrolled conventional car, in the case of the human we have a priori accepted the property of the cognitive being and in the case of the computer or the 'clever' car there is at least a *suspicion* thus a still open possibility of cognition.[53]

---

in the field of AI Ethics and Applied Ethics in general) not by providing answers to the questions raised but by negating the whole context within which these questions are born and raised. Nevertheless, in the preset analysis we adopt a compatibilist view supporting that the natural laws and the criteria of moral status attribution belong to discrete conceptual fields (namely the ontological and the evaluative).

[52] At this point, recall the described above difference between the philosophical and the 'technical' (technological) use of the term "autonomy."

[53] Although regarding the humans we have definitely accepted the property of the cognitive being which, of course, also implies intelligence, this is not the case with the 'smart' machines. For them the question of cognition remains open even though we answer positively regarding their ability to present intelligence (even in various levels). On the contrary, in the case of a heteronomous machine like the conventional car we a priori answer negatively both for the property of cognitive being and the ability of intelligence. Therefore, it seems that we have three levels in the attribution of the property of the cognitive being and the AI systems are placed in a middle ground (some prefer to call it a 'grey area') somewhere in between the full attribution of the property of the cognitive being (the case of humans) and the total rejection of this possibility (the case of conventional machines). We would like to stress here that the AI systems are not placed towards the negative end together with the rest of the machines due to

It seems, then, that in general we accept that the autonomous agency can't be but a *cognition-related* agency. Therefore, with regard to question 3 ('Is the autonomous action – especially the action of a moral agent – *necessarily* linked to the property of the cognitive being?'), we would answer that based on the dominant views in the fields of Ethics and AI Ethics (but also on the dominant views in our everyday life) the autonomous action *is indeed necessarily linked* to the concept of cognition. However, this concept is not treated in a uniform and unambiguous way, as an 'all or nothing' feature but rather as something that presents quantitative and qualitative variations.[54] Hence, there are still cases of human beings to which we deny the attribution of autonomous agency and therefore the attribution of a full-fledged or at least a partial moral status. Infants, certain categories of mental patients, humans in a comatose or vegetative state are only some of the cases of human beings for which we find it difficult to reach universally accepted and final answers regarding the attribution of cognitive agency and finally of moral status. Consequently, although we think of the autonomous agency as necessarily linked to cognition, the latter seems to be characterized by many different levels and instances which finally lead to speculation on and questioning of the need for adopting different levels in the attribution of moral personhood via the criterion of autonomy. Thus, we would like to complete our answer to question 3 ('Is the autonomous action – especially the action of a moral agent *necessarily* linked to the property of the cognitive being?') as follows: Based on the currently dominant views in the field of AI Ethics and Applied Ethics in general, the autonomous agency is necessarily linked to the property of cognition, but given the quantitative and qualitative differences that we acknowledge in the latter, this

---

the fact that they 'behave' (or behave?) in an intelligent way which creates a suspicion that this could be something more: a cognitive way. Could we ever manage to have something more than just a simple suspicion?

[54] The problem of conceptual vagueness comes forth again here with regard to the concept of cognition.

necessary link leads to a *non-unequivocal* correlation between the autonomous agency and the attribution of moral status. In the end, considering all the above analysis with regard to the Other Minds Problem, this unbreakable link between the autonomous agency and the cognition bequeaths to the first with all the conceptual, ontological and epistemological problems of the latter. As a result of this, the autonomous agency becomes a criterion quite difficult to use for the attribution of moral personhood.

Here, it is also worth mentioning that – at least under the Coherentist View – the autonomous agency becomes difficult and problematic to use as a criterion due to its connection with some other concepts. Specifically, the coherentist account constitutes a point of intersection between the discussion for autonomous agency and the traditional and arduous reflections regarding the concept of the person. This happens in three ways: (i) The demand for the existence of goals and mental states under the point of view of an agent is equivalent to the demand for a delimitation of a *personal* point of view (ii) The existence of long-term goals, plans and motives as the essential features of autonomous agency presupposes the "diachronic unity" of this personal point of view. Thus, it presupposes the continuity, the survival through time of the agent's identity, therefore the survival of *the same* person.[55] (iii) The acknowledgement of the

[55] At this point it becomes obvious that especially the version of the Coherentist View which proposes the long-term intentional mental states as essential for the autonomous agency asks for a "psychological continuity" which is equivalent to the psychological consistency needed for the preservation of the 'sense of the self' and finally of the person's identity. Here, the discussion for the delimitation of the concept of the autonomous agency overlaps with the problems of the preservation in time of the property and of the identity of the person. In other words, this coherentist account of the autonomous agency brings us against what is known as the 'persistence' and the 'characterization question' of personhood. For a detailed analysis of these two problems see Eric T. Olson, "Personal Identity," *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), ed.

long-term or short-term intentional mental states as being the essential ones for the attribution of autonomous agency might raise a question regarding the delimitation of the agent's will . Which manifestations of intention are thought of as extrinsic to the will, (i.e. as extraneous, as coming from outside the will and imposed on it) and which as intrinsic, namely as pure products of the will? Are there any completely intrinsic intentions? Which is the limit of distinction between the intrinsic and the extrinsic intentions? In other words, what is the border that distinguishes a person from the surrounding world? Moreover, can our impulses, our short-lived and very short-term strong desires be thought of as products of our will? Finally, are our personality traits endogenous or exogenous factors with regard to our will? This conjunction of the issue of autonomous agency with the question regarding the concept of the person is an example of the way in which the philosophical analysis and the conjunction of different concepts leads to an increase rather than a decrease of the philosophical problems, since any new concept (e.g. "person") that is introduced to help us clarify a previous concept (e.g. "autonomous agency") brings with it its own problems of delimitation.

The issue of the delimitation of the will and the inclusion (or not) of the impulses and the very short-term strong desires, brings forth question 4 as well ('Is the issue of autonomous agency attribution totally symmetrical to the issue of moral status attribution? Does the characterization of an entity as an autonomous agent necessarily entail that this entity can be also characterized as a moral agent?').

As one can easily see by looking to the relevant bibliography as well as from our everyday practice, different views regarding the limits of the will lead to respectively different answers to the above question. For example, we usually don't attribute full autonomy to drug addicts. As a consequence of this, we also don't attribute to them full moral status. The discussion

Edward N. Zalta, https://plato.stanford.edu/archives/fall2019/entries/identity-personal/.

over the limits of their moral responsibility has proven to be quite long and arduous. Thus, in the case of drug addicts, the autonomy - moral status relation seems to be symmetrical, namely the negation of full autonomy leads to a negation of a full moral status. Contrary to this, in other cases, for instance in the case of people that have undergone brainwashing or indoctrination, we usually don't attribute autonomy (the traditional bibliography on the issue of autonomous agency is quite clear with this) but we usually do attribute a moral status (for instance moral responsibility for their actions even if those were dictated by their indoctrination). Namely, while according to most of the philosophical accounts of autonomous agency these people are not considered to be fully autonomous agents, they are nevertheless acknowledged to have a full moral status. In this case the autonomy-moral status relation seems non symmetrical, since the negation of autonomy has not led to a respective negation of moral status. We see then that the symmetry of the relation between autonomy and moral status changes on a case-by-case basis; a fact that makes the use of the autonomy criterion even more problematic.

In conclusion, we would say that Dennett's invocation of autonomy does not provide his argument with robustness and clarity. Autonomy is a criterion that for now is characterized by conceptual vagueness – thus by ontological ambiguity – but also by epistemological difficulties due to its correlation (at least under the most popular in the field of AI Ethics trend of the coherentist approach) with the concepts of cognition and personhood.

## III. The Argument of Excessive Effectiveness

In his attempt to ground even more convincingly his argument in favor of attributing moral responsibilities to AI systems, Dennett supports the view that we recognize and admire the skill and the ability of the computer (i.e. Deep Blue) to win in chess and we congratulate its programmers for the achievement, but the victory belongs to the computer itself and not to its

programmers. *If the latter faced the world champion in chess (i.e. Kasparov), they would obviously lose within minutes*. At this point, Dennett seems to articulate an argument based on the excessive effectiveness of Deep Blue. This computer has indeed proven to be extremely effective in chess and of course it has been proven much more effective than its programmers (and most of the humans). According to Dennett, this effectiveness superiority of the computer over its human-programmers constitutes a sufficient reason for attributing the victory to the first and not to the latter. Could this specific argumentation by Dennett open the path for a successful answer to the responsibility gap question? Namely, could excessive effectiveness constitute a sound, sufficient and universally accepted criterion for the attribution of moral status – in this case, moral rights – to AI entities and even more generally to acting entities around us (humans, animals, machines etc.)? This possibility calls for an examination of the following question: Has, until today, existed any successful application of the excessive effectiveness criterion to humans, to animals, or to machines?

As seen at the beginning of the present article, Max Tegmark and Stuart Russell also refer to the criterion of excessive effectiveness, in this case in order to appeal for a limitation or even a prohibition of AI weapons. They do so by comparing AI weapons with weapons of mass destruction and stressing on their similarity in terms of their excessive effectiveness to kill. With the occasion of this appeal a question comes up: How come we don't attribute moral responsibility to nuclear or chemical weapons on the basis of their excessive effectiveness like Dennett suggests us to do in the case of Deep Blue? Both this supercomputer and the weapons of mass destruction present excessive effectiveness. Confining the discussion only to the level of effectiveness, we see that if Deep Blue is much more effective than its human-creators in winning a game of chess, the nuclear and the chemical weapons are similarly much more effective than their human-creators in killing. So, why hasn't until now any argument been articulated in favor of

a moral responsibility attribution to these weapons like it has been for Deep Blue? The above appeal by Tegmark and Russell equates the weapons of mass destruction with the AI systems (in this case AI weapons) on the basis of an analogous risk which in its turn implies an analogous excessive effectiveness. If the effectiveness superiority over the human-creators is analogous in the cases of Deep Blue and the weapons of mass destruction, why are we not ready to open a similar discussion for the attribution of moral status to the weapons of mass destruction like Dennett does with regard to the attribution of moral status to Deep Blue? It seems that the excessive effectiveness criterion is not applied in a consistent way to the machines.

At this point, one could answer that contrary to Deep Blue and most AI systems, weapons of mass destruction do not perform in an intelligent – or at least an intelligent-like way – and therefore there is not any issue of attributing moral responsibility to the latter.[56] However, we must point out that with such an argument: A) One has to define what does one mean with the term "intelligent" (or "intelligent-like") and thus one will again need to directly face the problem of defining the limits of the concept

[56] Of course, here we need to stress that nowadays most weapons of mass destruction are navigated and controlled by AI systems. Therefore, AI is now an integral part of weapons of mass destruction to the point that the latter can be classified as AI weapons. So the distinction between AI systems and weapons of mass destruction is no longer standing in practice. However, for the sake of the above discussion, let us assume that the weapons of mass destruction do not have AI features and belong to another class of machines. The very reference by Tegmark and Russell treats them in exactly this way in order to achieve the wanted comparison – and finally correlation – with AI weapons, not in the basis that the weapons of mass destruction employ AI but on the basis of an analogous risk. Moreover, if we prefer, we can confine our analysis and refer only to the older generation of weapons of mass destruction, for instance, to the first atomic bombs that were conventional bombs not having even the simplest system of guidance.

"intelligence" as well as "cognition."[57, 58] Thus, we seem to have here a behavioral criterion that not only fails to relieve us of the arduous problem of the delimitation of cognition, but it actually throws us back to it.[59] B) One diverts the whole analysis from the

[57] Already the distinction between an "intelligent" and an "intelligent-like" way again brings forth the Chinese Room Argument and the possibility of a simple imitation of intelligent behavior. In other words, it brings us against traditional questions of the Philosophy of Mind that we tried to bypass by introducing the excessive effectiveness criterion.

[58] For the differences between the terms "cognition" and "intelligence" see footnote 12.

[59] This can be easily seen form the fact that the above line of arguments and counter-arguments leads the advocate of Dennett's position to a circular argument and finally to a tautology. Specifically, Dennett's initial argument can be expressed with the following abstract statement: "We must attribute the moral responsibility of an action A to an entity E, if E performs A with excessive effectiveness." In order to face the counter-argument that entities to which we usually don't attribute moral responsibility also present an analogous excessive effectiveness, the above argument was rephrased as follows: "We must attribute the moral responsibility of an action A to an entity E, if E performs A with an intelligent (or intelligent-like) way [and with an excessive effectiveness]." However, in any case, even on the level of an everyday naïve psychology, the attribution of moral responsibility to an entity implies that this entity is intelligent (e.g. it is characterized by mental states of an intentional character). Thus, we have to ask: What more is added here (compared to the naïve psychology approach) with the criterion of excessive effectiveness? The above last version of Dennett's argument could be finally expressed as follows: "An entity E is intelligent if it acts in an intelligent (or intelligent-like) way." At this point we have to see that if we choose the version with the term "intelligent" we end up with a tautology (even if we distinguish between the terms "intelligence" and "cognition," the epistemological value of the above sentence can't surpass that of a tautology, since intelligence is a sub-set of cognition). On the other hand, if we choose the version with the term "intelligent-like," we avoid expressing a tautology, but we are confronted with the Chinese Room Argument. Note also that in the last abstracted version of Dennett's argument any reference to the excessive

criterion of the excessive effectiveness by introducing one more term – that of the intelligent (or the intelligent-like) way which after all seems to be finally more sufficient and decisive from the one that we initially tried to uphold (i.e. excessive effectiveness). In the end, if what distinguishes the AI systems from mass destruction weapons is the intelligent (or intelligent-like) way of their action, then what reasons do we have to refer to the criterion of the excessive effectiveness? The focus of our analysis has now been definitely shifted towards another criterion and any reference to the excessive effectiveness now seems redundant. In fact, if we carefully examine the way in which the above arguments were juxtaposed, the excessive effectiveness seems to be more of an element of similarity rather than of distinction between the AI systems and the conventional weapons of mass destruction. After all, it was this justified remark regarding the similar effectiveness that led to the adoption of the criterion of the intelligent (or intelligent-like) way in the first place.

Carrying on with our analysis regarding the application of the excessive effectiveness criterion, let us now, for the sake of the conversation, bypass the problem of defining what 'an intelligent way' is. Let us see that the inconsistent use of the excessive effectiveness criterion can be revealed even if our focus is confined only to the set of machines that we call "AI systems." Specifically, although an attribution of moral status is proposed for a super-computer like HAL 9000, this is not also the case with AI weapons. Excessive effectiveness is a feature that characterizes both the first and the latter. So, why do we start up a discussion of moral status attribution only for HAL? Which is the distinctive difference, the *differentia specifica* between them? Is it that HAL participates in a predominantly human activity as a member of a space expedition, while the weapons of mass destruction are not (killing is not an activity

---

effectiveness is completely missing, so it seems that the criterion for the attribution of moral status has been shifted from the concept of excessive effectiveness to the concepts of intelligence and cognition (see in the main text).

characteristic only of humans)? Nevertheless, we would answer that with this argument:

> A. One substitutes again the excessive effectiveness criterion with another criterion, namely the criterion of the field of human action.
> B. One accepts a delimitation of the term "cognition" that coincides exclusively with the delimitation of the term "human action." Therefore, one denies tacitly the attribution of the property of cognitive being to animals, an issue that is still debated and for which many of those who would like to deny the moral status of AI weapons answer positively supporting the possibility of animal moral rights.
> C. Therefore, we see that the inconsistency in the use of the excessive effectiveness criterion remains, even if we confine the discussion within the set of the AI systems.

Things are no better concerning the application of this criterion to humans. It is widely accepted and verified in practice that the human kind presents a remarkable diversity of skills which in any case are not distributed in a uniform way. People vary regarding their special abilities, their 'talents' as well as their weaknesses. However, we usually try not to have a similarly diverse attribution of moral status to them, although we don't always succeed in this task. Quite often people are considered morally responsible for their actions in fields in which they don't present an excessive effectiveness, whereas there are cases in which a mitigation of moral responsibility is attempted for actions in which people do present such effectiveness. A typical example of this is the case of people who have suffered damage in brain areas related to the triggering and the control of the so-called pro-social emotions. Usually, such individuals end up becoming serial killers and mass murderers since they combine a high level capacity to plan a murder – therefore an excessive

effectiveness of executing it – with a lack of moral restraints.[60] These people are frequently addressed as mental patients, thus as individuals having a reduced autonomy due to their mental illness. Eventually, we see that not only the attribution of moral status is not symmetrical to the attribution of excessive effectiveness (namely, the delimitation of moral status is not univocally related with the delimitation of any excessive effectiveness), but also it is rather based on other criteria like the criterion of autonomy (which brings us back to the previous discussion regarding the problems of the autonomous agency). So, if in the case of humans we avoid linking effectiveness to the attribution of moral status, why should we do so in the case of the machines?

In fact, in the case of machines – as well as animals – excessive effectiveness has been used sometimes as an indication of a non-cognitive, 'automatic' nature and, therefore, of a nature inferior to that of humans, and some other times as a proof of these entities' moral or cognitive superiority over humans. Respectively, René Descartes was the first who supported the view that an exhibition of an excessive effectiveness in certain actions on behalf of an entity is a safe indication – and thus a sound behavioral criterion – of the automatic nature of this entity.[61] For Descartes, the "automata" (animals and machines) function not based on rational mind but completely based on the specificities in the structure of their body. Therefore, they present an excessive effectiveness in certain areas of action due to the specific structure of what is nowadays called "hardware.""[62] This is a position that has also been adopted by

[60] Clare Allely et. al., "Neurodevelopmental and Psychosocial Risk Factors in Serial Killers and Mass Murderers," *Aggression and Violent Behavior* 19 (2014): 288-301.

[61] René Descartes, "Letter to the Marquess of Newcastle," in *The Philosophical Writings of Descartes: Volume 3*, trans. John Cottingham, Robert Stoothoff, and Murdoch Dugald, 302-304 (New York: Cambridge University Press, 2004), 304.

[62] René Descartes, "Discourse of the Method of Rightly Conducting One's

some modern researchers in the field of AI, in their effort to come up with a solid behavioral criterion for a safe judgment regarding cognition in the Turing Test context.[63] Here, the 'hardware' specificities have been substituted by what we could refer to as 'software' specificities, namely the specialization of the machine's program. A completely opposite use of the excessive effectiveness criterion has been made by a philosopher almost contemporary to Descartes, specifically by Michel de Montaigne in his support of the doctrine of *Theriophily*. According to Montaigne, the fact that the animals exhibit a remarkable and quite higher than humans effectiveness in certain actions, constitutes a sufficient proof of the animals' superiority over the humans and finally of the animals' right to have a moral status fully respected by the humans. [64]

Therefore, it seems that the philosophical analysis has not yet settled down with regard to the relation between excessive effectiveness and the attribution of cognitive abilities or finally the attribution of an ontological status that would be also related to a moral personhood attribution. On the contrary, up to now, the discussion is characterized from completely opposite ways of using the excessive effectiveness criterion. To the extend that moral responsibility is related to cognition, we could say that Dennett's view that excessive effectiveness constitutes a sufficient reason for attributing the win to Deep Blue and moral responsibility to HAL is diametrically opposite to the view of Descartes and to

---

Reason and of Seeking Truth in the Sciences," in *The Philosophical Writings of Descartes: Volume 1*, trans. by John Cottingham, Robert Stoothoff, and Murdoch Dugald, 11-151 (New York: Cambridge University Press, 2004), 141.

[63] Donald Michie, "Turing's Test and Conscious Thought," in *Machines and Thought. The Legacy of Alan Turing*, vol. 1, eds. Peter Millican, and Andy Clark, 27-51 (Oxford, New York: Oxford University Press, 2002).

[64] Michel de Montaigne, "An Apology for Raymond Sebond," in *Michel de Montaigne: The Complete Essays*, trans. Michael A. Screech, 489-683 (London, New York: Penguin Books, 2003).

the view of those modern AI researchers who try to ground the Turing Test on a correlation between excessive effectiveness and the total absence of cognitive abilities. According to the approach made by Descartes and all those who treat excessive effectiveness as an indication of an entity's 'automatic nature,' Deep Blue should never have the right to be attributed with the victory in a chess game. On the other hand, according to Dennett, excessive effectiveness constitutes a sufficient reason for attributing the victory to Deep Blue and moral responsibility to HAL. One could possibly support the view that Dennett's position seems more compatible with that by Montaigne. However, opposite to what Montaigne supports regarding the animals, Dennett does not use the excessive effectiveness criterion to support a superiority of Deep Blue and HAL over the humans. He rather argues for an equivalence between these super-computers and the humans. Under a rough description, we could say that until now we have three different uses of the excessive effectiveness criterion on behalf of the philosophers:

> 1) Descartes' use of excessive effectiveness as an evidence of other beings' (animals and machines) inferiority compared to the humans
> 2) Montaigne's use of excessive effectiveness as an evidence of the superiority of other beings (animals) over the humans
> 3) Dennett's use of excessive effectiveness as an evidence of equity between other beings (machines) and the humans.

Which of these uses is the correct one? For now, the only safe claim we can make is that the excessive effectiveness criterion is not being used with a constant, univocal and thus consistent way for the ontological comparison of humans with other entities. This inconsistency leads logically to a respective non-univocal and non-consistent use of the excessive effectiveness criterion for the attribution of moral status to these entities.

## IV. Conclusions

In this article we set out to examine the possibility of attributing moral personhood to AI systems. Our analysis focused exclusively on AI weapons, and this due to the severity of the consequences their use may result in; this severity is proportional to the sharpness and the intensity of the ethical issues that this use raises. In other words, we referred specifically to the case of AI weapons because it constitutes the most pressing of all the contexts in which the philosophers and the AI researchers find themselves confronted with the problem of moral status attribution to AI entities. Nevertheless, we think that the arguments and the conclusions that we have presented in the above text have a rather general validity – namely, they can be applied to any machine characterized as an "AI system" – since they are not based on aspects that are specific only to AI weapons. On the contrary, they can also apply to any other machine. Moreover, we have chosen to base our analysis on a scepticist response to the arguments supported by Daniel Dennett in his text *When HAL Kills, Who's to Blame? Computer Ethics*, which is considered to be a milestone of contemporary philosophical analysis in favor of the attribution of moral status to the machines. After all, the reference to HAL and to the murder that this system commits in the famous film *2001: A Space Odyssey* makes Dennett's analysis relevant to the ethical issues raised regarding AI weapons.

Specifically, we supported that Dennett's analysis is mainly based on three basic arguments: The analogy between the programmer – machine and the coach – athlete relations, the machine autonomy argument and the argument of excessive effectiveness (the last two as sufficient criteria for the attribution of moral status to an AI system).

With regard to the first argument, we showed that the support of an analogy in the programmer – machine and trainer – athlete relations as an argument in favor of the machine moral status is already a logical fallacy. First, because it constitutes a

circular argument given that it assumes the conclusion or in other words it presupposes what is to be proven, namely the ontological equivalence between the human and the machine. Second, because in the case that one considers this analogy as a functionalist one, one is confronted with the logical problems inherent in the foundations of *Machine Functionalism* as well. In addition, this view of a functionalist analogy faces also the ontological problems of *Machine Functionalism*.

Concerning the argument of machine autonomy, we initially observed that Dennett's programmers-environment parallelism, thus challenging the unconditional, absolute human autonomy, is not totally groundless. However, we showed that Dennett's appeal to the criterion of autonomy faces the problem of conceptual vagueness which is raised by a plurality of autonomy definitions. Moreover, according to the most popular – at least in the field of AI Ethics – autonomy account, namely, according to the internalist view, one is inevitably confronted with the Other Minds Problem and also with certain well-known and traditional problems regarding the property of personhood like the 'persistence' and the 'characterization problem.' Moreover, we showed that the appeal to the criterion of autonomy pits one's analysis against the ambiguity of the delimitation of the will. It is this ambiguity that leads to a non-symmetry in the relation between the attribution of autonomy and the attribution of moral personhood, namely to the inconsistent use of the criterion of autonomy.

Until now, the use of the excessive effectiveness criterion has been proven to be similarly inconsistent, both regarding the human – machine and the human – animal distinction. After all, the application of this criterion seems to take place with an arbitrarily selective way not only regarding AI weapons, but also other machines like the weapons of mass destruction.

We think that our counter-arguments presented above respond to a large part of the contemporary discussion regarding the literal attribution of moral status – and thus of the property of moral personhood – to AI systems and especially

to AI war machines. We support the view that for now and until the ontological and epistemological issues related to human cognition and artificial intelligence are resolved in a satisfying way, any discussion towards this direction can be made only with a metaphorical use of the words "autonomy," "personhood," and "moral status." Besides, we should not overlook the fact that the exhibition of morally relevant actions (actions that can be morally evaluated) on behalf of the machines is something completely different from the attribution of moral responsibility to the machines for their actions.[65]

Dennett's view is in favor of the attribution of moral status to the AI systems. Our present analysis did not aim at supporting the opposite view, namely a view against the attribution of moral status to these systems. It rather aimed at demonstrating the fact that based on the dominant current argumentation in the field of AI Ethics, the question regarding the attribution of moral status to the machines can only remain *undecidable*. Thus, we are once again confronted with a contradiction well known to anyone working in the field of Applied Ethics, specifically with the contradiction between the demand for clear and sound moral decision criteria and the interminable nature of a philosophical contemplation that tries to be consistent.

## References

Allely, Clare, Helen Minnis, Lucy Thompson, Philip Wilson, and Christopher Gilberg. "Neurodevelopmental and Psychosocial Risk Factors in Serial Killers and Mass Murderers." *Aggression and Violent Behavior* 19 (2014): 288-301.

Anderson, Michael, and Suzan L. Anderson. "Machine Ethics: Creating an Ethical Intelligent Agent." *AI Magazine* 28, no. 4 (2007): 15-26.

Avramides, Anita. *Other Minds*. London, New York: Routledge, 2001.

[65] Michael Anderson, and Susan L. Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent," *AI Magazine* 28, no. 4 (2007): 19.

Benacerraf, Paul. "God, the Devil and Gödel." *The Monist* 51 (1967): 9-32.

Block, Ned. "Are Absent Qualia Impossible?" *Philosophical Review* 89 (1980): 257-274.

Block, Ned. "Troubles with Functionalism." In *Readings in Philosophy of Psychology*, vol. 1, edited by Ned Block, 268-305. Cambridge: Harvard University Press, 1980.

Bratman, Michael. "Practical Reasoning and Weakness of the Will." *Noûs* 13, no. 2 (1979): 153-171.

Bratman, Michael. *Structures of Agency: Essays*. Oxford: Oxford University Press, 2007.

Buss, Sarah, and Andrea Westlund. "Personal Autonomy." *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition). Edited by Edward N. Zalta. https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/.

Calverley, David. "Toward a Method for Determining the Legal Status of a Conscious Machine." In *Proceedings of the AISB 2005 Symposium on Next Generation Approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment*, edited by R. Chrisley, R. Clowes, and S. Torrance, 75-84. Hatfield: University of Hertfordshire, 2005.

Casti, John L., and Werner De Pauli. *Gödel, a Life of Logic*. Cambridge: Basic Books, 2000.

Christman, John, "Autonomy in Moral and Political Philosophy." *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition). Edited by Edward N. Zalta. https://plato.stanford.edu/archives/spr2018/entries/autonomy-moral/.

Churchland, Paul M. *Matter and Consciousness*. Cambridge, Massachusetts: MIT Press, 1988.

Clarke, Arthur C. *2001: A Space Odyssey*. New York: New American Library, 1968.

De Landa, Manuel. *War in the Age of Intelligent Machines*. New York: Swerve Editions, 1991.

De Quincey, Christian. "Switched-on Consciousness: Clarifying What it Means." *Journal of Consciousness Studies* 13, no. 4 (2006): 6-10.

Dennett, Daniel. "Quining Qualia." In *Consciousness in Contemporary Science*, edited by Antony J. Marcel, and E. Bisiach, 42-77. New York: Oxford University Press, 1988.

Dennett, Daniel. "When Hal Kills, Who's to Blame? Computer Ethics." In *Hal's Legacy: 2001's Computer as Dream and Reality*, edited by David G. Stork, 351-365. Cambridge, MA: MIT Press, 1997.

Descartes, René. "Discourse of the Method of Rightly Conducting One's Reason and of Seeking Truth in the Sciences." In *The Philosophical Writings of Descartes: Volume 1*, translated by John Cottingham, Robert Stoothoff, and Murdoch Dugald, 111-151. New York: Cambridge University Press, 2004.

Descartes, René. "Letter to the Marquess of Newcastle." In *The Philosophical Writings of Descartes: Volume 3*, translated by John Cottingham, Robert Stoothoff, and Murdoch Dugald, 302-304. New York: Cambridge University Press, 2004.

Dreyfus, Hubert L. *What Computers Still Can't Do: A Critique of Artificial Reason*. New York: MIT Press, 1992.

Erion, Gerald J. "The Cartesian test for Automatism." *Minds and Machines* 11, no. 2 (2001): 29-39.

Floridi, Luciano, and J.W. Sanders. "On the Morality of Artificial Agents." *Minds and Machines* 14 (2004): 349-379.

Frankfurt, Harry. "Freedom of the Will and the Concept of a Person." In *The Importance of What We Care About*, edited by Harry Frankfurt, 11-25. Cambridge: Cambridge University Press, 1988a.

Frankfurt, Harry. "On Caring." In *Necessity, Volition and Love*, edited by Harry Frankfurt, 155-180. Cambridge: Cambridge University Press, 1999.

Frankfurt, Harry. "The Importance of What We Care About." In *The Importance of What We Care About*, edited by Harry Frankfurt, 80-94. Cambridge: Cambridge University Press, 1988b.

Frankish, Keith. "Illusionism as a Theory of Consciousness." *Journal of Consciousness Studies* 23, nos. 11-12 (2016): 11-39.

Frankish, Keith. *Illusionism: As a Theory of Consciousness*. Exeter: Imprint Academic Publishing, 2017.

Gounaris, Alkis. "Human Cognition and Artificial Intelligence: Searching for the Fundamental Differences of Meaning in the Boundaries of Metaphysics." Accessed January 14, 2019. https://alkisgounaris.gr/gr/research/human-cognition-artificial-intelligence/.

Gunderson, Keith. "Descartes, La Mettrie, Language, and Machines." *Philosophy* 39, no. 149 (1964): 193-222.

Gunkel, David. *The Machine Question: Critical Perspectives on AI, Robots and Ethics*. Cambridge, Massachusetts: MIT Press, 2012.

Hajdin, Mane. *The Boundaries of Moral Discourse*. Chicago: Loyola University Press, 1994.

Hardcastle, Valerie. *The Myth of Pain*. Cambridge, MA. Massachusetts: MIT Press, 1999.

Haugeland, John. *Artificial Intelligence: The Very Idea*. Cambridge, Massachusetts: MIT Press. Chicago, 1985.

Hoffmann, Christian, and Benjamin Hahn. "Decentered Ethics in the Machine Era and Guidance for AI Regulation." *AI & Society* 35, no. 3 (2020): 635-644.

Jaworska, Agnieszka. "Caring and Full Moral Standing." *Ethics* 117, no. 3 (2007): 460-497.

Jaworska, Agnieszka. "Caring and Internality." *Philosophy and Phenomenological Research* 74, no. 3 (2007): 529-568.

Jaworska, Agnieszka. "Caring, Minimal Autonomy, and the Limits of Liberalism." In *Naturalized Bioethics: Toward Responsible Knowing and Practice*, edited by Hilde Lindemann, Marian Verkerk, and Margaret Walker, 80-105. Cambridge: Cambridge University Press, 2008.

Kim, Jaegwon. *Philosophy of Mind*. USA: Westview Press, 1998.

Kim, Jaegwon. *Supervenience and the Mind*. Cambridge: Cambridge University Press, 1993.

Lang, Fabienne. "AI Flawlessly Beats US Air Force F-16 Pilot in Simulated Dogfight." *Interesting Engineering*. Accessed August 21, 2020. https://interestingengineering.com/ai-flawlessly-beats-us-air-force-f-16-pilot-in-simulated-dogfight.

Levy, David. *Intimate Relationships with Artificial Partners*. Ph.D. Diss., Maastricht University, 2007.

Levy, David. "The Ethical Treatment of Artificially Conscious Robots." *International Journal of Social Robotics* 1, no. 3 (2009): 209-216.

Lucas, John R. "Minds, Machines and Gödel." *Philosophy* XXX-VI (1961): 112-127.

Matthias, Andreas. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6, no. 3 (2004): 175-183.

Michie, Donald. "Turing's Test and Conscious Thought." In *Machines and Thought. The Legacy of Alan Turing*, vol. 1, edited by Peter Millican, and Andy Clark, 27-51. Oxford, New York: Oxford University Press, 2002.

Ministry of Defense. "Joint Doctrine Note 2/11, The UK Approach To Unmanned Aircraft Systems." Accessed July 20, 2020. https://www.law.upenn.edu/live/files/3890-uk-ministry-of-defense-joint-doctrine-note-211-the.

Monroe, Andrew E., Kyle D. Dillon, and Bertram F. Malle. "Bringing Free Will Down to Earth: People's Psychological Concept of Free Will and its Role in Moral Judgment." *Consciousness and Cognition* 27 (2014): 100-108.

Montaigne, Michel de. "An Apology for Raymond Sebond." In *Michel de Montaigne: The Complete Essays*, translated by Michael A. Screech. Penguin Books, 2003.

Moore, Edward F. "Artificial Living Plants." *Scientific American* 195, no. 4 (1956): 118-126.

Müller, Vincent C. "Autonomous Cognitive Systems in Real-World Environments: Less Control, More Flexibility and Better Interaction." *Cognitive Computation* 4, no. 3 (2012): 212-215.

Müller, Vincent C. "Ethics of Artificial Intelligence and Robotics." *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition). Edited by Edward N. Zalta. https://plato.stanford.edu/archives/fall2020/entries/ethics-ai/.

Olson, Eric T. "Personal Identity." *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition). Edition by Edward N. Zalta. https://plato.stanford.edu/archives/fall2019/entries/identity-personal/.

Owen, Jonathan, and Richard Osley. "Bill of Rights for Abused Robots: Experts Draw up an Ethical Charter to Prevent Humans Exploiting Machines." *The Independent*. Last modified April 1, 2007. https://www.independent.co.uk/news/science/bill-of-rights-for-abused-robots-5332596.html.

Pagallo, Ugo. "Robots of Just War: A Legal Perspective." *Philosophy & Technology* 24, no. 3 (2011): 307-323.

Pinker, Steven. "Can a Computer Ever be Conscious?" *US News & World Report* 123, no. 7 (1997). Accessed July 28, 2020. https://stevenpinker.com/files/pinker/files/computer.pdf.

Putnam, Hilary. *Representation and Reality*. Cambridge: MIT Press, 1992.

Regan, Tom. *The Case for Animal Rights*. Berkeley & Los Angeles: University of California Press, 1983.

Rey, Georges. "A Question About Consciousness." In *Perspectives on Mind*, edited by Herbert R. Otto, and James A. Tuedio, 5-24. Dordrecht: D. Reidel Publishing, 1988.

Rey, Georges. "A Reason for Doubting the Existence of Consciousness." In *Consciousness and Self-Regulation*, vol. 3, edited by Richard J. Davidson, Gary E. Schwartz, and David Shapiro, 1-39. New York: Plenum, 1983.

Rucker, Rudy. *Infinity and the Mind: The Science and Philosophy of the Infinite*. Princeton, N.J.: Princeton University Press, 1982.

Russell, Stuart, Max Tegmark, et al. "Autonomous Weapons: An Open Letter from AI & Robotics Researchers." *Future of Life Institute*. Accessed July 25, 2020. https://futureoflife.org/open-letter-autonomous-weapons/.

Savova, Viginia, and Leonid Peshkin. "Is the Turing Test Good Enough? The Fallacy of Resource-Unbounded Intelligence." *International Joint Conferences on Artificial Intelligence Organization: Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, IJCAI-07: 545-550. Accessed August 29, 2020. https://www.ijcai.org/Proceedings/07/Papers/086.pdf.

Sealre, John. *Minds, Brains, and Science*. Cambridge Massachusetts: Harvard University Press, 1984.

Searle, John. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, no. 3 (1980): 414-457.

Searle, John. *The Mystery of Consciousness*. New York: The New York Review of Books, 1997.

Shoemaker, David. "Caring, Identification, and Agency." *Ethics* 114, no. 1 (2003): 88-118.

Shoemaker, Sydney. *Identity, Cause, and Mind*. Cambridge: Cambridge University Press, 1984.

Singer, Peter. *Animal Liberation: A New Ethics for Our Treatment of Animals*. New York: New York Review of Books, 1975.

Singer, Peter. *Practical Ethics*. Cambridge: Cambridge University Press, 1993.

Singer, Peter. *Wired for War*. New York: Penguin Press, 2009.

Sloman, Aaron. "A Systematic Approach to Consciousness (How to Avoid Talking Nonsense?)." Accessed July 28, 2020. http://www.cs.bham.ac.uk/research/projects/cogaff/misc/consciousness.rsa.text.

Solum, Lawrence. "Legal Personhood for Artificial Intelligences." *North Carolina Law Review* 70, no. 4 (1992): 1231-1287.

Sparrow, Robert. "Killer Robots." *Journal of Applied Philosophy* 24, no. 1 (2007): 62-77.

Tegmark, Max. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Knopf, 2017.

Torrance, Steve. "Could We, Should We, Create Conscious Robots?" *Journal of Health Social and Environmental Issues* 4, no. 2 (2004): 43-46.

Turing, Alan. "Computing, Machinery and Intelligence." *Mind* LIX (1950): 433-660.

Turing, Alan. "On Computable Numbers with an Application to the Entscheidungsproblem." *Proceedings of the London Mathematical Society* 42, Series 2 (1937): 230-265.

Turing, Alan. "On Computable Numbers with an Application to the Entscheidungsproblem." *Proceedings of the London Mathematical Society* 43, Series 2 (1938): 544-546.

Vincent, James. "Former Go Champion Beaten by DeepMind Retires after Declaring AI Invincible." *The Verge*. Accessed August 1, 2020. https://www.theverge.com/2019/11/27/20985260/ai-go-alphago-lee-se-dol-retired-deepmind-defeat.

Von Neumann, John. *Theory of Self-Reproducing Automata*. Urbana: University of Illinois Press, 1966.

Wallach, Wendel, and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press, 2009.

Walsh, Toby. *It's Alive: Artificial Intelligence from the Logic Piano to Killer Robots*. Hamburg: Edition Körber, 2017.

Watson, Gary. "Free Agency." *Journal of Philosophy* 72, no. 8 (1975): 205-220.

Weller, Chris. "Meet the First-ever Robot Citizen – A Humanoid Named Sophia that once Said it Would 'Destroy Humans.'" *Business Insider*. Accessed July 30, 2020. https://www.businessinsider.com/meet-the-first-robot-citizen-sophia-animatronic-humanoid-2017-10?r=UK.

Wilkes, Kathleen. "Yishi, Duh, Um and Consciousness." In *Consciousness in Contemporary Science*, edited by Antony Marcel, and Edoardo Bisiach, 16-41. Oxford: Oxford University Press, 1988.