Journal of Ethics and Emerging Technologies

IEET

Article

The cognitive capacity of free will: a specific space for the human being irreproducible in AI

Cristiano Calì [1,*]

[1] Università degli Studi di Torino – Institute for Ethics and Emerging Technologies (IEET); cristiano.cali@unito.it

* Correspondence: cristiano.cali@unito.it; Tel.: +39 3491489788 (Italy, Rome 00167)

Abstract: For several decades, artificial intelligence - understood as a discipline but also as a series of increasingly advanced products of robotic science - has contributed to rethinking certain concepts typical of anthropology, including that of free will, a cognitive capacity that has been seriously questioned by neurophysiology in the last fifty years, but which today takes center stage when algorithms are constructed to act. This paper aims to show the necessary preconditions for an analogous argumentation between humans and algorithms regarding freedom to demonstrate whether such an approach is at least methodologically suitable for subsequent ethical reflection. An attempt is made to define the points of contact between humans and artificial intelligence with regard to the question of freedom (understood as freedom of will and not just freedom of action) and then to determine whether a certain understanding of freedom is comprehensible. The final section attempts to define the distinction between humans and machines, namely whether it is to be found in consciousness or in freedom.

Keywords: free will; artificial agency; responsibility of AI; mind-body problem.

Citation: Calì, Cristiano, 2023. The capacity of free will: a specific space for the human being irreproducible in AI. *Journal of Ethics and Emerging Technologies* 33: 2. https://doi.org/10.55613/rgpyzx09

Received: 18/06/2023
Accepted: 29/01/2024
Published: 29/01/2024

Publisher's Note: IEET stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

For several decades, artificial intelligence – understood as a discipline but also as a series of increasingly advanced products of robotic science – has contributed to rethinking certain concepts typical of anthropology, including the problem of free will, a cognitive capacity that has been seriously questioned by neurophysiology over the last fifty years, but which today takes centre stage when algorithms are constructed to act.

This article aims to show the necessary preconditions for conducting analogical reasoning between humans and algorithms regarding freedom in order to demonstrate whether such an approach is at least methodologically suitable for subsequent ethical reflection. In the final section, an attempt is then made to define what the discrimination between humans and machines might consist of and whether this is to be found specifically in consciousness or in freedom.

Two brief preliminary remarks are necessary. I will not consider freedom in its social declension. Therefore the relationship between freedom and artificial intelligence will not be understood through the significant changes that AI has brought about in the way freedom is exercised (Calì 2021; 2022), nor by considering the ethical implications of the important restrictions that AI imposes on freedom in specific contexts (Di Nonno 2022;

Grisanti 2022; Cucci 2020; Krienke 2020). I will confine myself to understanding the problem of freedom in its substantive aspect rather than in its exercise[1].

2. Strong AI

Although I am not a fan of apocalyptic scenarios and of that literature on artificial intelligence that relies on catastrophic predictions, in order to develop a reasonable analysis of freedom (in the strong sense I have assumed) concerning AI (an analysis that can then serve as an ethical premise), it would be useful to start by working with the imagination, imagining a context in which robots are fully integrated into our society, operating in our work contexts and acting and dialoguing just like humans[2]. It must, therefore, be assumed that the term AI will, from now on, be understood in the broadest sense. However, even this approach requires further justification.

The distinction between weak and strong AI is well known, but it is easy to see that an argument about the possible freedom of weak AI makes little sense[3]. Indeed, to speak of a narrow AI (or symbolic AI) would be to ignore aspects such as identity, consciousness, autonomy, and freedom. A narrow AI is, in fact, simply a refined version of Archimedes' lever: a tool to which humans delegate their efforts concerning simple tasks. These AIs are also called reproductive because they merely reproduce human activity. Although the best results to date have been achieved in this area, when we refer to weak AI, we should remember Edsger Dijkstra's famous *dictum*: 'The question whether a computer can think is no more interesting than the question whether a submarine can swim"[4]. I will, therefore, always refer to strong AI in the following.

While we can see weak AI immediately when we open our web browser or any application on our smartphone, we can only hint at strong AI (Bostrom 2014) or general AI (AGI). This AI would be an intelligence as flexible as human intelligence, capable of combining different concepts from different domains. This AGI is also referred to as productive artificial intelligence (to distinguish it from reproductive intelligence), as it «seeks to obtain the non-biological equivalent of our intelligence, regardless of the greater or lesser applicative success of the result» (Floridi 2021, p. 140). It would not only reproduce certain functions but also produce the mind. However, this productive dimension does not yet exist today, and as a branch of cognitive science, «it is a complete failure» (op. cit., p. 143); nevertheless, like Bostrom, I believe that strong AI cannot be too easily dismissed as a science fiction dream, not least because the extraordinary computing power of the new quantum computers seems to be an excellent support to continue on the path towards AGI:

> «It may be that the technological innovations are not so rapid-fire that they lead to a full-fledged singularity in which the world changes almost overnight. But this shouldn't distract us from the

[1] More specifically, I will defend that freedom that has been called free will in the specialist debate and, within it, that type of free will that we can define - following Mele's distinction (2014) - as strong free will and not just weak free will.

[2] It should be noted that the so-called narrow artificial intelligence also poses serious problems to the question of freedom but it would be impossible to investigate them here. On this topic cf. Calì (2024, submitted). While it is true that the realization of perfectly functioning cyborgs in the normal world, like Ava in Alex Garland's film *Ex Machina*, is a long way off, I do not consider such a scenario impossible, although not attainable in the short term. If we consider that until a little less than ten years ago, many of the algorithmic applications that we make use of daily did not exist but had already been preconceived by TV series or films, there is no reason to rule out such a scenario. Today, moreover, several robots have attained abilities that were once the exclusive preserve of humans, but above all they have developed techniques to enable humans to achieve results that were previously unthinkable.

[3] For this debate, cf. Bostrom (2014).

[4] I am not trying to argue that it would be possible to build an AI like human beings. In this article, I try to explain how, theoretically, the reductionist paradigm is erroneous because there is at least one irreducible and, therefore, irreproducible element.

> larger point: we must come to grips with the likelihood that as we move further into the twenty-first century, humans may not be the most intelligent beings on the planet for that much longer. The greatest intelligences on the planet will be synthetic» (Schneider 2019, p. 21).

To believe that these problems will always remain at a hypothetical level is, therefore, undoubtedly legitimate but certainly not justified. Suffice it to recall two aspects. Firstly, in 2017, some of the most authoritative representatives of research in this field predicted that it would be possible to achieve this in 2030 (or 2060) (Dilmegani, 2023); secondly, Microsoft, one of the giants of the sector, which has financed an investment of billions, has moved in this direction. It can be said that «researchers are striving to turn science fiction into science fact» (Schneider 2019, p. 25).

As much as strong AI is only hypothetical today, I follow Susan Schneider's assessment in my approach:

> «I suspect it's not that far away, though […]. Billions of dollars are now pouring into constructing smart household assistants, robot supersoldiers, and supercomputers that mimic the workings of the human brain» (op. cit., p. 19).

The fact that she proposes here to take a strong AI perspective is therefore not aimed at formulating utopian ideas but rather corresponds to the approach also proposed by Schneider, according to which it is better from an ethical point of view to assume *a priori* that a highly developed artificial intelligence can have a consciousness, at least until there is a test that can prove the opposite.

Linked to the realization of AGI is the so-called dream of the *singularity*, to use the now classic expression of the writer Vernor Vinge. This intention, contemplated by renowned scientists, philosophers and engineers, aims at an intelligence capable of developing the ability to find solutions to problems still denied to the limited human intellect (Kurzweil 2009). The singularity would be when machines reach a level of intelligence far superior to that of humans, a moment when the latter will not be able to understand the decisions and behavior of machines, which could endanger the survival of humans.

At the moment, I would not subscribe to any of these catastrophic scenarios. Still, at the base of these projects – which aim to reproduce an artificial mind capable of surpassing the human one – there is indeed an explicit consensus that the human mind is reducible to its functions and mechanisms, which could also run on an artificial mind (although it is by no means essential that artificial intelligence be human-like to possess all or most human capabilities). Hawking's words are emblematic of this orientation: «I think the brain is like a program [...] so theoretically it is possible to copy the brain onto the computer and thus realize a form of life after death» (in Schneider 2019, p. 145) .

It would be impossible at this point to even hint at the very long debate that has been and is associated with the *brain-hardware* and *mind-software* metaphors; I will merely point out two reasons proposed by Schneider – that show that some of the premises on which the brain-computer theory is based are completely false. Some proponents of this orientation claim that even if we improve hardware-brain artificiality by leaps and bounds, it will still be possible to run the same mind-software[5]. Another element by which I consider this analogy deeply flawed is the fact that one of the basic assumptions of this metaphor is that the software mind has nothing to do with the biological constitution (let alone the knowledge of this constitution) of the hardware, i.e. the brain. While this approach is profoundly reductionist (since it reduces the complex mind to mere functions), it also seems to strongly contradict another reductionist assumption: We are no more than machines, and biological machines at that:

[5] A researcher such as Schneider invited us to note the impossibility for our mind to be an abstract entity on a par with software (Schneider, 2019, chap. 8).

> «The worlds of artificial intelligence, biology, and even neuroscience are inebriated with this notion. It is acceptable to say, without qualifications, that organisms are algorithms and that bodies and brains are algorithms. This is part of an alleged singularity enabled by the fact that we can write algorithms artificially and connect them with the natural variety, and mix them, so to speak. In this telling, the singularity is not just near: it is here» (Damasio 2018, p. 230).

The dream of productive AI therefore seems to rest on a weak philosophical foundation. Thus Damasio states:

> «Once we would remove the current chemical substrate for suffering and for its opposite, pleasure and flourishing, we would remove the natural grounding for the moral systems we currently have. Of course, artificial systems could be built to operate according to "moral values." That would not mean, however, that such devices would contain a grounding for those values and could construct them independently» (op. cit., p. 233).

However, this critical reference to the brain-computer metaphor is not intended to express a judgment on the theory that our brain apparatus functions largely (but not exclusively) according to computational schemes. Assuming that the analogy between brain and hardware works – assuming we view the brain solely as an information processor – the other part of the analogy, that between mind and software, is much more complex. On the one hand, it must be said that hardware has never been seen to produce software by itself, not least because «in a computer, software and hardware are separate, while the mind, on the other hand, is already contained in the physical body and inseparable from the brain» (Chiariatti 2021, p. 45); on the other hand – and here lies the second paradox – such a theory, although reductionist in its premises, becomes dualistic in its conclusions. Rita Levi Montalcini's brief comment on this question is noteworthy: «Can circuits of silicon neuroids process mental states? If a complex of neu-roids were capable of possessing mental faculties, such as those of consciousness, we would be faced with a dualist theory re-proposed in an entirely new guise» (2004, p. 34).

However, I believe that my attempt to identify a non-reproducible human characteristic is valid because, according to authoritative scholars, artificial intelligences are 'configured «in discontinuity with all the other types of machines that preceded them, both in the ancient world and in the world that emerged from the industrial revolutions, with AI technologies being proposed, in effect, as machines for "augmenting" our intellectual capacities» (Cabitza 2021, p. 39). Thus, if there is such a close connection between artificial intelligence and cognitive abilities - also and especially with regard to the already existing applications of weak AI - one cannot escape the artificial intelligence argument when trying to defend freedom as a cognitive ability.

3. The myth of the *man-machine*

If you take AI in its strongest sense, you are catapulted into the dream or nightmare (depending on your perspective) of the man-machine, a mirage that has been present since the earliest civilizations: think of the rabbinical tale of the Golem or Homer's myth of the three golden virgins built by the Greek god Hephaestus. In the history of philosophy, too, there have been repeated attempts to equate man with an automaton, whether in the matrialistic or the spiritual variant (Leibniz comes to mind). However, one difference must be noted. If in those mythological tales human reproduction was the primary goal, today, in the wake of materialism, this element forms the fundamental premise: «Radical materialism insists that humans and other mammals are no different. Although they are more complex than those simpler systems, and folk psychology treats them as agents, they are ultimately nothing more than elaborate biophysical machines» (List 2019, p. 58) .

The relationship between humans and machines, in particular, becomes crucial for considerations of freedom, as they can essentially be reformulated as follows: Are we willing humans or pre-programmed automatons? This question has indeed been the subject of the *querelle* between determinists and libertarians since the days of the atomistic school. Therefore, It is necessary to understand whether the basic assumption that man is

just a machine deserves attention by noting similarities and differences between these two realities.

There is undeniable correspondence between man and machine, and it is certainly no coincidence that machines have been used metaphorically to speak of man. On the contrary, it must be recognized that the greatest developments in 21st-century science are linked to the discovery that the formation of physical structures and the transmission of information depends on algorithms that rely on codes. «Using an alphabet of nucleic acids – says Damasio – the genetic code helps living organisms assemble the basics of other living organisms and guide their development» (2018, pp. 229-230).

This first aspect is fundamental: humans have a code within them that roughly determines certain organism characteristics. Humans would thus be sophisticated machines equipped with a *basic code*, and «we are each of us composed of trillions of robotic cells, each with its own complete set of genes and an impressive array of internal life-support machinery» (Dennett 2003, p. 150). But it is still a machine. This code – like an algorithm – is subject to certain laws, and this element would suffice – since Hobbes' time – to substantiate the claim that man is an automaton without freedom: «Our thoughts and actions are the outputs of a computer made of meat our brain a computer that must obey the laws of physics. Our choices, therefore, must also obey those laws. This puts paid to the traditional idea of . . . free will (List 2019 p. 3).

Although this understanding of man as a machine is attractive, it remains a mirage because, according to current knowledge, there are more differences than similarities between man and the most modern machine that exists today. It is true that we have reached a stage in evolution where «many aspects of the assembly of natural organisms and of communication depend on algorithms and on coding, as do many aspects of computation as well as the entire enterprises of artificial intelligence and robotics» (Damasio 2018, p. 230). This fact must not be absolutized and must not lead to the already widespread radical notion that natural organisms are nothing more than algorithms.

It is astonishing, then, that engineers try to reproduce the mind on the basis of the computational assumption. At the same time, our experience suggests that the mind fundamentally differs from an algorithm. Consider the symbolic approach to AI, which was not only the first to be developed but also the one that has delivered excellent results: While a machine built according to the paradigm slavishly follows the laws of logic (which, incidentally, were identified by humans), those same laws do not govern the way humans live and think (or at least do not play a predominant role in the most common actions of our lives). These laws have served in computer programs to mimic the world of humans. Still, our minds do not follow the principles that apply to computers and do not do so because they are simultaneously superior and inferior: superior because they are much faster in performing certain processes and can consider much more marginal data simultaneously; inferior because they are much slower in the linear process and tire after a few steps (Boncinelli 2013, p. 63). However, such a solution encompasses both the human being and the machine in functional terms, and consequently, the only question that remains is: what is the discrimination (if any) between a human being and a machine?

4. The *discrimen* between human being and machine: consciousness or freedom?

The majority of scholars have recognized the substantial difference between artificial and The majority of scientists have recognized the essential difference between artificial and natural subjects in consciousness. However, this concept, which has already been described with considerable difficulty by philosophers of mind, does not have an easier fate in the field of AI. Consider, for example, the work of Benjiamin Libet, who summarizes the fundamental difference between two complementary factors (such as experience and consciousness) in the single concept of consciousness (Libet et al. 1983; Libet 2004). To this day, defining consciousness is an immeasurable problem, and consequently it is currently impossible to reproduce it in a machine. However, such a

formulation does not exclude the possibility that a machine cannot achieve this degree of complexity, but only that the machine is incapable of doing so at the present time since it does not know all the processes underlying consciousness.

It should also be noted that consciousness – understood as the possession of conscious mental states – would not in itself be sufficient to draw a line between persons and non-persons. Many mammals, according to Linne R. Baker,

> «have mental states of belief and desire; many mammals have conscious states. What marks persons off from everything else in the world, I shall argue, is that a person has a complex mental property: a first-person perspective […]. We human persons are animals in that we are constituted by animals, but, having first-person perspectives, we are not "just animals." (Baker 2000, p. 4)

The peculiar element in this reading would not be the consciousness that guarantees the existence of conscious mental states but, above all, that which also presupposes the ability to understand oneself in the first person. Baker's theory is certainly intriguing, although it has also been criticized. Still, it allows me here to limit the role that the attribution of consciousness plays in AI discourse in order to focus my attention on a complementary aspect.

Baker argues that the first-person perspective, which we might also call self-consciousness (to distinguish it from other kinds of consciousness), is fundamental since «only beings with a first-person perspective can conceive of their own future» (op. cit., p. 181). In other words, the peculiarity of human beings consists in imagining the future and orienting themselves accordingly. Mind you; this is not about foreseeing goals and orienting oneself towards them – an ability that can also be attributed to some algorithms – but about the ability to anticipate, in the sense of understanding, and thus also to be aware of the future and to act projected into the future. This element is completely absent from AI, and yet those who devote themselves to this field have no qualms about describing some algorithms as acting machines, even though there are three key differences to human action: The machine is not aware of what it is doing, the machine does not decide what activities it performs, and the machine cannot explain why it has chosen a particular path[6]. These three elements are not only prerequisites for free action but also prerequisites for acting out court. A text from one of Mark Twain's masterpieces is useful for understanding the difficulties in pursuing this assimilation:

> «So does a rat. […] He observes a smell, he infers a cheese, he seeks and finds. The astronomer observes this and that; adds his this and that to the this-and-thats of a hundred predecessors, infers an invisible planet, seeks it and finds it. The rat gets into a trap; gets out with trouble; infers that cheese in traps lacks value, and meddles with that trap no more. The astronomer is very proud of his achievement, the rat is proud of his. Yet both are machines; they have done machine work, they have originated nothing, they have no right to be vain; the whole credit belongs to their Maker. […] One is a complex and elaborate machine, the other a simple and limited machine, but they are alike in principle, function, and process, and neither of them works otherwise than automatically, and neither of them may righteously claim a personal superiority or a personal dignity above the other (1906, pp. 95-96).

However, a clarification is imperative if my words are not to be understood as devaluing consciousness in favor of freedom. The question of whether machines can ever be regarded as free is closely linked to the question of whether these machines can ever have consciousness, not only on a theoretical but above all on a practical level: we will be more inclined to regard a machine that kills consciously as free and responsible than a machine that does not kill consciously and would thus indirectly subordinate itself to its programmer. Nevertheless, I address the question of the *agere* of machines and the

[6] One of the most promising branches in this field is eXplainable Artificial Intelligence (XAI) and Algorithm fairness. Work is being done so that some algorithms are able to explain their "action" by at least giving an idea as to why a neural network has acted in one way or another (Adadi and Berrada 2018; Hind at al. 2019).

question of whether this *agere* can be free, in view of the fact that scholars already preach the *agere* of such machines, even though there is no conscious machine today. I would therefore like to suggest that the problem of the free agency of machines can be addressed independently of consciousness. However, discoveries in the latter area would be relevant to the former.

This approach, which I believe is also underpinned by the relationship between consciousness and free will, could be analyzed using Libet's experiments. The relationship between actions and consciousness is indeed not only very complex but also contingent[7]. In other words, machines do not necessarily have to have consciousness to perform complex actions, just as humans do not necessarily have to have perfect and complete consciousness to perform free actions[8]. Further research in this direction could show that consciousness (as Libet understood it) is not an essential component of human action and certainly not of AI action (Nahmias et al. 2019). In this respect, Martin Heisenberg, former lecturer at the Department of Biology at the University of Würzburg:

> «I maintain that we need not be conscious of our decision-making to be free. What matters is that our actions are self-generated. Conscious awareness may help improve our behaviour, but it does not necessarily do so and is not essential. Why should an action become free from one moment to the next simply because we reflect upon it»? (Heisenberg 2009, p. 165).

Suppose certain human actions are classified as free despite the absence of complete and perfect consciousness. In that case, there is no reason not to pose the problem of freedom independently of the *question of consciousness*, or even more so if today's actions that we would call complex (such as evaluating *curricula* or driving cars in privileged contexts) are performed by unconscious machines. So the moment a machine can "act", the problem of freedom in relation to AI does indeed arise and can – at least hypothetically – be addressed independently of the future attainment of consciousness.

5. The way of understanding the *agere* of AIs

Before I go into *medias res*, I would like to point out that – although studies on AI and machine ethics have increased greatly in recent times – an appropriate vocabulary is still completely lacking. Today, we even use the common vocabulary of agents for AI. Let me give you just one example: We need not be afraid of intelligent agents, which are undeniably becoming increasingly intelligent and autonomous (Turi et al. 2019, p. 105). To refer to AI, we use a term like agent or adjectives like autonomous, assuming that they have the same meaning for humans as they do for machines. Certainly, this approach is a symptom of the everyday dualism and "animistic" attitude we have already spoken of, but I agree with Massimo Chiriatti when he sees in this manipulation of terms the eternal dream of man to be a producer or creator. This dream «is made explicit in language when we assign names to objects, as if, for example, the term "learning" had the same meaning for machine learning as it does for us. Unfortunately, in our relationship with machines, we lack a neutral vocabulary with which to describe artificial phenomena» (2021, p. 21). However, one should not believe that this is a purely linguistic convention.

The standard work on AI – edited by Stuart Russell, professor at Berkeley, and Peter Norvig, research director at Google, relies on the (no less complicated) notion of rationality to get around the problem of defining intelligence concerning machines, and

[7] Even an author like Galen Strawson does not support the indispensability of consciousness for free will. Although commonly accepted, in fact, the connection between consciousness and free will still remains obscure, and the imperative to clarify it is doubly crucial precisely in light of artificial intelligence (Caruso, 2016).

[8] I note in passing how curious it is that the very people who advocate a close relationship between consciousness and freedom in strong AIs are also those who regard human freedom as an illusion guaranteed solely by our sensation of conscious will.

the entire book assumes that a rational agent is a system that is capable of understanding what the best decisions are to solve a problem or achieve a self-set goal.

I think that a reformulation or reformulation of the vocabulary we use concerning AI is urgently needed, at least until we have achieved the longed-for strong AI, or at least to «distinguish the human quality of action from machines that are [only] endowed with operational capabilities» (Casalone 2020, p. 38). This vocabulary becomes even more necessary if one adopts the paradigm recently proposed by Luciano Floridi, who has indicated a new way in which the acronym for artificial intelligence, commonly used today, should be reformulated.

According to Floridi, we should move from AI to AA, an acronym that no longer refers to *artificial intelligence* but to *artificial agency*. This new form of agency would be unique in that it would be identified with an *agere sine intelligere*. Such a configuration of the question presents itself as *hapax* in both a technical and philosophical context, as action has always been linked to *intelligere* (understood in its etymological sense). This inseparable relationship has also been maintained by those working materially on the development of machines that do things (I use this term provisionally to avoid the acronyms AI, AA, agent machines, or other similar ones), who recognized that intelligence is the prerequisite for action; hence the main task of reproducing intelligence (or at least some of its functions) so that an adequate behavioral output can be obtained from the machines. However, this primitive attempt was abandoned: Intelligence has not been achieved (at least not in its most complex definition), and through the mimetic process of human intelligence, the world has radically changed so that AI is no longer the link between intelligence and action, but the proof that one does not have to be intelligent to be endowed with agency. Today, therefore, we must view AI as a growing resource of agency that is interactive, autonomous, and often self-learning, capable of tackling an ever increasing number of problems and tasks that would otherwise require human intelligence and intervention (and possibly an unlimited amount of time) to be performed successfully (2021, p. 150).

All of these tasks, even particularly complex ones, are excellently performed by AI without having the intelligence that we can still only preach about sentient beings today, and this is really a subversion of our way of understanding AI:

> «This divorce between artificial agency and natural intelligence, between agere and intelligere, is revolutionary. [...] We have modified one of the fundamental equations on which human history and moral evaluation has always been based, the one that identifies acting with natural, at least biological if not human, acting» (op. cit., pp. 150-151).

Even if this radically new form of *agere* takes shape, the difference between *artificial agent* and *real agent* remains. In the infosphere, in a world where the boundaries between real and virtual, human and digital, are no longer tangible, there is still a *default* assumption, a basic idea: the natural agent can imitate an *artificial* agent in the moment. The example given by Floridi is particularly illustrative to get to the heart of the matter:

> «This is why we are regularly asked to prove that we are not robots by clicking on so-called CAPTCHAs, the Completely Automated Public Turing test to tell Computers and Humans Apart. The test consists of slightly altered strings of letters, possibly mixed with other graphic elements, which we have to decipher to prove that we are not an artificial agent but human. It is a trivial test for a human being capable of even the slightest intelligence, but apparently an insurmountable task for AI that only knows how to act: that is how little progress there has been in the cognitive area of the production of non-biological intelligence» (op. cit., p. 160).

Although the distinction between natural and artificial remains, the noun participle is always the same agent, which must call freedom – at least misunderstood as autonomy – and responsibility into question. Since the Middle Ages and even in modern times, freedom has always been understood in connection with the mind: The will could sometimes even deviate from it, since it was broader, but it was always inclined (when

not determined) by it. Today, even this binomiality has been drastically overturned. Floridi's words explain the radical nature of this process:

> «Agere has always been associated with intentionality, either in the sense of acting for an end (intentionality in agere) or in the sense of self-conscious acting (intentionality in agere). In the past, we have treated acting, intelligere and intentionality as three inseparable aspects of the same phenomenon. But in the face of new forms of agere devoid of intelligere and thus a fortiori devoid of intentionality, the question of accountability (of giving an account of agere itself as the cause of something) is separated from that of responsibility (understood as the duty to do or control something, even when there is no direct causal relationship with that something): an artificial agent may be 'causally' accountable for an evil it has caused, but not 'morally' responsible for it, rather like an earthquake. The difference with the earthquake is that the artificial agent is autonomous, capable of learning and modifying its behaviour, and can therefore be causally accountable for an evil, but it is designed, produced, used and controlled by human beings, on whom indirect responsibility is shifted [...], a bit like what happens with a dog, which although it is an agent [...] is never morally responsible for the effects of its actions, as its master can be. This is because a dog, like an artificial agent, lacks the intelligere and intentionality required by moral agere, which remains only human in terms of duties and rights. In other words, the agere of the AI is not reducible to a simple facere of an earthquake and is much more like the "behaving" of a biological agent, with the fundamental difference that it is an agent whose basic characteristics are designed and approved by other human agents, upon whom the responsibility for the artificial agere is thus transferred. The consequence is that AI does not shift or diminish human responsibility, but magnifies it enormously. AI is the child of a lesser god, humanity, which often knows how to create more than it knows how to manage» (op. cit., pp. 151-153).

It is not possible to discuss here all the issues that such a new understanding of AI entails. Still, I believe that there are solid elements for a recalibration of the current approach to AI ethics, since it seems that the capacity to be free is a *constitutivum* of the human being that allows *agere* to be derived exclusively from esse. Thus, since, as Kant said, freedom is the *ratio essendi* of the moral law, it will be necessary to reconceive the foundations of AI ethics before sclerotizing reflection exclusively in the field of applied ethics.

References

(Adadi & Berrada 2018) Adadi A. - Berrada M., *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*, IEEE Access, 6 (2018), pp. 52138-52160;

(Baker 2000) Baker L.R., *Persons and Bodies: A Constitution View*, Cambridge University Press, Cambridge-New York 2000;

(Boncinelli 2013) Boncinelli E., *Quel che resta dell'anima*, BUR Rizzoli, Milano 2013;

(Bostrom 2014) Bostrom N., *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Oxford-New York 2014;

(Cabitza 2021) Cabitza F., *Deus in machina? L'uso umano delle nuove macchine, tra dipendenza e responsabilità*, in Id. - Floridi L., *Intelligenza artificiale. L'uso delle nuove macchine*, Bompiani, Milano 2021, pp. 9-111;

(Calì 2022) Calì C., Id., *Algoritmi e processo decisionale. Alle origini della riflessione etico-pratica per le IA*, Scienza & Filosofia, 27 (2022), pp. 69-87;

(Calì 2021) Calì C., *L'imparzialità del giudicante. Alcune implicazioni etiche dell'utilizzo dell'intelligenza artificiale in giurisprudenza*, in Alù A. - Ciccarello A. (eds.), *La pubblica amministrazione del futuro. Tra sfide e opportunità per l'innovazione del settore pubblico*, Editoriale Scientifica, Napoli 2021, pp. 121-134;

(Calì 2024) Calì C., *How Technology Changes Us. The Agency of AI and the Cognitive Ability to Make Rational Decisions*, Journal of Responsible Technology, International Journal of Technoethics, 15 (2024), submitted.

(Caruso 2016) Caruso G.D., *Consciousness, Free Will, and Moral Responsibility*, in Gennaro R.J. (ed.), *The Routledge Handbook of Consciousness*, Routledge, London 2016, pp. 78-90;

(Chiariatti 2021) Chiriatti M., *Incoscienza artificiale. Come fanno le macchine a prevedere per noi*, Luiss University Press, Roma 2021;

(Cucci 2020) Cucci G., *Per un umanesimo digitale*, La Civiltà Cattolica, I, 2020, pp. 27-40;

(Damasio 2018) Damasio A.R., *The Strange Order of Things: Life, Feeling, and the Making of Cultures*, Pantheon Books, New York 2018;

(Dennett 2023) Dennett D.C., *Freedom Evolves*, Viking Press, New York 2003;

(Di Nonno 2022) Di Nonno E., *Cina: il contenimento del Covid che mette in pericolo la privacy*, in https://masterx.iulm.it/news/esteri/cina-privacy-abolita-con-il-pretesto-del-covid/ (cons. 15/08/2022);

(Dilmegani, 2023) Dilmegani C., *When will singularity happen? 1700 expert opinions of AGI*, 2023, https://research.aimultiple.com/artificial-general-intelligence-singularity-timing/;

(Floridi 2021) Floridi L., *Agere sine intelligere. L'intelligenza artificiale come nuova forma di agire e i suoi problemi etici*, in Id. - Cabitza F., *Intelligenza artificiale. L'uso delle nuove macchine*, Bompiani, Milano 2021, pp. 115-183;

(Grisanti 2022) Grisanti C., *La lezione della Corea del Sud nella lotta al Covid-19*, in https://www.internazionale.it/notizie/claudia-grisanti/2020/03/18/lezione-corea-sud-covid-19/ (cons. 15/08/2022);

(Heisenberg 2009) Heisenberg M., *Is Free Will an Illusion?*, Nature, 459 (2009), pp. 164-165;

(Hind et al. 2019) Hind M. - Wei D. - Campbell M. - Codella N.C.F. - Dhurandhar A. - Mojsilović A. - Ramamurthy K.N. - Varshney K.R., *TED: Teaching AI to Explain its Decisions*, AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (2019), pp. 123-129;

(Krienke 2020) Krienke M., *I robot distinguono tra bene e male? Aspetti etici dell'intelligenza artificiale*, Aggiornamenti Sociali, (2020) 4, pp. 315-321;

(Kurzweil 2009) Kurzweil R., *The Singularity Is Near: When Humans Transcend Biology*, Duckworth, London 2009;

(Levi Montalcini 2004) Levi Montalcini R., *Abbi il coraggio di conoscere*, Mondadori, Milano 2004;

(Libet et al. 1983) Libet B., - Gleason C.A. - Wright E.W. - Pearl D.K., *Time of Conscious Intention to Act in Relation of Cerebral Activity to Onset of Cerebral Activity (Readiness-potential): The Unconscious Initiation of Freely Voluntary Act*, Brain, 106 (1983) 3, pp. 623-642;

(Libet 2004) Libet B., *Mind Time: The Temporal Factor in Consciousness*, Harvard University Press, Cambridge/MA-London 2004;

(List 2019) List C., *Why Free Will Is Real*, Harvard University Press, Cambridge/MA-London 2019;

(Mele 2014) Mele A. *Free: Why Science Hasn't Disproved Free Will*, Oxford University Press, Oxford-New York 2014;

(Nahmias et al. 2019) E. Nahmias - C. Hill Allen - B. Loveall, *When Do Robots Have Free Will? Exploring the Relationships between (Attributions of) Consciousness and Free Will*, in B. Feltz - M. Missal - A.C. Sims (eds.), *Free Will, Causality, and Neuroscience*, Brill - Rodopi, Leiden 2019, pp. 57-80;

(Schneider 2019) Schneider S., *Artificial You: AI and the Future of Your Mind*, Princeton University Press, Princeton/NJ 2019;

(Turi et al. 2019) Turi N. - Gori M. - Landi M., *Guida per umani all'intelligenza artificiale. Noi al centro di un mondo nuovo*, Giunti, Firenze-Milano 2019;

(Twain 1917) Twain M., *What Is Man? and Other Essays*, Floating Press, Boston, 1917.