RICERCHE

Artificial intelligences as extended minds. Why not?

Gianfranco Pellegrino[α] & Mirko Daniel Garasic[β]

Ricevuto: 30 novembre 2019; accettato: 25 giugno 2020

Abstract Artificial intelligences and robots increasingly mimic human mental powers and intelligent behaviour. However, many authors claim that ascribing human mental powers to them is both conceptually mistaken and morally dangerous. This article defends the view that artificial intelligences can have human-like mental powers, by claiming that both human and artificial minds can be seen as extended minds – along the lines of Chalmers and Clark's view of mind and cognition. The main idea of this article is that the *Extended Mind Model* is independently plausible and can easily be extended to artificial intelligences, providing a solid base for concluding that artificial intelligences possess minds. This may warrant viewing them as morally responsible agents.

KEYWORDS: Artificial Intelligence; Mind; Moral Responsibility; Extended Cognition

Riassunto *Intelligenze artificiali come menti estese. Perché no?* – Intelligenze artificiali e robot simulano in misura sempre crescente le capacità mentali e i comportamenti intelligenti umani. Molti autori, tuttavia, sostengono che attribuire loro capacità mentali umane sia concettualmente errato e moralmente pericoloso. In questo lavoro si difende l'idea per cui le intelligenze artificiali possano avere capacità mentali simili a quelle umane, sostenendo che menti umane e artificiali possano essere considerate come menti estese – sulla scorta della prospettiva di Chalmers e Clark circa la mente e la cognizione. L'idea principale alla base di questo lavoro è che il *Modello della Mente Estesa* abbia plausibilità a prescindere e che possa essere facilmente esteso alle intelligenze artificiali, fornendo una base solida per concludere che le intelligenze artificiali possiedano delle menti e si possano considerare come agenti moralmente responsabili.

PAROLE CHIAVE: Intelligenza artificiale; Mente; Responsabilità morale; Conoscenza estesa

1 Introduction

ROBOTS AND OTHER FORMS OF artificial intelligence (AIs) seem unable to replicate human intelligence in general. AIs seem to lack mental powers. If so, they cannot have moral responsibility, hence humanhood. On the basis of this argument, some authors claim that considering AIs as moral agents or patients is conceptually inappropriate and morally dubious, because it would amount to unduly extending moral status to entities lacking the mental powers necessary to qualify as moral agents or patients. This could have the effect of weakening our awareness of what makes human beings morally worthy of respect,

[α]Dipartimento di Scienze Politiche, Università LUISS "Guido Carli", viale Pola, 12 - 00198 Roma (I)

[β]Dipartimento di Giurisprudenza, Economia, Politica e Lingue Moderne, LUMSA, via Marcantonio Colonna, 19 - 00192 Roma (I)

E-mail: gpellegrino@luiss.it (✉); m.garasic@lumsa.it

thereby eroding respect for human beings.[1] Let's call this set of claims the *Artificial Moral Responsibility* puzzle (for short, AMR).

Notice that AMR rests on an implicit argument, that can be put as follows. Only human beings can be genuine *responsible* moral agents, because responsibility entails moral agency, and the latter requires some specifically human characteristics, in particular the capacity of giving shape to intentions and free will, understood as mental powers. In a nutshell, moral responsibility requires having a mind.[2] Artificial agents cannot have a mind. Therefore, they cannot be regarded as responsible moral agents.

The argument above could give rise to the following objections. First, a distinction can be made between responsible moral agency and humanhood. According to some authors, children are not fully responsible moral agents, however they are obviously regarded as human. Second, some features required by responsible moral agency can be shared by non-human creatures and be found lacking in human creatures. Superior animals can have human-like cognitive powers, and they can form intentions. Some humans fail to have the cognitive and the rational mental powers needed to form intentions and act autonomously. Moreover, mental capacities very often come in degrees. As a consequence, moral responsibility can be a gradual feature, as well. These two objections can be stated in a general form, as follows: The fact that entities lacking mentality cannot be regarded as humans does not entail that entities with minds are *ipso facto* human. The claim that "mindless" creatures cannot be human does not entail that "minded" creatures are necessarily human. Humanhood is more than having a mind.

An answer to AMR has been provided by Luciano Floridi. He claims that there is a level of description and a general ethical view (that Floridi calls *information ethics*) according to which

> *artificial* informational entities [such as software or robots, but also corporations, organizations and other kinds of legal persons], insofar as they can be agents, can also be accountable *moral* agents.[3]

Floridi describes an artificial agent as displaying three characteristics: interactivity (the ability to interact with its environment), autonomy (the capacity to «change its state without direct response to interventions», by performing internal transitions), and adaptability (the ability to change its internal states depending on external inputs, so as to be viewed as «learning its own mode of operation in a way that depends critically on its experience»[4]). Floridi claims that artificial agents can be morally accountable, namely they can be considered sources of moral goods and evils, or of morally right or wrong actions, even though they are not morally responsible. Their status, Floridi suggests, is like the condition of children or superior animals, whose actions can be regarded as morally assessable, even though full moral responsibility should be ascribed to their parents and owners. Artificial agents cannot be ascribed full-fledged moral responsibility because they lack mental powers, such as the capacity to form intentions. For this reason, Floridi concludes, the morality of artificial agents is a "mindless morality."[5] However, Floridi remarks, artificial entities still have a moral status, as agents and patients, on account of the consequences of their actions and their capacity to interact autonomously with and adapt to their environment. «There is plenty of room», Floridi suggests, «for prescriptive discourse that is independent of responsibility assignment».[6]

This seems to us like a hasty surrender, also because Floridi does not consider the view of mind assumed by the authors who deny the possibility of ascribing moral responsibility to artificial agents. This denial presupposes a certain view of the nature of mind. For instance, if mentality necessarily requires a biological brain, or phenomenological consciousness, then it follows that non-biological arrangements unable to experience conscious states cannot have mental states. But is this brain-based, consciousness-requiring view the only plausible account of what having a mind means? Floridi seems to take this view for granted. This is clear in the following passage, for instance: information ethics, Floridi says,

> complements the more traditional approach, common at least since the Stoics and revived by Montaigne and Descartes, which considers whether non-human (animal or artificial) agents have mental states, feelings, emotions, and so on. By focusing directly on "mindless morality", one is able to avoid that question, as well as many of the concerns of AI, and tackle some vital issues in contexts where artificial agents are increasingly part of our everyday.[7]

The main aim of this paper is to challenge and supplement Floridi's approach by endorsing a view of the mind that allows ascribing a kind of mental life to artificial agents. The view we will consider concerns the *location* and the *constitution* of the mind. In the so-called *Extended Mind Model* (EMM), put forward by Andy Clark and David Chalmers, our minds spill out into the world, meaning that normal cognitive processes heavily rely on external devices and tools.[8] The mind goes beyond the skull and the brain contained in it and spreads across the world. In this paper, we suggest that endorsing EMM can be a helpful move in the ethics of artificial intelligence. Our reasoning will go as follows. Let's assume that a sufficient condition to ascribe humanhood is possessing a knowing mind. Notice that this assumption takes a lot for granted – indeed, it takes for granted at least what is assumed in most formulations of AMR (assumptions about humanhood and mentality that can elicit the objections raised above). We take for granted (*i*) that some humans can have mental powers; (*ii*) that these mental powers are cognitive powers; and (*iii*) that possession of mental powers is required in order to have moral responsible agency. We are not taking for granted that each and every human being must have cognitive powers, in order to be considered a member of the human species, nor do we assume that animals cannot have cognitive powers. And, of course, we are not claiming that mentality fully overlaps with cognition. Mental activity can fall short of cognition, and it can encompass non-cognitive processes.[9] Nor are we claiming that possession of a mind is a necessary and sufficient condition for being human. Even if mindless creatures are not human – assuming that this is the case – this does not imply that minded creatures are necessarily human. Super-human and non-human agents can have minds. In this respect, our general claim here is to be understood in a broad way. We are not claiming that AIs can have a distinctively human intelligence, but that they can have a form of intelligence that confers genuine moral responsibility, as in morally responsible human agents.

Notice also that our argument here is conditional. We are not claiming that AIs can surely have a mind and be considered genuinely responsible moral agents. Rather, we are claiming that, if EMM is defensible, and if certain other seemingly plausible claims about collective persons and their collective intentions are granted, then the issue about artificial responsible moral agents has not yet been settled, and it is conceptually possible that these agents exist. It is not our aim here to provide an overall defense of EMM and collective intentionality.[10] Likewise, we are not going to provide a fully developed assessment of Floridi's mindless ethics. We do not claim that intentions are crucial or indispensable in ethics. We are simply suggesting that it may be premature to dispense with intentions and minds when considering artificial agents. Also, we are not going to consider empirical arguments, to the effect that genuinely responsible artificial agents will be possible in the future, or concerning the elements needed to implement such agents. We remain at a conceptual level. We claim that responsible artificial agents are not conceptually impossible.

The core idea of EMM is that very often knowledge in humans emerges from a system, constituted of human brains, external features of their environments, and their interactions.[11] If so, the mind itself can be considered to emerge from a system. Let's call this claim the *System view of mind* (from now on, SVM).

It will be our contention that SVM can be generalized, by claiming that knowledge and mind *in general* emerge from systems. However, these systems can be wholly constituted of non-biological parts, for instance they can be made of AIs, the features of their environments, and their interactions (let's call this the *Extended parity* thesis). If so, robots and AIs can have knowledge. Hence, they can be considered, at least *prima facie*, to qualify for moral responsibility.

As said, ours is an internal challenge to Floridi's view of artificial moral agency. We shall build on his claims, going beyond mere accountability and showing that artificial agents can be genuinely morally responsible – or at least that this is a conceptual possibility that cannot be definitely ruled out.

The paper unfolds as follows. In §2, we give details on AMR and on the main kinds of response to AMR in the literature. In §3, we set the stage for arguing that EMM can be applied to AIs – if human minds are extended minds, then AIs can be considered to have human-like minds. If so, AMR can be dispelled. AIs might be responsible moral agents, after all. The main argument in favor of this claim appears in §4, while §5 concludes.

2 Artificial intelligence and the artificial moral status puzzle (AMR)

2.1 Unpacking AMR

AI is sometimes presented as aiming to build artificial persons – or artificial creatures that appear to be persons. This may amount to building creatures able to think like humans, or to have an artificial correlate of a human mind, including capacities for responsible moral agency.[12] This aim can elicit two reactions, both of which question the very possibility and the ethical plausibility of artificial humanhood, and the lack of what John Tasioulas calls "the human factor" in robots and AIs.[13] First, it might be argued that the attempted enterprise is conceptually or empirically impossible. Let us call this the *Impossibility Objection*. Second, it might be contended that, because of this impossibility, regarding AIs as human persons is morally wrong or bad, because it is hubristic, disrespectful of our humanhood, or even dangerous for humans. Let's call this the *Moral Objection*. Sometimes, the *Impossibility* and the *Moral Objection* are not clearly distinguished, especially in non-scholarly discussions addressing the general public.

Even in scholarly discussions, though, it often turns out that the *Impossibility Objection* inherits its urgency from concerns relating to the *Moral Objection*. Sometimes, the Impossibility Objection takes a particular form. Robots or AIs cannot be considered *moral agents,* some suggest, and they cannot be given moral status, because they lack the features that make some human persons capable of moral action, namely *sentience* – i.e. the capacity for phenomenal experience or qualia, and notably the capacity to feel pain and suffer – and *sapience* – i.e. higher intelligence, especially self-awareness and reason-responsivity.[14] Let's call this the *Moral Status Objection*.[15] Often, authors do not distinguish clearly which objection they are raising, and they use arguments in support of one objection as if they were also grounds for other similar objections. What we have called AMR above is the union of these objections.

As noted, AMR embeds complex assumptions. A way to unpack them is by considering the following argument:

The AMR argument

1. *External likeness*: Robots and AIs behave

in ways that are strikingly similar to intentional and mind-driven human conduct, and they are destined to become increasingly similar to us in the distant future, at least in these respects.

2. *Mentality ascription*: As a consequence of 1., we may be tempted to ascribe mental powers to robots and AIs. The reasoning is as follows. If the behaviour of robots and AIs can only be explained by mental causes, then we should posit these causes, thereby ascribing mental powers to robots and AIs.

3. *Internal differences*: However, robots and AIs have different substrata from human beings. At the very least, they have a different brain architecture and a non-biological body. As a consequence, their behaviour can be explained by appealing to non-mental causes.

4. *Insufficiency of external likeness*: (*a*) In virtue of 3. above, we cannot ascribe mental powers to AIs and robots only on the basis of observing external conduct for which the causes are typically mental. (*b*) The possibility that non-mental causes produce the same effects as mental causes is real. Hence, (*c*) the external behaviour of robots and AIs is insufficient evidence for positing internal mental causes of it.

5. *No external likeness without internal likeness (and vice versa)*: In normal cases, external likeness involves internal likeness, and vice versa.[16] If an entity behaves like a human being, this should be because it has a human mind (and a human brain architecture) and a human body. If, however, the entity has a different substratum, its behavior cannot be human, notwithstanding the seeming likeness.

6. *Vs. mentality ascription*: In virtue of 4. and 5. above, 2. is false. We should not ascribe mental powers to robots and AIs.

7. *Morally inappropriate ascription*: 2. is not only false, but also morally inappropriate, because treating robots and AIs as humans, despite their lack of human features, risks inducing blindness to the morally salient and valuable features of human moral agents.

1. to 4. constitute what we called the *Impossibility Objection*, whereas 7. expresses the *Moral Objection*. But, as noted, very often the different parts of AMR, and the two objections, are not disentangled, but rather presented in a conflated form. Consider, for instance, these passages, taken from a recent overview article by John Tasioulas:

> Although robots and artificial intelligence can achieve complex goals – such as recognizing a face in the crowd or translating a document from one natural language to another – they have nothing like the ability to deliberate on the ultimate ends. For some philosophers, this faculty of rational autonomy is the source of the special dignity inherent in human beings, which makes them different from non-human animals. [...] moral decision-making confronts a potential infinity of relevantly different situations that no algorithm or process of machine learning is sensitive enough to engage with adequately. [...] Sound moral reasoning requires the cultivation of emotional responses on the part of the reasoner, such as guilt, indignation, and empathy, that are properly attuned to their objects. It is these responses that enable us to register the moral significance of certain situations [...]. But, arguably, they are inherently beyond the capacities of beings that do not share a human consciousness and way of life.[17]

The claim stated in the passage above is that robots and AIs cannot replicate what makes humans the kind of beings they are and what makes them especially worthy – the features making them human and giving them "special dignity", i.e. their sentience, their deliberative faculties and their capacity as moral agents. Relying on David Wiggins, Tasioulas affirms that for some people, interactions with robots and AIs can give rise to complaints; because they lack human qualities, they cannot engage in mutual understanding, solidarity and responsibility. The unemployed person whose job application is rejected by an automated system may find this dehumanizing and disrespectful. People involved in intimate relationships with robots may experience a sense of revulsion.[18] This view seems to ground the idea that people have a right to meaningful human contact, and that certain interactions between robots or AIs and humans – especially when robots make significant decisions about humans – can jeopardize this right.[19]

This idea may be used to claim that the attempt to build AIs with these distinctively human features is both impossible and debasing. Indeed, it is debasing because it is impossible: calling those imperfect robots and AIs "humans" amounts to disrespecting genuine (i.e. biological) human beings. Moreover, the futuristic scenario of artificial super-intelligences may elicit worries about the prospect of AI's domination of normal human beings. In this perspective, pursuing the project of building increasingly intelligent and human-like robots and AIs could seriously limit the liberties of future generations of biological humans.[20] The connection between the *Impossibility*, the *Moral*, and the *Moral Status Objections* is apparent here. This way of framing the various concerns raised by the prospect of humanoid robots and AIs is a paradigmatic instance of AMR.

The *Impossibility Objection* arises from a specific view of humans and their minds, according to which humans have a specific kind of mind, capable of introspective, phenomenal awareness – i.e. sentience and conscience – and a higher form of intelligence – i.e. sapience.[21] While each of us can entertain doubts about other minds, assuming the principles stated in 4. (*a*) and (*b*) above, we can safely ascribe to our fellow humans the same mental states to which each of us has introspective access. But it is not clear that we can do the same with robots and AIs. It is not clear that they have both phenomenal consciousness and introspective access to it. It is not clear that something without a human mind – or the physical substratum over which a human mind supervenes – can be a genuinely intelligent being.[22]

The *Moral Objection* is backed by two deep concerns. Giving robots and AIs some or all of our humanhood would mean losing the last bastion of the alleged exceptionality that has characterized human self-understanding. For centuries, at least in the West, we have considered ourselves children of God and masters of a world at the center of the universe. Galileo questioned both our divine descent and the centrality of our world in the universe, along with any hope that the world would be governed by an intentional project. Darwin completed the work by showing us how close we are to animals. Freud dissolved the Cartesian picture of a transparent mind. The irreducibility of human consciousness and mind to matter, as well as the idea of a transparent self, were the last remnants of a gap between us humans and the rest of the world. If one admits that consciousness is a property of matter – even of inanimate matter – then no distinction between world and mind, or between human and non-human, remains standing. And this might, of course, be seriously disturbing.[23]

Moreover, our common-sense morality gives essential importance to the links between mind, freedom, and responsibility. Only those who have a mind can be free, because only minds can make decisions, can alter the course of events, and so on. (Of course, free will can be conceptualized in more or less strong ways. However, the claim that the completely deterministic behavior of AIs is a departure from humanhood can be shared by both libertarians and compatibilists). Only those who can be free can be responsible for their actions and can therefore be the object of moral judgments, of praise or blame. Robots and AIs are troubling, because their seeming humanhood may lead us into the temptation to attribute responsibility to them, thereby treating them as moral agents. But, on the one hand this undermines our pride, the pride of being the only moral agents in a world of moral patients, and on the other hand it continues to seem to us an undue extension, a sort of illicit anthropomorphization of inanimate matter: how can a machine be responsible? Would we put the crazy electric saw that cuts the worker's hand in jail, to punish him for the crime?[24] These general worries underlie AMR.

2.2 Responses to AMR

Three responses to AMR appear in the literature. Some authors start from the past. We had artificial persons even before the rise of robots and AIs. We gave legal personality to artificial persons such as states, corporations, and the like. Recently, some scholars have proposed we give legal personality to natural non-human entities, such as rivers or trees.[25] Likewise, we can extend the framework of legal personality to robots and AIs, thereby considering them responsible agents, at least under certain circumstances. Robots and AIs can be legally liable or morally accountable, even though not morally responsible, for the harms they are causally responsible for.[26] Let us call this the *Legal Personality Answer to AMR*.[27]

Other authors start from the future. They predict that we will increasingly see the appearance of human/machine hybrids or cyborgs. As a consequence, the human/machine divide is destined to a gradually blur. Hence, AMR will soon lose its significance. We may have the impression that machines will never replicate the features that make us the humans that we are. But this is only a temporal bias, due to the fact that our conception of humanhood has been shaped by familiarity with purely biological humans. Human beings will increasingly mix their biological parts with non-biological additions, and a parity between biological and non-biological components of the human body is in the offing. As a consequence, the notion of a "human being" where a pure biological constitution is a necessary feature is soon to be abandoned. Then, whatever AIs will be in the future, they will not be relevantly unlike human cyborgs. One day, it will be impossible to tell the difference between compounds of biological parts constituting a specimen of *Homo Sapiens* (wholly biological instances of *Homo Sapiens*) and non-biological assemblages being a specimen of AIs (wholly non-biological instances of robots and AIs).[28] Let's call this the *Cyborg* or *Transhumanist Answer* to AMR.[29]

Finally, as said in §1, Luciano Floridi claims that artificial agents can be considered moral agents because morality does not necessarily require mind – or, better, moral agents do not need a mind. Let's call this the *Mindless Morality Answer* to AMR.

Here, we depart from these answers. We shall argue that AIs can display a minded morality, if we endorse the view of mind embedded in EMM. This move will address AMR, because it will show that AIs have minds; hence, they can be considered, at least in principle, responsible moral agents. Let's call this view the *Extended Mind Answer* to AMR.

3 System views of mind, intention, and morality

The main idea we defend here is that artificial minds can be likened to human minds, provided that the latter are understood according to EMM. This opens the door to ascribing genuine responsibility to artificial agents, at least as a conceptual possibility

In this section, we present two ideas. First, minds emerge from a system, and not necessarily from individual biological brains. Second, intentions and responsibility (and morality) can also emerge from systems. These ideas will turn out to be the first step in our answer to AMR. §4 will be devoted to EMM as a general model of the mind and a specific model of artificial minds.

3.1 The system view of mind (SVM)

Premise 1 and 2 of the AMR argument above may be seen as a statement of the Turing test.[30] Floridi's *Mindless Morality Answer* to AMR is also based on a Turing-like test. Floridi compares two agents – two healthcare assistants in a hospital. One is an artificial agent (say a webbot), the other is a human nurse. He assumes that both agents can

> respond to environmental stimuli – for example the presence of a patient in a hospital bed – by updating their states (interactivity), for instance by recording some chosen variables concerning the patient's health [...], change their states according to their own transition rules and in a self-governed way, independently of environmental stimuli (autonomy), for example by taking flexible decisions based on past and new information, which modify the environment temperature; and [...] change the transition rules by which their states are changed according to the environment (adaptability), for example by modifying past procedures to take into account successful and unsuccessful treatments of patients.[31]

If so, Floridi concludes, both agents can be the source of morally relevant actions. For instance, one of them can kill the patient. But what if the killing is done by the artificial agent? According to Floridi, the killing is still wrong, because both agents

> acted interactively, responding to the new situation with which they were dealing, on the basis of the information at their disposal. They both acted autonomously: they could have taken different courses of actions, and in fact we may assume that they actually changed their behaviour several times in the course of the action on the basis of new available information. They both acted adaptably: they were not simply following orders or predetermined instructions. On the contrary, they both had the possibility of changing the general heuristics that led them to make the decisions they did, and we may assume that they took advantage of the available opportunities to improve their general behavior.[32]

The two agents, according to Floridi, are both accountable, even though, lacking mental states, the webbot is not morally responsible.[33] Their behavioural likeness is sufficient to consider them liable for punishment or preventive action – prison for the human agent, discontinuance for the webbot. Of course, there is no point in blaming the webbot, i.e. in considering it morally responsible. It has no previous intentions, nor can it join the moral game of praise and blame.

As said, this is a hasty conclusion, because it rests on a contentious view. To state it again, the view taken for granted both by Floridi and supporters of AMR is that

i. genuine responsibility requires intentions;

ii. intentions can be formed only by creatures having a mind;

iii. (*a*) mind is necessarily a feature of biological brains; moreover, (*b*) a human mind is necessarily a feature of biological human brains.

John Searle, who famously raised an objection to the Turing test, endorsed *iii.* above. [34] For Searle, mind and knowledge are phenomena internal to the human brain – to a brain made as human brains are made, i.e. with a biological constitution and with a certain conformation. To claim that computers can think, can be intelligent, or can know would mean to admit that the mind can also belong to inanimate matter – or, more precisely, to non-biological agglomerations. But this is not possible: thought must necessarily be embodied in pieces of living biological matter. Searle's position presupposes not only that thought is a property only and necessarily of certain biological entities, but also that the mind and knowledge are placed completely inside the brain. Andy Clark calls this vision of the human mind "BRAINBOUND" and describes it as follows:

> This is the model of mind as essentially inner and, in our case, always and everywhere neurally realized. It is, to put it bluntly, the model of mind as brain (or perhaps brain and central nervous system): if BRAINBOUND is correct, then all human cognition depends directly on neural activity alone.[35]

If this model is correct, then only if robots and artificial AIs are able to *literally reproduce*, and not just simulate, human neural activity, can they be considered minds. The AMR argument incorporates and expands Searle's view of the nature of human mind and his claim that this nature can never be reproduced by robots and AIs. In limiting himself to accountability, Floridi takes for granted Searle's view of mind. However, this view can be challenged.

In §1, we introduced SVM, i.e. the idea that mind supervenes on certain specific kinds of systems, constituted by individual brains, their environment, and brain/environment interactions. A specific version of this view is endorsed in Clark and Chalmers' first presentation of EMM. In many cases, Clark and Chalmers point out,

> the human organism is linked with an external entity in a two-way interaction, creating a *coupled system* that can be seen as a *cognitive system* in its own right. All the components in the system play an active causal role, and they jointly govern behavior in the same sort of way that cognition usually does. If we remove the external component the system's behavioral competence will drop, just as it would if we removed part of its brain. Our thesis is that this sort of coupled process counts equally well as a cognitive process, whether or not it is wholly in the head.[36]

In *Supersizing the Mind* (2008), Clark states SVM with the utmost clarity:

> Possessing a contentful mental state is most plausibly a property of a *whole active system*, perhaps in some historical and/or environmental context.[37]

EMM, then, amounts to claiming that the mind is realized in distributed systems, of which the biological brain is only one of the components – systems that may be composed of human brains and pens, paper, computers, other devices, but also books, memorable places, and so on. According to EMM, the mind is literally constituted by active features of the environment.[38]

Searle locates the mind in a unified, homogenous place. Moreover, he claims that the mind's location is the human biological brain. By contrast, SVM locates minds in systems: minds can supervene on complex systems. But if so, why not say that wholly nonbiological systems can be the realization base of minds? This idea will be developed in §4.

3.2 The system view of intentions

AMR embeds the idea that intentions can be had only by creatures with minds, and responsibility requires intentions. If only creatures with biological brains can have minds, then AIs cannot have intentions. As a consequence, AIs cannot be genuinely responsible agents. In §§ 1 and 3.1. we introduced SVM, i.e. the idea that minds can emerge from systems. Can intentions emerge from systems as well? And if intentions emerge from systems, can we conclude that the latter are morally responsible? And if systems can be morally responsible, what about entirely artificial systems?

In *The Ethics of Information* (2013), Floridi includes organizations in the set of nonhuman moral agents, on account of the fact that these entities are standardly regarded as *legal persons*.[39] Moreover, he considers cases in which the combined action of different agents – constituting «a multi-system agent, which might be human, artificial or hybrid» – yields morally significant outcomes, and specifically cases in which morally neutral actions, if combined, bring about morally bad or wrong results. He calls them cases of *distributed morality*. An obvious example of a distributed morality case is the tragedy of the commons.[40] Notice that in these cases, the morally relevant outcome emerges from a system. For instance, in the tragedy of the commons, a set of actions whose impact in isolation would not yield the spoiling of the resource brings about the spoilage, *qua* set. What Floridi calls distributed morality can be regarded as a *system view of morality*, perfectly parallel to SVM.

Floridi notes that cases of collective responsibility (when «a whole group of people is held responsible for some of its members' actions, even when the rest of the group has had no involvement at all [...] in such actions») are well-known instances of distributed morality.[41] However, he claims that intentions can be irrelevant in the most important cases of distributed morality he focuses on. He intimates that multi-agent systems «might be totally mindless, so that any talk of beliefs, desires, intentions and motivations would be merely metaphoric».[42] In a later work, Floridi clarifies that distributed moral actions cannot be intentional, because it is not the case that, if agent *A* means to cause outcome *a*, and *B* means to cause *b*, and *a* and *b* cause *c*, then it follows that A and B mean to cause *c*. As a consequence, distributed agents cannot be ascribed distributed moral responsibility.[43]

However, in recent discussions, the possibility of collective intentions has been explored. While many authors stick to the idea that collective intentions are cases of shared intentions – i.e. intentions of individuals having overlapping or shared contents – other authors maintain that collective intentions are to be predicated of collective subjects.[44] One of these theories has been provided by Bryce Huebner.[45] He gives an account of distributed cognition, or macrocognition, i.e. of cognition emerging from sets of individual cognizers. Importantly, he connects macrocognition to collective intentionality, in a cautious way, grounding his view on both conceptual arguments and cognitive science data. We have no room here to assess this view. It is enough to say that the conceptual possibility that intentions can emerge from a system cannot be ruled out. Moreover, this possibility is connected in obvious ways with SVM. Collective intentions can be a by-product of distributed cognition. Hence, distributed responsibility – i.e. genuine responsibility grounded in previous intentions – can be a by-product of distributed cognition. To put it otherwise, SVM (and EMM) can be the framework within which one can account for the genuine moral responsibility of multi-agent systems. However, multi-agent systems can be hybrids, i.e. in part artificial, in part human. What about entirely artificial multi-agent systems? Can they have genuine moral responsibility for the same reasons? This would be a large inquiry. For now, let's go back to EMM and to the prospect of extending it to AIs.

4 Extended artificial minds

Is there anything in Clark's conception that makes it necessary for at least one component of the system to have a biological nature? If a combination of human biological brain + pencil + paper produces knowledge, then why should a combination with no biological components not be able to produce knowledge? In a paper, Clark writes:

> It seems possible (for example) to ascribe representational contents, in ways that are not obviously conventional or derivative, to the states and processes of artificially evolved creatures [...]. Or, if simple artificial creatures do not move you, take any inner neural structure deemed [...] to

> be the vehicle of some intrinsic content X. Can we not imagine replacing part or all of that structure with a functionally equivalent silicon part? [...]. Unless we question-beggingly assert that only neural stuff can be the bearer of intrinsic content, then surely we should allow that the siliconized vehicle, or at least the hybrid circuit that now includes it, is as capable of supporting intrinsic content as was its biological predecessor?[46]

In his *Natural-Born Cyborgs* (2003), Clark defends the claim that mixing our biological cognitive parts with non-biological cognitive devices is not a future prospect, but our inherent nature as cognitive beings. We are natural-born cyborgs, Clark claims: hybridization for cognitive purposes is an aspect of our humanhood. This claim is at the same time a consequence of and an evidence for EMM. We are natural-born cyborgs because our minds extend beyond the brain, and our cognition supervenes on several parts of the world that we use as cognitive devices and tools. But, again, if parts of the outside world interact with our brains, and knowledge emerges from this interaction, why can knowledge not emerge from interactions between non-biological parts? Why should SVM not be a general conception of knowledge? In this section, we are going to consider arguments in favor of a positive answer to these questions.

As highlighted, we are claiming that EMM should be extended to supposedly artificial minds. The reasoning we employ to support this claim goes as follows:

(1) *Mentality as a sufficient condition of responsibility*: A sufficient condition for ascribing responsibility is possession of mind.

(2) *EMM*: Mind extends beyond human brain and supervenes on several features of the environment.

(3) *Extended parity*: There are no substantial differences between cognitive systems constituted by human brains and their environment and cognitive systems entirely made up of AIs and their environment.

As a consequence, we get

(4) *Artificial minds*: AIs can have minds.

Hence, we get

(5) *Artificial responsibility*: AIs can be considered responsible moral agents.

As said above, we are not tackling the complicated issues concerning (1) in this paper.[47] Here, we focus on *Extended parity*. It is our contention that this claim can be derived from some of the arguments Clark used to support EMM – in particular, by arguments he employed to defend EMM from criticisms. As a consequence, if EMM is independently plausible – a big if, of course – then Extended Parity is plausible as well, and (4) and (5) above follow, or at least they gain support from the plausibility of both EMM and Extended Parity.

Fred Adams and Ken Aizawa, in a series of papers, levelled the following critiques at EMM.[48] First, EMM incurs a fallacy: a "coupling-constitution fallacy". The mistake lies in moving from

1. *Coupling claim*: an object O or process P is coupled in some fashion (for instance, causally) to the cognitive process CP of some cognitive agent CA,

to

2. *Constitution claim*: O or P are parts of CP or of CA.

According to Adams and Aizawa, coupling is not constitution, and what is coupled to a given system is not necessarily part of it. As a consequence, moving from 1. to 2. is a logical and category mistake:

> coupling relations are distinct from constitutive relations, and the fact that object or process *X* is coupled to object or process *Y* does not entail that *X* is part of *Y*. The neurons leading into a neuromuscular junction are coupled to the muscles they innervate, but the neurons are not a part of the muscles they innervate. The release of neurotransmitters at the neuromuscular junction is coupled to the process of muscular contraction, but the process of releasing neurotransmitters at the neuromuscular junction is not part of the process of muscular contraction.[49]

Adams and Aizawa's second objection relates to their first. They suggest that Clark fails to provide a view of the "mark of the cognitive", i.e. of «what makes a process a cognitive process rather than a noncognitive process».[50] According to Adams and Aizawa, a view of what marks the cognitive is the only ground for ascribing parts to cognitive wholes. Once you have a view of what is cognitive, you can establish whether a certain object or process is a genuine part of a cognitive system. However, according to Adams and Aizawa, Clark has no view, or a defective view, of the mark of the cognitive. He simply affirms that «a cognitive process is one that is coupled to a cognitive agent».[51] But this "only pushes back the question" of what makes something a cognitive agent.

Clark reacts to the first objection by denying that he and Chalmers derive a claim about constitution from a claim about coupling. Rather, he suggests, their claim is simply about the conditions under which such constitution is possible. Clark and Chalmers' view concerns the *integration* or *incorporation* of certain parts within larger cognitive systems. The idea is that, in certain conditions of reliable, portable, and automatic coupling, certain objects and process are to be considered as parts of larger cognitive wholes. A successful case of note taking and recollection with the help of notes is a case in which the notebook, the notes written in it, and the brain of the reader constitute a larger system able to produce cognition. As a consequence, Clark is not claiming that when O or P are coupled with CP or CA they are part of CP or CA. Rather, he is maintaining that successful coupling – i.e. a coupling the obtaining of which produces cognition – makes O, P, CP and CA integrated in a larger system. This is what we called above the *System view of mind*, i.e., the view that mental powers emerge out of complex systems.[52]

Moreover, Clark responds to Adams and Aizawa's second critique by rejecting an implicit assumption in it, namely the idea that

> some objects or processes, *in virtue of their own nature* [...] are [...] *candidate parts* (for inclusion in a cognitive process), whereas other objects or processes, still in virtue of their own nature, are not.[53]

Coupling, Clark claims, is

> intended to make some object, which in and of itself is not usefully (perhaps not even intelligibly) thought of as *either cognitive or noncognitive*, into a *proper part of some cognitive system*, such as a human agent.[54]

This remark rests on a general emergentist principle that Clark explicitly put forward later in his 2010 response to Adams and Aizawa. The thought is that knowledge is an emergent property of systems, whose parts must not necessarily be cognitive in themselves. Here is Clark's statement of this principle:

> In general, for some *X* to be part of the supervenience base of some *Y*, where that *Y* must (to count as *Y* at all, let's assume) exhibit some property *Z*, there is no requirement *that Z be in addition a property of the putative part X*.[55]

Then, for some O or P to be part of the supervenience base of some cognitive system, where that system must (to count as cognitive at all) exhibit some property – for instance, being a case of successful cognition – there is no requirement that this property be in addition a property of O or P. Non-cognitive objects and process can be the supervenience base of cognition.

The principle above can be seen as a kind of precisification of SVM. Knowledge supervenes on, or rather emerges out of, systems, and the parts of these systems need not necessarily be cognitive. Non-cognitive parts can yield cognition. Of course, this view rests on the larger functionalist approach that Clark repeatedly advocates in his writings. If cognition is an achievement, a product of certain activities, then we don't need to dig into the nature of what performs the function of producing knowledge. It is enough that something is able to produce knowledge to declare it a part of a cognitive system, or a base of knowledge. Knowledge is nothing more than an activity, a result, an emergent property of inferior level parts. And the only unifying trait of the parts of a cognitive system is that they produce knowledge together. There is no need for anything more profound than looking at the results of the operation of the system.

The logic underlying the thoughts above leads to the following conclusion. There are no principled exclusionary rules concerning what can and cannot be part of a cognitive system. As a matter of experience, we can have pure cognitive systems, where knowledge supervenes entirely on human brains – let's call them *purely biological cognitive systems*. We can also have *impure cognitive systems*, where human brains and non-biological objects and processes are the realization base of cognition. But perhaps we can also have other kinds of pure cognitive systems, where knowledge supervenes entirely on non-biological matter – let's call them *purely artificial cognitive systems*. As noted, there are no principled reasons to exclude the possibility that these systems can be fully cognitive. If an object plays a cognitive function in an impure cognitive system, why exclude the possibility that the same object can play the same function in an artificial pure cognitive system? The possibility of artificial pure cognitive systems vindicates *Extended parity*.

Adams and Aizawa claim that the fact that cognition lies entirely within the human brain is a matter of experience and science. Our best scientific account of knowledge, they suggest, shows that human brains can know, whereas non-biological matter alone cannot. Matter does not think:

> the empirical evidence we have indicates that the brain processes information according to different principles than do common brain-tool combinations. Think of consumer electronics devices. We find that DVD players, CD players, MP3 players, tape recorders, caller ID systems, personal computers, televisions, AM/FM radios, cell phones, watches, walkie talkies, inkjet printers, digital cameras, and so forth, are all information processors. The preponderance of scientific evidence, however, indicates that they process information differently than does the brain. That is why, for example, the brain is capable of linguistic processing, whereas these other devices are not. That is why, for example, the brain is capable of facial recognition over a range of environmental conditions, whereas these other devices are not. This is why the brain is crucial for humans' ability to drive cars, whereas these other devices are not. The differences in information-processing capacities between the brain and a DVD or CD player is part of the story of why you cannot play a DVD or CD with just a human brain. These differences are part of the reason you need a radio to listen to AM or FM broadcasts. It is these differences that support the defeasible view that there is a kind of intracranial processing, plausibly construed as cognitive, that differs from any extracranial cra-

nial or transcranial processing.[56]

If this is the ground for intracranialism, then EMM and its extension to AIs are obviously plausible. Our daily experience shows that AIs can perform many of the operations that the quote above ascribes exclusively to human brains, such as linguistic processing, facial recognition, and driving cars. As a consequence, today's experience and our best account of AIs' potentialities vindicate, at least *prima facie*, the ascription of mental powers to robots and AIs.

Extended parity establishes the independence of cognition and mind possession from substrata. If two systems perform the same function – producing knowledge and displaying mental powers – and differ only in the substratum of their implementation, then they are both cognitive. If two systems perform the same cognitive function and differ only in how they came into existence, then they are on a par. AIs and human minds are not different, in so far as cognition and mental powers are considered.[57]

5 Conclusion

The case for extending EMM to AIs stated above is conditional. The thought is that if EMM is a plausible model of human mind, then AIs have minds. Of course, many objections can be and have been raised against EMM. [58] Some of them can be levelled also against the further extension we put forward here. But there is at least a specific objection to our argument. We assume that behavioral likeness – human beings and AIs can both achieve knowledge – is a necessary and sufficient condition to ascribe mental powers to AIs. However, it might be objected that, whereas we have access to our own minds, we cannot have access to AIs' minds. It is indeed not clear whether robots have intentional and conscious mental states such as those that each of us has and attributes to others.

However, this view rests on implausible assumptions. To decide whether or not robots have intentional mental states. when they seem to have them, is too ambitious a claim, which would require a kind of privileged access (Cartesian? Or telepathic?) to the minds of others: from a certain point of view, we cannot be sure that even other human beings have intentional mental states, yet, we do not deny that they are responsible moral agents. If robots behave as such, if they consistently play the game of knowledge and of morality, why should they be considered any less than cognitive agents? As Alan Turing remarked,

> The only way in which one could be certain that a machine thinks is to be the machine, and to feel oneself thinking. [...] Likewise according to this view the only way to know that a *man* thinks is to be that particular man.[59]

If skepticism about other human minds is to be rebutted, then skepticism about artificial minds should be avoided too.[60]

Two similar skeptical objections might still be raised. On the one hand, it might be argued that when responsibility is at stake, while we are sure about human moral responsibility, we have many doubts about artificial moral responsibility. On the other hand, we can say that what we regard as responsibility when we face AIs' behavior should be ascribed to their human creators and not to the artificial agents.

The first objection is rather question begging. It amounts to saying that we have clear ideas about human responsibility, while we will never have clear ideas about artificial responsibility. Both claims are false.

The second objection can be rebutted as follows. Software engineers often work in teams. Moreover, much of the machine's behaviour depends on other software, software which often works probabilistically, and on interfaces and external factors, multiple interactions, successive maintenance and management regimes, and machine learning. These are factors completely out of the control of the original creators. If so, it is far from clear that each and every component in the conduct of an AI can be ascribed to its creator.[61]

Perhaps what leads to denying that robots can have a mental life, and therefore a moral life, is a mistake in our view of the mind. Mental abilities are not the exclusive property of human biological brains, but they can also be the property of certain clusters of matter. After all, what are our brains, if not clusters of matter? But we unhesitatingly ascribe mentality to them.[62] In his notebooks from 1836-1844, Darwin writes: «If all men were dead, then monkeys would make men – men would make angels». Darwin seems to mean that, in the event of the extinction of our species, other species would occupy our ecological niche. But he also seems to suggest that human beings may themselves evolve into something different. If there is no *Scala Naturae*, as Darwin taught us, we cannot see why certain machines could not occupy, perhaps with us, our ecological niche and we could not occupy theirs.

Acknowledgements

Ancestors of this paper have been presented at 11th *International Scientific Conference on Neuroethics* and 6th *Conference of the Italian Society for Neuroethics* (*SINe*) and at the *Italian Association of Cognitive Science (AISC)* – both in 2019. The authors thank the audience for questions and suggestions, and in particular Mario De Caro, Michele Di Francesco and Massimo Reichlin. Marco Viola read and commented a later version. We are grateful to him for his remarks and criticisms. We thank also two anonymous referees for this journal, who raised helpful criticisms and attracted our attention to points that needed further clarification. The usual caveats hold. The authors discussed, reviewed, and approved the entire manuscript together. Despite being the outcome of a collective effort, the division of labour in this paper was as follows: G. Pellegrino wrote the second part of the paper and half of the first, M.D. Garasic the first part.

Notes

[1] Cf. N. BOSTROM, E. YIDKOWSKY, *The ethics of artificial intelligence*, in: K. FRANKISH, W.M. RAMSEY (eds.), *The Cambridge handbook of artificial intelligence*, Cambridge University Press, Cambridge 2014, pp. 316-334, here pp. 321-322; L. FLORIDI, J.W. SANDERS, *On the morality of artificial agents*, in: «Minds & Machines», vol. XIV, n. 3, 2004, pp. 349-379, here pp. 349-352; see also L. FLORIDI, *The ethics of information*, Oxford University Press, Oxford 2013, pp. 148-152.

[2] For a naturalist argument about the connection between membership in human species and possessing intentionality, see M. TOMASELLO, *The origins of human communication*, MIT Press, Cambridge (MA) 2009.

[3] L. FLORIDI, *The Ethics of Information*, cit., p. 110.

[4] *Ibid*., p. 140. Adaptability is connected with machine learning; see *ibid*., p. 145.

[5] Cf. *ibid*., pp. 135, 151, 157.

[6] *Ibid*., p. 151.

[7] *Ibid*., pp. 159-160.

[8] Cf. A. CLARK, D. CHALMERS, *The extended mind*, in: «Analysis», vol. LVIII, n. 1, 1998, pp. 7-19; A. CLARK, *Supersizing the mind: Embodiment, action, and cognitive extension*, Oxford University Press, Oxford 2008; A. CLARK, *Memento's revenge: The extended mind extended*, in: R. MENARY (ed.), *The extended mind*, MIT Press, Cambridge (MA) 2010, pp. 43-66.

[9] On these issues, see D. DE GRAZIA, *Human identity and bioethics*, Cambridge University Press, Cambridge 2005; J. MCMAHAN, *The ethics of killing: Problems at the margins of life*, Oxford University Press, New York 2002.

[10] These topics have elicited a growing scholarship, where defenses and objections have been exchanged; see, for instance, M. COLOMBO, E. IRVINE, M. STAPLETON (eds.), *Andy Clark and his critics*, Oxford University Press, Oxford 2019, and P. FRENCH, *Collective and corporate responsibility*, Columbia University Press, New York 1984; M. GILBERT, *Sociality and responsibility. New essays in plural subject theory*, Rowman & Littlefield, Lanham 2000; C. LIST, P. PETTITT, *Group agency: The possibility, design, and status of corporate agents*, Oxford University Press, New York 2011.

[11] In this paper, we assume that knowledge can be obtained, and that we can recognize a case of knowledge when we see it. Both assumptions are controversial. Skeptics deny the first assumption,

and well-known puzzles about the necessary conditions of knowledge, in particular the ones elicited by a famous paper by E. Gettier, beset the second assumption: see E.L. GETTIER, *Is justified true belief knowledge?*, in: «Analysis», vol. XXIII, n. 6, 1963, pp. 121-123. Here, we do not tackle these issues. Also, we do not tackle the relations between knowledge, intelligence, understanding, intentionality, and conscience. It will be evident that we do not ascribe conscience or sentience to robots and AIs, and that our answer to AMR is not meant to establish whether robots and AIs could ever have self-awareness. Likewise, we assume that knowledge and intelligence without conscience, or sapience without sentience, are possible. Again, these are contentious issues, which we cannot deal with here.

[12] Cf. S. BRINGSJORD, N. SUNDAR GOVINDARAJULU, *Artificial Intelligence*, in: E.N. ZALTA (ed.), *Stanford encyclopedia of philosophy*, 2018 [accessed 2 February 2019].

[13] Cf. J. TASIOULAS, *First steps towards an ethics of robots and artificial intelligence*, in: «Journal of Practical Ethics», vol. VII, n. 1, 2019, pp. 61-95, §4.2. For a general presentation of the ethics of artificial intelligence, see also N. BOSTROM, E. YIDKOWSKY, *The ethics of artificial intelligence*, cit.

[14] Cf. N. BOSTROM, E. YIDKOWSKY, *The ethics of artificial intelligence*, cit., p. 322; G. OPPY, D. DOWE, *The Turing test*, in: E.N. ZALTA (ed.), *The Stanford encyclopedia of philosophy*, §2.4 [accessed 12 October 2019].

[15] Cf. N. BOSTROM, E. YIDKOWSKY, *The ethics of artificial intelligence*, cit., pp. 321-324.

[16] The principle stated here is not uncontroversial. It may be interpreted as a strong form of supervenience, and it can be challenged as a form of reductionism. In this paper, we are not going to consider these and similar issues.

[17] J. TASIOULAS, *First steps towards an ethics of robots and artificial intelligence*, cit., pp. 64, 70-71.

[18] Cf. N. BOSTROM, E. YIDKOWSKY, *The ethics of artificial intelligence*, cit., p. 316; J. TASIOULAS, *First steps towards an ethics of robots and artificial intelligence*, cit., pp. 76-80; D. WIGGINS, *Continuants: Their activity, their being, and their identity*, Oxford University Press, Oxford 2016, p. 91.

[19] Cf. J. TASIOULAS, *First steps towards an ethics of robots and artificial intelligence*, cit., p. 79; UNESCO, *Report of COMEST on robotics ethics - UNESCO Digital Library*, Paris 2017, p. 39 [accessed 11 October 2019].

[20] Cf. N. BOSTROM, *Superintelligence: Paths, dangers, strategies*, Oxford University Press, Oxford 2014; N. BOSTROM, E. YIDKOWSKY, *The ethics of artificial intelligence*, cit., pp. 328-332; G. OPPY, D. DOWE, *The Turing test*, cit., § 2.2; J. TASIOULAS, *First steps towards an ethics of robots and artificial intelligence*, cit., pp. 88-89.

[21] A paradigmatic defense of this view is in J.R. SEARLE, *The rediscovery of the mind*, MIT Press, Cambridge (MA) 1992.

[22] Cf. R.M. FRENCH, *Subcognition and the limits of the Turing test*, in: «Mind», vol. XCIX, n. 393, 1990, pp. 53-65.

[23] Cf. A.M. TURING, *Computing machinery and intelligence*, in: «Mind», vol. LIX, n. 236, 1950, pp. 433-460, here p. 444; L. FLORIDI, *The ethics of information*, cit., pp. 13-14.

[24] Cf. L. FLORIDI, J.W. SANDERS, *On the morality of artificial agents*, cit., pp. 366-369.

[25] Cf. T. NAGEL, *Justice and nature*, in: «Oxford Journal of Legal Studies», vol. XVII, n. 2, 1997, pp. 303-321; C.D. STONE, *Should trees have standing? Law, morality, and the environment*, Oxford University Press, Oxford 2010; L.H. TRIBE, *Ways not to think about plastic trees: New foundations for environmental law*, in: «Yale Law Journal», vol. LXXXIII, 1974, pp. 1315-1346.

[26] Cf. J. TASIOULAS, *First Steps towards an ethics of robots and artificial intelligence*, cit., pp. 82-83, L. FLORIDI, *The ethics of information*, cit., p. 154.

[27] Cf. *ibid*., p. 146. This answer can be off the mark, as it is not uncontroversial that human beings are necessarily persons. As said in §1 above, it is logically possible to have non-human persons or human non-persons. Here, we do not tackle these issues.

[28] Cf. R. KURZWEIL, *We are becoming cyborgs* [accessed 11 October 2019]; R. KURZWEIL, *The singularity is near: When humans transcend biology*, Duckworth, London 2006; J. TASIOULAS, *First steps towards an ethics of robots and artificial intelligence*, cit., pp. 81-82.

[29] Cf. A. CLARK, *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*, Oxford University Press, Oxford 2003; M. MORE, N. VITA-MORE, *The transhumanist reader: Classical and contemporary essays on the science, technology, and philosophy of the human future*, Wiley-Blackwell, London 2013.

[30] Cf. G. OPPY, D. DOWE, *The Turing test*, cit.; A.M. TURING, *Computing machinery and intelligence*, cit.

[31] L. FLORIDI, *The ethics of information*, cit., pp. 146-147.

[32] *Ibid.*, p. 147.

[33] Floridi better clarifies his view on causal accountability in a later article, where he regards causal accountability as an instance of strict liability or faultless responsibility; see L. FLORIDI, *Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions*, in: «Philosophical Transactions of the Royal Society», 2016, Art.Nr. 20160112.

[34] Cf. J. SEARLE, *Minds, brains, and programs*, in: «Behavioural & Brain Sciences», vol. III, n. 3, 1981, pp. 417-457; J. SEARLE, *Minds, brains and science: The 1984 Reith Lectures*, Harvard University Press, Cambridge (MA) 1984.

[35] A. CLARK, *Supersizing the mind*, cit., p. xxvii.

[36] A. CLARK, D. CHALMERS, *The extended mind*, cit., pp. 8-9 - emphases added.

[37] A. CLARK, *Supersizing the mind*, cit., p. 76 - emphasis added.

[38] Cf. R. MENARY, *The extended mind*, cit., p. 2.

[39] Cf. L. FLORIDI, *The ethics of information*, cit., pp. 137-138, 146.

[40] Cf. *ibid.*, chap. 13. Floridi discusses his preferred instances of distributed morality in §13.4.

[41] *Ibid.*, p. 262.

[42] *Ibid.*, p. 265.

[43] L. FLORIDI, *Faultless responsibility*, cit., p. 4.

[44] For references, see *supra*, fn. 10.

[45] Cf. B. HUEBNER, *Macrocognition. A theory of distributed cognition and collective intentionality*, Oxford University Press, Oxford 2014.

[46] A. CLARK, *Intrinsic content, active memory and the extended mind*, in: «Analysis», vol. LXV, n. 285, 2005, pp. 1-11, here p. 4.

[47] Cf. *supra*, § 1.

[48] Cf. F. ADAMS, K. AIZAWA, *The bounds of cognition*, in: «Philosophical Psychology», vol. XIV, n. 1, 2001, pp. 43-64; F. ADAMS, K. AIZAWA, *Why the mind is still in the head*, in: P. ROBBINS, M. AYDEDE (eds.), *The Cambridge handbook of situated cognition*, Cambridge University Press, Cambridge 2009, pp. 78-95; F. ADAMS, K. AIZAWA, *Defending the bounds of cognition*, in: R. MENARY (ed.), *The extended mind*, cit. pp. 66-80.

[49] F. ADAMS, K. AIZAWA, *Defending the bounds of cognition*, cit., p. 68.

[50] *Ibidem.*

[51] *Ibidem*.

[52] Cf. A. CLARK, *Coupling, constitution, and the cognitive kind: A reply to Adams and Aizawa*, in: R. MENARY (ed.), *The extended mind*, cit., pp. 81-99, here p. 84.

[53] *Ibid.*, pp. 84-85.

[54] *Ibid.*, p. 83.

[55] *Ibid.*, p. 89.

[56] F. ADAMS, K. AIZAWA, *Defending the bounds of cognition*, cit., p. 75.

[57] Cf. N. BOSTROM, E. YIDKOWSKY, *The ethics of artificial intelligence*, cit., pp. 322, 323.

[58] Cf. R. MENARY (ed.), *The extended mind*, cit.

[59] A.M. TURING, *Computing machinery and intelligence*, cit., p. 446.

[60] Cf., for a similar argument, L. FLORIDI, *The ethics of information*, cit., p. 148.

[61] Cf., for a similar argument, *ibid.*, p. 154.

[62] Cf. A. CLARK, *Surfing uncertainty. Prediction, action, and the embodied mind*, Oxford University Press, Oxford 2016, pp. xiii-xiv.

References

ADAMS, F., AIZAWA, K. (2001). *The bounds of cognition*. In: «Philosophical Psychology», vol. XIV, n. 1, pp. 43-64.

ADAMS, F., AIZAWA, K. (2009). *Why the mind is still in the head*. In: P. ROBBINS, M. AYDEDE (eds.), *The Cambridge handbook of situated cognition*, Cambridge University Press, Cambridge, pp. 78-95.

BOSTROM, N. (2014). *Superintelligence: Paths, dangers, strategies*, Oxford University Press, Oxford.

BOSTROM, N., YIDKOWSKY, E. (2014). *The ethics of artificial intelligence*. In: K. FRANKISH, W.M. RAMSEY (eds.), *The Cambridge handbook of artificial intelligence*, Cambridge University Press, Cambridge, pp. 316-334.

BRINGSJORD, S., SUNDAR GOVINDARAJULU, N. (2018). *Artificial intelligence*. In: E.N. ZALTA (ed.), *Stanford encyclopedia of philosophy*, Spring Edition, URL: <https://plato.stanford.edu/entries/artificial-intelligence/>.

CLARK, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*, Oxford University Press, Oxford.

CLARK, A. (2005). *Intrinsic content, active memory and the extended mind*. In: «Analysis», vol. LXV, n. 285, pp. 1-11.

CLARK, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*, Oxford University Press, Oxford.

CLARK, A. (2010). *Memento's revenge: The extended mind extended*. In: R. MENARY (ed.), *The extended mind*, MIT Press, Cambridge (MA), pp. 43-66.

CLARK, A. (2010). *Coupling, constitution, and the cognitive kind: A reply to Adams and Aizawa*. In: R. MENARY (ed.), *The extended mind*, MIT Press, Cambridge (MA), pp. 81-99.

CLARK, A. (2016). *Surfing uncertainty. Prediction, action, and the embodied mind*, Oxford University Press, Oxford.

CLARK, A., CHALMERS, D. (1998). *The extended mind*. In: «Analysis», vol. LVIII, n. 1, pp. 7-19.

COLOMBO, M., IRVINE, E., STAPLETON, M. (eds.) (2019). *Andy Clark and his critics*, Oxford University Press, Oxford.

DEGRAZIA, D. (2005). *Human identity and bioethics*, Cambridge University Press, Cambridge.

FLORIDI, L. (2013). *The ethics of information*, Oxford University Press, Oxford.

FLORIDI, L. (2016). *Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions*. In: «Philosophical Transactions of the Royal Society», vol. CCCLXIV, n. 2083, Art.Nr. 20160112 – doi: 10.1098/rsta.2016.0112.

FLORIDI, L., SANDERS, S.W. (2004). *On the morality of artificial agents.* In: «Minds & Machines», vol. XIV, n. 3, pp. 349-379.

FRENCH, P. (1984). *Collective and corporate responsibility*, Columbia University Press, New York.

FRENCH, R.M. (1990). *Subcognition and the limits of the Turing test*. In: «Mind», vol. XCIX, n. 393, pp. 53-65.

GETTIER, E.L. (1963). *Is justified true belief knowledge?*. In: «Analysis», vol. XXIII, n. 6, pp. 121-123.

GILBERT, M. (2000). *Sociality and responsibility. New essays in plural subject theory*, Rowman & Littlefield, Lanham.

HUEBNER, B. (2014). *Macrocognition. A theory of distributed cognition and collective intentionality*, Oxford University Press, Oxford.

KURZWEIL, R. (2002). *We are becoming cyborgs*, URL:<https://www.kurzweilai.net/we-are-becoming-cyborgs>.

KURZWEIL, R. (2006). *The singularity is near: When humans transcend biology*, Duckworth, London.

LIST, C., PETTITT, P. (2011). *Group agency: The possibility, design, and status of corporate agents*, Oxford University Press, New York.

MCMAHAN, J. (2002). *The ethics of killing: Problems at the margins of life*, Oxford University Press, New York.

MORE, M., VITA-MORE, N. (2013). *The transhumanist reader: Classical and contemporary essays on the science, technology, and philosophy of the human future*, Wiley-Blackwell, London.

NAGEL, T. (1997). *Justice and nature*. In: «Oxford Journal of Legal Studies», vol. XVII, n. 2, pp. 303-321.

OPPY, G., DOWE, D. (2019). *The Turing test.* In: E.N. ZALTA (ed.), *The Stanford encyclopedia of philosophy*, Spring Edition, URL: <https://plato.stanford.edu/archives/spr2019/entriesuring-test/>.

SEARLE, J.R. (1981). *Minds, brains, and programs*. In: «Behavioural & Brain Sciences», vol. III, n. 3, pp. 417-457.

SEARLE, J.R. (1984). *Minds, brains and science: The 1984 Reith Lectures*, Harvard University Press, Cambridge.

SEARLE, J.R. (1992). *The rediscovery of the mind*, MIT Press, Cambridge (MA).

STONE, C.D. (2010). *Should trees have standing? Law, morality, and the environment*, Oxford University Press, Oxford.

TASIOULAS, J. (2019). *First steps towards an ethics of robots and artificial intelligence*. In: «Journal of Practical Ethics», vol. VII, n. 1, pp. 61-95.

TOMASELLO, M. (2009). *The origins of human communication*, MIT Press, Cambridge (MA).

TRIBE, L.H. (1974). *Ways not to think about plastic trees: New foundations for environmental law*. In: «Yale Law Journal», vol. LXXXIII, pp. 1315-1346.

TURING, A.M. (1950). *Computing machinery and intelligence.* In: «Mind», vol. LIX, n. 236, pp. 433-460.

UNESCO (2017). *Report of COMEST on robotics ethics - UNESCO Digital Library*, Paris, URL: <https://unesdoc.unesco.org/ark:/48223/pf0000253952>.

WIGGINS, D. (2016). *Continuants: Their activity, their being, and their identity*, Oxford University Press, Oxford.