Philosophy & Technology
https://doi.org/10.1007/s13347-021-00462-7

RESEARCH ARTICLE

Can Autonomous Agents Without Phenomenal Consciousness Be Morally Responsible?

László Bernáth[1,2]

Received: 11 February 2021 / Accepted: 16 June 2021

Abstract
It is an increasingly popular view among philosophers that moral responsibility can, in principle, be attributed to unconscious autonomous agents. This trend is already remarkable in itself, but it is even more interesting that most proponents of this view provide more or less the same argument to support their position. I argue that as it stands, the Extension Argument, as I call it, is not sufficient to establish the thesis that unconscious autonomous agents can be morally responsible. I attempt to show that the Extension Argument should overcome especially strong ethical considerations; moreover, its epistemological grounds are not too solid, partly because the justifications of its premises are in conflict.

Keywords Autonomous agents · Consciousness · Consequentialism · Extension Argument · Moral responsibility

1 Introduction

Autonomous agents (AAs) are programs or robots that can successfully carry out relatively complex problem-solving tasks without the intervention of a human agent. These AAs — such as self-driving cars, caring robots, autonomous weapon systems, and sophisticated automated trading systems — are intelligent in the sense that their problem-solving methods are rather effective and flexible, and superficially imitate human problem-solving processes. However, the consensual opinion is that it would

✉ László Bernáth
bernathlaszlo11@gmail.com

[1] Institute of Philosophy, Research Centre for the Humanities, Tóth Kálmán u 4, 1097 Budapest, Hungary

[2] Faculty of Humanities, Institute of Philosophy, Eötvös Loránd University, Múzeum krt. 4, 1088 Budapest, Hungary

Published online: 11 July 2021

be an exaggeration to attribute phenomenal consciousness (hereafter consciousness) to them. Thus, virtually everyone agrees that it would be mistaken to suppose that the world or any event in it *appears* to these AAs in a qualitative way.[1]

At first glance, the lack of consciousness rules out the possibility that these AAs have more than mere *causal* agency and responsibility (which can be attributed even to earthquakes and tornados). Even if one believes that groups can be agents and responsible beings in a robust sense, it seems to be intuitively plausible that the consciousness of the members of the group is a condition for the group to have robust agency and responsibility.

Nevertheless, intuitions and seemings can be misleading, and a growing number of philosophers claim that this is precisely the case with the robust agency and responsibility of AAs. They claim that nothing in principle rules out the possibility that AAs are agents and responsible beings in some (but still somewhat limited) sense (Floridi & Sanders, 2004; Stahl, 2006 Hellström 2013). And since they do not argue that this is so because these AAs can be regarded as conscious beings,[2] they believe that the unconsciousness of AAs is not something that rules out their robust agency and (some kind of) responsibility.

In this paper, I will focus on the part of the literature that is even more ambitious. Some philosophers claimed that not only some kind of agency and responsibility but even moral responsibility can in principle be attributed to AAs. However, I will argue that the Extension Argument (EA), as I call it — which is deployed by philosophers (Bechtel, 1985; Dennett, 1997; Hage, 2017; Sullins, 2006) who welcome the attribution of moral responsibility to unconscious AAs — is not sufficiently strong to outweigh the considerations against the attribution of moral responsibility to them because the epistemic bases of the EA are shaky.

Firstly, I will outline two basic considerations against attributing moral responsibility to unconscious AAs. Secondly, I will reconstruct the Extension Argument. Thirdly, I will argue that there is an internal epistemic tension between the premises of the Extension Argument, and one of them is particularly vulnerable to a general epistemic objection that is often raised against consequentialist approaches. Finally, I will conclude that the EA — as it stands — is not strong enough, and it is not easy to see how one could revamp it so that it can overcome the two basic ethical considerations against attributing moral responsibility to unconscious AAs.

[1] I do not use the term "qualia" because many philosophers define it in a way that is too restrictive for the purposes of the present paper. For example, Dennett (1990) defines "qualia" as intrinsic mental properties which cannot be described from a third person point of view and which can be known to their subjects without error. However, when I talk about phenomenal events/states, I refer to mental events/states that have phenomenal character. If a mental event/state has phenomenal character, there is something it is like for the subject to have the experience. Furthermore, an agent has phenomenal consciousness if she experiences such phenomenal events/states on some occasions. (See Tye 2018 for the different definitions of 'qualia').

[2] It is true that some proponents of the (possible) moral responsibility of AAs claim that consciousness is necessary for moral responsibility. However, they do not refer to phenomenal consciousness by using the term "consciousness." For example, Daniel Dennett uses the term "consciousness" to refer to a hierarchical structure of propositional attitudes. According to Dennett, an agent is conscious if she has beliefs about beliefs, desires about desires, and so on in a specific, complex way (Dennett 1991, 1997).

1.1 Two Basic Considerations Against Attributing Moral Responsibility to Autonomous Agents

An agent is morally responsible if she (or it) can, in principle, be appropriately blamed (or praised)[3] for violating (or discharging) a moral obligation.[4] Based on Tigard 2020[5] that categorizes those different notions of moral responsibility which can be found in the literature about AI, this definition defines a "descriptive" notion of moral responsibility because it is about what is needed for having a moral status that has in itself a neutral value rather than a notion of moral responsibility that implies obligations or possessing a virtue.

Furthermore, this definition uses a concept of moral blame that is neither a mere evaluation of an action, nor a characteristic trait from, say, a legal, an epistemic, or a prudential point of view. Rather, moral blame is a reactive attitude which targets the agent based on her moral misbehavior or vice in a hostile kind of way. Blame is often colored by anger, indignation, condemnation, or even a desire for punishing the agent. At the least, the blamer prefers that the blamed one feels remorse about her wrongdoing/vice, which gives grounds for blaming her (on this point, see Fricker, 2016).

Note, the definition ties moral responsibility not to the accessible possibility of appropriate blame but to its "in principle" possibility. It can be the case that blaming an agent in an appropriate way is impossible due to the circumstances. For example, the wrongdoer cannot be appropriately blamed if no-one knows about her wrong deeds or if all of who could blame her committed the same moral mistake. Nevertheless, it is clear that if someone had the sufficient epistemic or moral position to blame the wrongdoer, she could appropriately blame her and this is why she is morally responsible. Thus, the condition of moral responsibility is not the accessible possibility of appropriate blame, but the in principle possibility of it.

The reason why the attribution of moral responsibility to AAs seems, even in principle, to be absurd to many is not only that blaming mundane machines and

[3] I put the term "praise" into brackets because the paper, alike to the literature on moral responsibility in general, focuses on blame-related aspects of moral responsibility rather than its more inspiring side. However, in some context, moral praiseworthiness has a central role in attributing moral responsibility to agents. For example, in philosophy of religion, it makes sense to attribute moral responsibility to God who cannot be blamed (because He cannot violate any moral obligation) since God can, in principle, appropriately praised and worshipped for being perfectly good. Still, praise and praiseworthiness are much less relevant than blame and blameworthiness from the perspective of the paper because misdirected praise is much less worrying than misdirected blame. The latter results in much more injustice than the former. I would like to thank an anonymous reviewer that s/he pointed out the relevance of praiseworthiness in the general context of responsibility-attribution.

[4] This is a Strawsonian definition of moral responsibility, which is rather neutral until one specifies the sense in which the morally responsible agent is, in principle, an appropriate target of blame (see, among others, Strawson 1962, Smith 2008, Scanlon 2008). Some theories claim that the agent has to be able to *deserve* blame (Pereboom 2014) or be a *fair* target of blame (Wallace 1994) in order to be morally responsible. Other theories — such as consequentialist ones — claim that a sufficient condition for being morally responsible is that the blaming of the agent should tend to produce positive consequences (Smart 1961; Vargas 2013).

[5] I would like to thank an anonymous reviewer for drawing my attention to this paper.

software is a ridiculous practice in our culture. Rather, the close connection between moral blame and moral responsibility makes it difficult to conceive how attributing moral responsibility to objects would make any sense at all. In the literature, two complementary aspects of this problem are captured by two arguments against AAs' responsibility. Robert Sparrow puts the first aspect of the problem in the following way.

> Why should it be so hard to imagine holding a machine responsible for its actions? One reason is that it is hard to imagine how we would hold a machine responsible — or, to put it another way, what would follow from holding it to be responsible. To hold that someone is morally responsible is to hold that they are the appropriate locus of blame or praise and consequently for punishment or reward. A crucial condition of the appropriateness of punishment or reward is the conceptual possibility of these treatments. Thus in order to be able to hold a machine morally responsible for its actions it must be possible for us to imagine punishing or rewarding it. Yet how would we go about punishing or rewarding a machine? (Sparrow, 2007, 71)

Blame can be regarded as a mild form of punishment, or an attitude that goes hand in hand with a disposition to punish the blamed one — at least — in a mild manner (Bernáth, 2020). Furthermore, as Sparrow points out, if to be morally responsible is to be, in principle, an appropriate target of punishment such as blame or forms of punishments that are intimately related to blame, then it makes no sense to attribute moral responsibility to beings that are immune to punishment. However, unconscious beings cannot, even in principle, be punished in any way because they are unconscious and do not feel pain or suffering.[6] So, unconscious beings cannot have moral responsibility.[7]

This argument is rather strong because one can resist it only if she denies that blame aims to change behavior through inducing moral emotions like shame and guilt (Carlsson, 2019). Denying this is a costly maneuver because it is a robust moral intuition that seems to be independent of the theoretical framework on which one relies. Imagine that you blame someone for her misconduct, but she reacts in the following way. She tells you that she knows that you blame her and she gets that you have a good reason to be angry with her. However, she does not feel anything, neither shame nor blame, she just sees that she should have acted differently and she promises that she will act better in the future. Regardless of

[6] One of the anonymous reviewers of a previous version of the paper objected that even dead people cannot be punished. Thus, according to Sparrow's and my definition of moral responsibility, they should not be morally responsible because they cannot be the appropriate target of blame-related punishments. My reply to this worry is that the individuals who are dead now are, in principle, appropriate target of blame because it could have been happened in the past that someone appropriately punishes them in a blame-related way. I would like to thank the anonymous reviewer to press me to clarify this point.

[7] However, the argument does not imply that unconscious AAs cannot have moral rights, and we are not obligated to avoid harming them. The relation between moral rights and consciousness is less clear than the relation between moral responsibility and consciousness. Gunkel (2014) points out the problems of the theories of moral rights with regard to whether we should attribute moral rights to AAs.

your moral theoretical framework, you would be not only baffled due to the oddness of the scene, but you would think about whether there was a point in blaming her at all since, it seems, it would be sufficient to *tell* her why her action was unacceptable. Or, if her action was not only morally bad but evil, your anger would not evaporate and you still wanted that she feels herself worse. Of course, I do not claim that one cannot deny that blaming primarily aims at inducing painful moral emotions. What I claim is that one has to provide a strong argument against Sparrow's consideration because it is based on a moral intuition that seems to be rather reliable and not theory-driven.

As far as I know, the second aspect of the problem that arises from the lack of consciousness is not outlined as clearly as the previous one. However, Duncan Purves, Ryan Jenkins, and Bradley J. Strawser come pretty close to it when they argue as follows.

> If either the *desire-belief* model or the predominant *taking as a reason* model of acting for a reason is true, then AI cannot in principle act for reasons. Each of these models ultimately requires that an agent possess an attitude of belief or desire (or some further propositional attitude) in order to act for a reason. AI possesses neither of these features of ordinary human agents. AI mimics human moral behavior, but *cannot take a moral consideration such as a child's suffering to be a reason for acting*. AI *cannot be motivated to act morally*; it simply manifests an automated response which is entirely determined by the list of rules that it is programmed to follow. Therefore, AI cannot act for reasons, in this sense. Because AI cannot act for reasons, it cannot act for the right reasons. (Purves et al., 2015, 896; italics added)

Purves, Jenkins, and Strawser argue that autonomous weapon systems (AWSs) should not be deployed in war because they cannot reason morally. Nevertheless, it is obvious that if something cannot reason morally, then it cannot be morally responsible, because this kind of inability exempts it from having moral obligations. Moreover, Purves, Jenkins, and Strawser apparently argue merely for the conclusion that the lack of specific propositional attitudes (such as beliefs and desires) makes AWSs incapable of moral reasoning. But if propositional attitudes are mere dispositions — as many argue — then AAs (including AWSs) can in principle have beliefs and desires precisely because they can mimic the behavior of human agents. On closer inspection, however, it is clear that the authors rely on an approach to reasons which claims that having a reason (or taking something as a reason) requires a *non-automatic* or *non-programmed* recognition of something as a reason. And this non-automatic recognition of reasons cannot be conceived without the reasoner being (consciously) aware of such events as a child's suffering.

I would like to put this point in another way. Even if human behavior is predetermined by past events and physical laws, and even if humans uncritically adopt many moral principles, their moral systems and their attitudes toward these systems are shaped (to some extent) by their conscious experiences. They do not only follow moral rules but — due to their conscious experiences of suffering, compassion, guilt, pain, joy, feeling of autonomy and freedom — can also understand *why and*

for what reason these moral rules should be followed. For example, an unconscious caring robot can follow or (if it malfunctions) disregard the rules of ethical caring, but it does not have a clue as to *why and for what reason* following the rules is good.[8] However, every human knows to some extent why it is so important to appropriately care about other people, since they know from experience how much suffering indifference and rudeness can cause and how much joy proper caring produces. Besides that, people are capable of modifying their inherited moral systems in the light of conscious experiences.

I regard the argument from moral blame and the one from moral reasons as the two strongest arguments against the possibility of morally responsible unconscious AAs. Both of them arise from pre-philosophical and fundamental considerations about the role of consciousness in morality.[9] Of course, even fundamental considerations can be wrong and others can outweigh them. However, these other considerations should be particularly strong because the pre-philosophical considerations for the impossibility of morally responsible unconscious AAs are rather plausible in themselves.

This is why it is not surprising that the most common reasoning in favor of the (possible) moral responsibility of unconscious AAs almost completely casts doubt on the pre-philosophical understanding of morality. More precisely, the argument — which I call Extension Argument (EA) — is based on two approaches that are used for casting doubt on our intuitive understanding of morality in many other contexts: consequentialism and scientism. However, I will argue that the consequentialist and the scientistic bases of the EA are in epistemological tension with each other, and that one of its premises is *especially* vulnerable to an epistemic worry about consequentialism.

2 The Extension Argument

In this chapter, I reconstruct the Extension Argument, which is apparently the most popular reasoning among those philosophers who argue that there is, in principle, no obstacle to attributing moral responsibility to AAs. One can find fully fledged versions

[8] It is worthwhile to note that the incapability of unconscious AAs to understand why and for what reason following the rules is good does not imply that they have to be unable to appropriately follow these rules and properly bring decisions in morally complex situations. Even though there are numerous challenges, it may be feasible to produce AAs that are sufficiently sensitive to moral considerations. (See, for an excellent overview, Wallach et al., 2008. I would like to thank an anonymous reviewer for drawing my attention to this paper).

[9] There are other objections against the moral responsibility of AAs but I do not regard them as strong as the two objections that I highlighted. Based on incompatibilist considerations, some argue that AAs cannot be morally responsible because their behavior is determined by external forces or deterministic natural laws (Bringsjord, 2008; Hew 2014). Others argue on the basis of particular compatibilist theories that AAs cannot be morally responsible (Hakli & Mäkelä 2019). The problem with these arguments is that it is hard to tell to what extent they are based on theory-driven assertions, and in which regards they are based on real moral experiences. I believe that the arguments from moral blame and moral reasoning much less likely have theoretical or perspectivistic roots than the arguments from incompatibilism or historical compatibilism.

of the Extension Argument (Hage, 2017 and Dennett, 1997 combined with Dennett, 1971, 1984, and 1991)[10] and incomplete ones as well (Bechtel, 1985; Sullins, 2006). To reconstruct the Extension Argument, I will start with showing how these versions attempt to overcome the ethical considerations against attributing moral responsibility to unconscious AAs.

Insofar as phenomenal consciousness has an indispensable role in moral reasoning and responsibility practices, the moral responsibility of unconscious AAs seems to be inconceivable. Thus, a proponent of the (possible) moral responsibility of unconscious AAs has to downplay the role of consciousness on solid grounds. In most cases, these solid grounds are the sciences. Here is a telling quote from Jaap Hage.

> There are many reasons to assume that intention and free will do not play a role in the chains of facts and events that constitute the physical aspects of acts. The physical research paradigm which assumes that physical events are only linked to other physical events in a law-like fashion, works quite well and leaves no room for intervening mental phenomena like intention or will. It is completely unclear how physical events might be influenced by mental events as such, and there is no evidence that such an influence exists. It should be emphasized in this connection that the two words 'as such' in the previous sentence are crucially important. The possibility should not be excluded beforehand—and actually it is quite likely—that intentions and will have counterparts in brain states. These brain states play a role in the processes of the physical world. However, this does not mean that intention and will as mental phenomena (as 'qualia'; […]) influence the physical world.
> Therefore, even if humans act on the basis of intentions and free will—whatever that might mean—while autonomous agents do not, this does not make a difference for what happens physically. It is for this reason not at all obvious that the alleged difference between humans and autonomous agents should mean that only human agents are held responsible for their acts and autonomous agents not (Hage, 2017, 258)

Hage argues that phenomenal consciousness (and qualia) does not have any causal effects on our behavior, and our knowledge of it is based on our best scientific theories. Thus, neither the lack of phenomenally conscious free will or intention, nor any conscious experience can make a relevant metaphysical difference between AAs and humans with regard to moral responsibility. If conscious experiences as such do not cause anything in the physical world, they certainly do not cause our attribution of moral responsibility to each other. In other words, if anything at all provides a good reason for attributing moral responsibility to each other, it cannot be phenomenal consciousness because it is causally inert.

Hage's argumentation is admittedly inspired by Daniel Dennett's approach to consciousness. Let us see a passage from Dennett's *Consciousness Explained*

[10] One has to read Dennett's other works to get the full line of argumentation besides Dennett 1997 that addresses the moral responsibility of AAs directly.

that reveals the broader scientific and philosophical background of Hage's argumentation.

> The Cartesian Theater may be a comforting image because it preserves the reality/appearance distinction at the heart of human subjectivity, but *as well as being scientifically unmotivated*, this is metaphysically dubious, because it creates the bizarre category of the objectively subjective — the way things actually, objectively seem to you even if they don't seem to seem that way to you! […] Some thinkers have their faces set so hard against "verificationism" and "operationalism" that they want to deny it even in the one arena where it makes manifest good sense: the realm of subjectivity. […] We might classify the Multiple Drafts model [Dennett's model for consciousness], then, as *first-person operationalism*, for it brusquely denies the possibility in principle of consciousness of a stimulus in the absence of the subject's belief in that consciousness. (Dennett, 1991, 132. Italics are added)

The Cartesian Theater — in Dennett's terms — is the apparent place in which phenomenal experiences appear to a unified and responsible agent: that is, to a self. However, Dennett takes scientific results to show that the supposition of such a Cartesian Theater is unmotivated, just as the supposition of such phenomenal experiences that occur *independently* of *our beliefs*. Dennett and Hage agree that if one can, from a physical point of view, successfully explain actions as physical events without reference to any conscious experience, then phenomenal consciousness as such (or in itself) has no role in the explanation of actions. Rather, at best, *beliefs* about phenomenal experiences are the ones which can have any role in any plausible causal story about actions. So, we *"merely" attribute* (rather than discover) not only causally efficient phenomenal experiences, intentions, unified agency, self, and a Cartesian Theater but also moral responsibility to ourselves. As Hage explains:

> If agency and causation are a matter of attribution, then responsibility must be a matter of attribution too. If one considers responsibility as a mind-independent entity in the 'outside world', this responsibility does not exist. There is no responsibility to be discovered in the 'outside world' analogously to the way we can discover a pond in the forest or a birthmark on somebody's skin. We cannot discover that somebody was responsible for some act, although we can discover facts that are grounds for attributing responsibility to somebody. Responsibility is best accounted for with the attributivist view, and is then the result of attribution, rather than a 'real' phenomenon. (Hage, 2017, 261)

To put it another way, from a physical and metaphysically objective point of view, there are no morally responsible agents. Thus, if one would like to see morally responsible agents, one has to adopt a different perspective. According to Dennett and Hage, even if AAs are not morally responsible but humans are, it cannot be because AAs do not possess some fancy metaphysical features which humans do. The lesson drawn from this way of thinking about scientific results and moral responsibility can be summarized as follows.

No Difference Premise There are no moral responsibility-relevant metaphysical differences between highly developed unconscious AAs and fully developed human beings.

However, even if it is *possible* to regard some physical systems as morally responsible entities, the question still remains: why should we do this after the recognition of the fact that the perspective of science and objective metaphysics does not contain moral responsibility? Note that this is a relevant problem for anyone wishing to argue for the possible moral responsibility of unconscious AAs, even if some proponents of the AAs' responsibility by and large ignore it (Bechtel, 1985; Sullins, 2006). This is because the lack of moral responsibility-relevant metaphysical differences between humans and AAs leaves open the possibility of putting humans into the category of non-responsible bio-robots instead of elevating some machines to the ranks of responsible humans. Dennett provides a classic consequentialist reasoning for why we should stick to the practice of attributing moral responsibility to humans even if moral responsibility is not part of the metaphysical landscape.

> Any finite control system (such as a human brain) will always be prone to making mistakes or arriving at decisions that a more leisurely analysis would condemn; it is an inevitable feature of human character, even perfected to its limit. Original Sin, naturalized. It is wise, however, to adopt policies that minimize the bad effects of these inevitable defects of character. […] By somewhat arbitrarily holding people responsible for their actions, and making sure they realize that they will be held responsible, we constrain the risk-taking in the design (and redesign) of their characters within tolerable bounds. When in spite of these best measures people get caught in wrong deeds, their gambles (wise or foolish) have simply lost and they ought not to object to paying the assigned penalty. (Dennett, 1984, 165)

In short, we should attribute responsibility to each other because that practice has good consequences. To serve as the second premise of the Extension Argument, I can put it as follows:

Consequentialist Responsibility Premise If there are no moral responsibility-relevant metaphysical differences between unconscious AAs and humans, then we should attribute (or not attribute) moral responsibility to unconscious AAs and humans based on the overall value of outcomes of such a practice.

In the light of the Consequentialist Responsibility Premise, it is straightforward that if the attribution of moral responsibility to AAs has good consequences, one should do so (since even in the case of humans, the fact that they do not fulfill fancy metaphysical conditions of responsibility is quite unimportant). As Hage explains:

> The practice of attributing responsibility and liability may be justified if this attribution brings desirable consequences. […] An intelligent program may possess knowledge about its potential responsibility and take this knowledge into account in deciding what it will do. This knowledge may be generally

> available, but may also be the result of being having [*sic*] been held responsible on a particular occasion. The adaptation of behavior to potential or actual responsibility presupposes that the agent is not focused on a single task such as taking a particular kind of administrative decisions or conducting e-trade, but that it performs tasks like that in the context of wider tasks such as contributing to the well-being of society, or the maximization of its profits. Presently there are, to the author's knowledge, no practically functioning systems which can do this, but for the theoretical question that is not very relevant. If such systems would exist—and it is quite likely that they can already be created—it would make sense to hold them responsible for their doings, both in the abstract as well as in concrete cases. (Hage, 2017, 268-269)[11]

So the third premise of the Extension Argument can be formulated as follows:

Possibility Premise It may, in principle, turn out that the outcomes of attributing moral responsibility to unconscious AAs would be positive overall.

Still, in a sense, there seems to be a difference between being an appropriate target of responsibility-attribution and being morally responsible. On the face of it, it is conceivable that attributing moral responsibility to an agent is appropriate because doing so is beneficial for everyone while the agent in question is not morally responsible at all. However, the proponents of the Extension Argument suggest that this alleged difference is, even if it exists, ultimately irrelevant. Dennett dedicated a full monograph to demonstrating this irrelevance, and he summarizes his view as follows.

> Surely, it seems, we can make a distinction between the question of why we *hold* people responsible, or *take* responsibility ourselves for various things, and the question of why or whether we actually *are* responsible. […]
> But whatever responsibility is, considered as a metaphysical state, unless we can tie it to some recognizable social desideratum, it will have no rational claim on our esteem. Why would anyone care whether or not he had the property of responsibility (for some particular deed, or in general)? […] Instead of investigating, endlessly, in an attempt to *discover* whether or not a particular trait is of someone's making – instead of trying to assay exactly to what degree a particular self is self-made – we simply *hold* people responsible for their conduct (within limits we take care not to examine it too closely). And we are rewarded for adopting this strategy by the higher proportion of "responsible"

[11] Even though I cannot show texts to prove that, but I have the impression that many who like the idea of attributing moral responsibility to AAs do that not because they believe that attributing moral responsibility to AAs makes these machines better beings or improves moral behavior in any other way but because they worry about the responsibility gap. That is, they worry about that if AAs are not morally responsible for the harms they cause, nobody will be responsible for them. I would like to mention that one can deal with this issue without promoting the need of attributing moral responsibility to AAs. See, for instance, Schulzke 2013 and Champagne & Tonkens 2015.

> [that is, morally acceptable] behavior we thereby inculcate. (Dennett 1984, 164)

In other words, Dennett says that being morally responsible has any relevance inasmuch as it has social consequences. Because "metaphysical" moral responsibility does not have such relevance, we should regard socially relevant "consequentialist" moral responsibility as moral responsibility *simpliciter*. Thus, the fourth and final premise of the Extension Argument can be formulated in the following way.

Revised Responsibility Premise If the outcomes of attributing moral responsibility to a being-type *B* are positive overall, then *B* is morally responsible.

Based on the four reconstructed premises, one can conclude that unconscious AAs can, in principle, be morally responsible. To sum up the Extension Argument, I provide its full form below. (As I previously mentioned, the full form of the argument is endorsed by Hage, 2017 and Dennett, 1997 combined with Dennett, 1971, 1983, and 1991, whereas incomplete versions of it can be found in e. g. Bechtel, 1985, and Sullins, 2006).

(a) There are no moral responsibility-relevant metaphysical differences between highly developed unconscious AAs and fully developed human beings. [No Difference Premise]
(b) If there are no moral responsibility-relevant metaphysical differences between unconscious AAs and humans, then we should (or should not) attribute moral responsibility to unconscious AAs and humans based on the overall value of the outcomes of such a practice. [Consequentialist Responsibility Premise]
(c) It may, in principle, turn out that the outcomes of attributing moral responsibility to unconscious AAs would be positive overall. [Possibility Premise]
(d) If the outcomes of attributing moral responsibility to a being-type *B* are positive overall, then *B* is morally responsible. [Revised Responsibility Premise]

∴ It may, in principle, turn out that unconscious AAs are morally responsible.[12]

In the next section, I will argue that there is an epistemological tension between the first, science-based premise and the other consequentialist premises. In other words, I will attempt to show that it is hard to ground all premises in a coherent way.

[12] It is interesting that a recent outstanding paper (Behdadi & Munthe 2020) that excellently reviews the arguments and theories with regard to the moral agency and responsibility of AAs does not recognize the argumentation and the position of the proponents of EA. I think this is because it is difficult to see the difference between the position of the proponents of EA and the positions of two other camps. There is a camp of which members argue that we should attribute something alike to moral responsibility that has less robust normative implications (for example, as I mentioned, Floridi & Sanders 2004, Stahl 2006, Hellström 2013). There is another approach that seems to suggest that it is open to attribute a kind of responsibility to AAs that has the same normative role as moral responsibility has, but it is rather clear about that it does not identify this kind of responsibility with full-blown moral responsibility (Coeckelbergh 2009, Coeckelbergh 2010). Beyond the similarity between those positions and the approach of the proponents of EA, the fact that there is only one recent paper (Hage 2017) that openly defends EA makes it rather difficult to recognize its distinctive features.

Furthermore, the consequentialist premises of the Extension Argument are especially vulnerable to general objections against consequentialism.

3 Problems with the Extension Argument

3.1 The Problem of Epistemic Coherence

It is easy to see that proponents of the Extension Argument justify the No Difference Premise and the other premises in different ways. On the one hand, they support the first premise by appealing to *scientific* findings in order to reject *ethical considerations* in favor of the idea that phenomenal consciousness is the precondition of being morally responsible. On the other hand, the Consequentialist Responsibility Premise and the Revised Responsibility Premise hinge on *ethical* considerations, while the Possibility Premise is based on *epistemic* ones without any reference to scientific results. Given that ethical considerations about the indispensable role of consciousness with regard to moral responsibility are *prima facie* plausible and widely shared yet (allegedly) *unreliable* in light of various *scientific findings*, the problem arises why one who accepts the No Difference Premise should rely on *any* ethical considerations, including those that support the Consequentialist Responsibility Premise and the Revised Responsibility Premise. To put it differently, if the ethical considerations about the role of consciousness in constituting morally responsible agents are unreliable, how could one trust in the seemingly less robust and less widely shared consequentialist considerations about the truth of the Consequentialist Responsibility Premise and the Revised Responsibility Premise?

As far as I can tell, no proponents of the Extension Argument have addressed this worry. Nonetheless, one can reasonably figure out what they could say in defense of the Extension Argument. In the next subsection, I will briefly analyze two possible strategies against this objection which, as I will argue, do not have a real chance of success. After that, in a separate subsection, I will examine the most promising way of responding to the epistemic challenge, and attempt to show why this defense is not so plausible either.

3.1.1 Two Possible Strategies for Overcoming the Epistemic Challenge to EA — and Why They Are Unsuccessful

First, the proponents of EA could argue that they can back up the Consequentialist Responsibility Premise and the Revised Responsibility Premise with scientific results. However, I do not have any hope in the success of this strategy. The Consequentialist Responsibility Premise seems to be a purely ethical proposition that has nothing to do with any field of scientific research, while the Revised Responsibility Premise can be based on a conceptual analysis that is driven, as Dennett's

reasoning demonstrates, by convictions about the (ir)relevance of certain factors/considerations. Both of them are based on the *ethical* idea that if most of our views about the key notions of ethics are inapplicable to our ethical practices in the light of scientific findings, then we should rethink these key notions and our ethical practices themselves from the only remaining ethical perspective: namely, from the consequentialist one. Of course, no scientific research can tell us what the *main theoretical grounds* of any ethical reform should be.

Second, the proponents of EA could claim that the Consequentialist Responsibility Premise and the Revised Responsibility Premise have stronger ethical support than the belief in the indispensable role of consciousness. However, in the light of the two arguments for the indispensable role of consciousness in constituting morally responsible agents, this strategy seems to be doomed to failure. On the one hand, the arguments based on blame and moral reasoning rely on fundamental ethical considerations about moral responsibility which have a strong grip on us and are rarely contested. On the other hand, ethical intuitions in favor of the Consequentialist Responsibility Premise and the Revised Responsibility Premise are strongly contested, and it is less clear how compelling they are. This is because there are strong and widely shared ethical considerations against the Consequentialist Responsibility Premise and the Revised Responsibility Premise, and it is not so clear whether the consequentialist arguments can outweigh them.

Against the Consequentialist Responsibility Premise, there is an ethical consideration which says that if there is no moral responsibility-relevant metaphysical difference between AAs and humans, then we should (or should not) attribute moral responsibility to humans and AAs only if they both have the metaphysical property of being morally responsible. And this remains the case even if neither of them has this metaphysical property regardless of the consequences, because it would be unfair to attribute moral responsibility to agents who are in fact not morally responsible and do not have the metaphysical basis for deserving blame and punishment.

Against the Revised Responsibility Premise, there is a strong intuition that "being morally responsible" and "being someone to whom attributing moral responsibility is useful" are distinct properties partly because "being morally responsible for event *E*" does not seem to be a future-directed property. In *everyday moral practice*, if one tries to answer the question whether one is morally responsible for *E*, one usually asks questions about what the case was when or before *E* happened, but no one asks what *will happen* if the wrongdoer is blamed or punished for *E*.

I do not claim that the consequentialist arguments for the truth of the Consequentialist Responsibility Premise and the Revised Responsibility Premise are wrong. Rather, I am arguing that the ethical considerations for the Consequentialist Responsibility Premise and the Revised Responsibility Premise do not seem to be more reliable than the ethical considerations against the moral responsibility of unconscious agents, because the former seem to be shakier from an ethical point of view than the latter. So, if one cannot trust in ethical considerations for the indispensability of phenomenal consciousness, one cannot trust in the Consequentialist Responsibility Premise and the Revised Responsibility Premise either.

3.1.2 The Defeaters Strategy

Finally, proponents of the Extension Argument could argue that there is no epistemic tension between the scientific justification of the No Difference Premise and the justifications of the two ethical premises because ethical considerations are in general *reliable*. Scientific findings are "only" possible *defeaters* of ethical considerations. Since there are no actual scientific defeaters of the latter two premises, whereas certain scientific results defeat the view that there are moral responsibility-relevant differences between highly developed unconscious AAs and humans, one can accept all premises of the Extension Argument without any epistemological tension.

This is by far the most promising strategy for the proponents of EA, but its use creates a problem. Insofar as ethical considerations are in general reliable, the scientific basis for the No Difference Premise needs to be extremely solid because it has to override ethical considerations that have seemingly rather solid bases (recall the arguments from blame and moral reasoning for the claim that having consciousness is relevant with regard to moral responsibility). However, there are reasons to doubt that the scientific results provide such a strong basis for denying the relevance of phenomenal consciousness to moral responsibility.

At this point, it is worthwhile to recall that the defense of the No Difference Premise rests on the claim that the causal efficiency of phenomenal consciousness is allegedly overridden by the findings of cognitive science. Once again, as Hage puts it: "The physical research paradigm which assumes that physical events are only linked to other physical events in a law-like fashion, works quite well and leaves no room for intervening mental phenomena like intention or will. It is completely unclear how physical events might be influenced by mental events as such, and there is no evidence that such an influence exists." (Hage, 2017, 258). Now, if phenomenal consciousness has no causal role whatsoever, its presence cannot make any moral responsibility-relevant metaphysical difference as compared to unconscious AAs.

To begin with, there are notable proponents of forms of non-reductive physicalism (or physicalist property dualism; for a useful overview see Robb & Heil, 2019) and even substance dualism (see Moreland, 2018; Swinburne, 2018) who accept the causal efficiency of phenomenal consciousness. They argue either that the most plausible form of the causal closure of the physical realm is compatible with the causal efficiency of the mental or that the causal closure of the physical is not proven and we have good reasons for denying it. For example, the obviousness of mental causation can be such a reason for denying the causal closure of the physical. The evaluation of contemporary forms of dualism lies outside the scope of the present paper, but to say the least, the mere existence of high-quality defenses of interactionist dualism already casts doubt on the claim that science can provide rationally quasi-irresistible arguments against the causal efficiency of phenomenal consciousness.

Moreover, there are detailed objections to the claim that cognitive science has shown that phenomenal consciousness plays no causal role in human behavior (see Mele, 2009, Walter, 2011, Shields, 2014, Brass, Furstenberg & Mele 2019, Bernáth, 2019). For example, they argue that neuroscientific experiments did not show that

unconscious brain states determine voluntary actions because they did not close out the possibility that these unconscious brain states prepare the voluntary action without deciding which possible course of action will be done. Unfortunately, once again, this topic lies outside the scope of the present paper. But the mere existence of high-quality scholarly criticisms of neuroscience-based arguments against the causal efficiency of phenomenal consciousness casts doubt on the claim that science has the final word on this matter.

Instead of surveying the scholarly objections to neuroscience-based arguments for the inefficiency of phenomenal consciousness, I would like to put forward a simple argument in order to make it a bit clearer how problematic it is to draw any science-based conclusions about the causal inefficiency of phenomenal consciousness. Let us modify Hage's quote in the following way.

> The physical research paradigm which assumes that physical events are only linked to other physical events in a law-like fashion, works quite well and leaves no room for emerging mental phenomena like intention or will. It is completely unclear how physical events give rise to phenomenal mental events as such, and there is no evidence that phenomenal mental events exist.

The problem is that Hage *could* claim that all of the above is true for the same reasons as he asserts the causal inefficiency of consciousness. From a purely scientific perspective, there is no reason to suppose the existence of phenomenal consciousness. Like in the case of causal efficiency, it is rather unclear (and not only from a cognitive science perspective) why and how phenomenal consciousness comes into existence. Thus, if the explanatory success of the physical research paradigm and the lack of scientific evidence for the causal efficiency of phenomenal consciousness provide sufficient reason to deny the causal efficiency of phenomenal consciousness, then the success of this research paradigm and the lack of scientific evidence provide sufficient reason for denying the existence of phenomenal consciousness.

To avoid this unappealing conclusion, the proponent of EA could claim that the first-person perspective provides sufficient evidence for the existence of phenomenal consciousness, since we experience phenomenal mental events rather obviously. However, if one regards the first-person perspective as an epistemic one that can provide sufficient evidence for a belief even in the face of notable scientific difficulties, then the first-person perspective can also justify the belief in the causal efficiency of phenomenal consciousness. This is because it seems to be rather obvious from a first-person perspective that when I hit my finger with a hammer, the unpleasant phenomenal characteristic of pain is one of the causes of my pulling back my hand.

Even though most philosophers regard the denial of phenomenal consciousness as an absurd position, many of them hold absurd views with regard to this or that (at least those who elaborate their philosophical systems in detail). So why should we not allow the proponents of EA to deny even the existence of phenomenal consciousness, if this is what they need to be able to defend their view on moral responsibility? To put it bluntly, because it makes not much sense to endorse consequentialism if there are no ethically relevant consequences. Hedonistic

consequentialism, which is probably the most popular form of naturalistic consequentialism, claims that one should evaluate an action based on how many negative or positive phenomenal mental events (such as pain, suffering, joy, pleasure) it causes. And if there are no such mental events with their distinctive phenomenal qualities, then there are no ethically relevant differences between actions from a consequentialist perspective. Even if one endorses preference consequentialism (this is the main alternative to the hedonistic one, which claims that the value of an action is based on whether it satisfies preferences or not), it is hard to see why any action could be good or bad if satisfying a preference is *never* satisfactory. Now, it seems that the most full and coherent defense of EA should include an ethical argument for this rather curious type of preference consequentialism according to which satisfying a preference is good/bad completely regardless of what it is like to satisfy a preference. However, I do not know how one could do it in a plausible way if she admits at the same time that ethical considerations are, in general, reliable (as the proponent of the defeater strategy does). Because it is a rather robust ethical consideration that the goodness of satisfying a preference has to do something with the what-it-is likeness of it.

To sum up, if the proponents of EA regard ethical and other commonsense considerations as rather reliable in order to preserve the epistemic credibility of their consequentialist premises, they have to provide *extremely solid evidence* for the causal inefficiency of phenomenal consciousness, a thesis that justifies the No Difference Premise. However, in the light of some high-quality defenses of interactionalist dualism and heavy criticism of the arguments for the inefficiency thesis, it is unlikely that they can produce such a strong evidence. Insofar as they argue that the lack of scientific evidence for the causal efficiency of phenomenal consciousness — together with the theoretical difficulties of integrating a causally efficient phenomenal consciousness into the scientific picture of the world — provides, in itself, extremely strong evidence against the causal efficiency of phenomenal consciousness, they should deny the mere existence of phenomenal consciousness as well. However, this maneuver undermines their consequentialist premises and attitude, or at least, it makes very difficult to defend them.

3.2 Epistemic Problems with the Third Premise of EA

Let me repeat the third premise of EA.

Possibility Premise It can, in principle, turn out that the outcomes of attributing moral responsibility to unconscious AAs would be positive overall.

First off, it should be pointed out that the Possibility Premise claims not only that attributing moral responsibility to unconscious AAs can be positive overall (which is rather trivial), but also that it can *turn out* that this is the case. No proponent of EA claims that we should start to attribute moral responsibility to unconscious AAs blindly. Rather, they insist that if we have sufficient evidence that this practice has good consequences, then we should encourage it.

However, there is a good reason to doubt whether the overall value of the consequences of attributing moral responsibility to unconscious AAs could turn out either way. One of the most formidable arguments against consequentialism is the epistemic objection. According to this, because in most cases it is impossible to know which course of action will have the most positive consequences, consequentialist considerations are useless for someone who wishes to decide how one should act (see Lenman, 2000; Elgin, 2015).

Some respond to the epistemic objection by claiming that consequentialism gives the criteria for moral goodness, and not practical guides for making morally right decisions. Given that the proponents of EA talk about the possible revision of our responsibility-practices *in the light of the* consequences, this kind of defense is irrelevant in this context.

Others claim that agents should choose those options which *seem to* have the most positive consequences *from the agent's epistemic point of view*. The proponents of EA should endorse something like that. Nevertheless, the epistemic problem of estimating the consequences of attributing moral responsibility to unconscious AAs is still daunting. How could we make a *reasonable* estimation of whether the practice of holding unconscious AAs morally responsible would have positive consequences or not? To be sure, attributing moral responsibility to unconscious agents would be a huge change in morality, with wide-ranging effects both on our understanding of morality and everyday life. How could we reasonably believe that we are able to estimate the later consequences of such a change?

In making this point, it is worthwhile to stress that philosophers cannot settle even the debate about whether attributing moral responsibility to humans has good consequences. Some philosophers argue that it would be better to abandon practices that are tied to moral responsibility, such as blaming and retributive punishment (see, for example, Waller, 2011; Pereboom, 2014), while others recommend that we should stick to the concept of moral responsibility even if moral responsibility is not real at all, because abandoning responsibility-practices would have unwelcome effects (see, for instance, Smilansky, 2000).

I think that part of the best explanation why philosophers cannot agree whether abandoning responsibility-practices would be beneficial or harmful is that this matter cannot be settled from an armchair. Rather, if we attempt to estimate the overall value of the consequences of attributing moral responsibility to humans or AAs, we should do it empirically. Yet, even using empirical tools, the success of this undertaking is highly doubtful because we cannot investigate the effects at issue in a laboratory, but only by collecting data about social processes and phenomena. Before doing these empirically based social investigations, no one has any good basis for arguing for or against the thesis that these empirical investigations are sufficiently reliable regarding the overall value of responsibility attribution. So, instead of baldly asserting the Possibility Premise, we should be *agnostic about* whether the overall value of attributing moral responsibility to AAs (and for that matter, even to humans) can *turn out either way*. Thus, not only there is an epistemic tension between the justification of the No Difference Premise, the Consequentialist Responsibility Premise and the Revised Responsibility Premise, but also the Possibility Premise is based on problematic epistemic assumptions.

I consider the epistemological issues with the Extension Argument so problematic that I would suggest the proponents of the moral responsibility of phenomenally unconscious AAs an alternative strategy.[13] Instead of relying on the alleged causal inertness of phenomenal mental events and consequentialist considerations, they should deploy a Strawsonian strategy (see Strawson, 1962). That is, they should argue that if people spontaneously adopted a practice that morally blames AAs, AAs would be morally responsible because the moral practice of the moral community determines which types of entities are morally responsible irrespectively of their metaphysical features (Coeckelbergh, 2014 claims something very similar about the moral status of AAs while Coeckelbergh, 2009 argues in a somewhat Strawsonian fashion that not full-blown, virtual responsibility can be attributed to AAs). This strategy would be less ambitious because it cannot aim at the philosophy-based reform of our moral practice. It can justify only the possible future changes of moral attitudes toward AAs. Moreover, it is not even sure that it could secure the possibility of morally responsible, but phenomenally unconscious AAs. This is because it can be the case that adopting a moral stance, in which we attribute moral responsibility toward an entity, necessarily includes regarding the entity in question as phenomenally conscious (Coeckelbergh, 2009 points out the importance of that, and perhaps it is not a coincidence that Coeckelbergh who has the most similar approach to this sketched Strawsonian strategy does not want to attribute full-blown, "real" moral responsibility to AAs). Still, I would suggest the proponents of the view according to which AAs without phenomenal consciousness may be morally responsible beings that they give this Strawsonian strategy a shot rather than salvaging the Extension Argument.

4 Conclusion

Given the strength of two basic considerations against attributing moral responsibility to unconscious AAs and the epistemic problems of EA, there are no sufficiently strong arguments for the thesis that unconscious AAs can be morally responsible. One can find other, somewhat similar arguments in the literature, but they are less ambitious than EA, since they do not argue for the possibility of attributing full-blown *moral* responsibility to AAs, but only for attributing less robust kinds of responsibility to them (Stahl, 2006; Coeckelbergh, 2009, Hellström, 2013). They should be examined separately. But even if they successfully establish some kind of responsibility of AAs, their success does not necessary mean that one can provide an argument for the *moral responsibility* of unconscious AAs, because the argument has to overcome powerful ethical considerations against it. I hope that the analysis of EA shows why it is so difficult to undermine the reliability of rather strong and basic ethical considerations and, at the same time, to provide *other ethical* considerations in favor of the moral responsibility of AAs. Not to mention the problem that there are good reasons for thinking that no one knows how the attribution of moral responsibility to unconscious AAs would play out in the long run. Thus, it is hard to

[13] I would like to thank an anonymous reviewer for pointing out this alternative strategy.

provide any consequentialist argument for the possibility of the moral responsibility of unconscious AAs.

Acknowledgements I would like to thank Boldizsár Eszes, Alena Popa, Anna Réz, Judit Szalai, and Zsófia Zvolenszky for their suggestions and the stimulating exchanges we had on many occasions. I owe special thanks to the research project entitled "Meant to be: Resuscitating the Metaphysics of Teleology" that is hosted by the CEU and supported by the Ian Ramsey Centre and the John Templeton Foundation.

Funding Open access funding provided by Eötvös Loránd University. The research was supported by János Bolyai Research Scholarship of the Hungarian Academy of Sciences (grant no. BO/00432/18/2), the OTKA (Hungarian Scientific Research Fund by the National Research Development and Innovation Office) Postdoctoral Excellence Programme (grant no. PD131998), the Higher Education Institutional Excellence Grant (Autonomous Vehicles, Automation, Normativity: Logical and Ethical Issues) at the Eötvös Loránd University, and two other OTKA research grants (grant no. K132911, K123839).

References

Bernáth, L. (2019). Why libet-style experiments cannot refute all forms of libertarianism. In B. Feltz, M. Missal, A. C. Sims (Eds.), *Free will, causality, and neuroscience*, 97–119. Leiden, The Netherlands: Brill. https://doi.org/10.1163/9789004409965_007.

Bernáth, L. (2020). Blame and fault: Toward a new conative theory of blame. *Disputatio: International Journal of Philosophy, 12*(59), 371–394.

Bechtel, W. (1985). Attributing responsibility to computer systems. *Metaphilosophy, 16*(4), 296–306.

Behdadi, D., & Munthe, C. (2020). A normative approach to artificial moral agency. *Minds and Machines, 30*, 195–218.

Bringsjord, S. (2008). Ethical robots: The future can heed us. *AI & Society, 22*(4), 539–550.

Brass, M., Furstenberg, A., & Mele, A. (2019). Why neuroscience does not disprove free will. *Neuroscience and Biobehavioral Reviews, 102*, 251–263.

Carlsson, A. B. (2019). Shame and attributability. In D. Shoemaker (Ed.), *Oxford Studies in Agency and Responsibility* (6th ed., pp. 112–139). Oxford University Press.

Champagne, M., & Tonkens, R. (2015). Bridging the responsibility gap in automated warfare. *Philosophy & Technology, 28*(1), 125–137. https://doi.org/10.1007/s13347-013-0138-3.

Coeckelbergh, M. (2014). The moral standing of machines: Towards a relational and non-Cartesian moral hermeneutics. *Philosophy & Technology, 27*, 61–77. https://doi.org/10.1007/s13347-013-0133-8.

Coeckelbergh, M. (2009). Virtual moral agency, virtual moral responsibility: On the moral significance of the appearance, perception, and performance of artificial agents. *AI & Society, 24*(2), 181–189. https://doi.org/10.1007/s00146-009-0208-3.

Dennett, D. C. (1984). *Elbow room: The varieties of free will worth wanting*. Cambridge, Mass, MIT Press.

Dennett, D. C. (1990). Quining Qualia. In W. Lycan (Ed.), *Mind and Cognition* (pp. 519–548). Blackwell.

Dennett, D. C. (1997). When HAL Kills, Who's to blame? In D. G. Stork (Ed.), *HAL's Legacy: 2001's Computer as dream and reality* (pp. 351–366). Massachusetts, MIT Press.

Dennett, D. C. (1991). *Consciousness explained*. Little, Brown and Company.

Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy, 8*, 87–106.

Lenman, J. (2000). Consequentialism and cluelessness. *Philosophy & Public Affairs, 29*, 342–370.

Elgin, S. (2015). The unreliability of foreseeable consequences: A return to the epistemic objection. *Ethical Theory and Moral Practice, 18*, 759–766. https://doi.org/10.1007/s10677-015-9602-8.
Floridi, L., & Sanders, J. (2004). On the morality of artificial agents. *Minds and Machines, 14*, 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d.
Fricker, M. (2016). What's the point of blame? A Paradigm Based Explanation. *Noûs, 50*(1), 165–183.
Gunkel, D. J. (2014). A vindication of the rights of machines. *Philosophy & Technology, 27*(1), 113–132. https://doi.org/10.1007/s13347-013-0121-z.
Hakli, R., & Mäkelä, P. (2019). Moral responsibility of robots and hybrid agents. *The Monist, 102*(2), 259–275.
Hage, J. (2017). Theoretical foundations for the responsibility of autonomous agents. *Artifical Intelligence and Law, 25*, 255–271. https://doi.org/10.1007/s10506-017-9208-7.
Hellström, T. (2013). On the moral responsibility of military robots. *Ethics and Information Technology, 15*, 99–107. https://doi.org/10.1007/s10676-012-9301-2.
Mele, A. (2009). *Effective intentions: The power of conscious will*. Oxford University Press.
Moreland, J. P. (2018). In defense of a Thomistic-like dualism. In J. J. Loose, A. J. L. Menuge, & J. P. Moreland (Eds.), *The Blackwell companion to substance dualism* (pp. 102–122). Wiley-Blackwell.
Pereboom, D. (2014). Free will, agency, and meaning in life. *Oxford, Oxford University Press*. https://doi.org/10.1093/acprof:oso/9780199685516.001.0001.
Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice, 18*(4), 851–872.
Robb, D. & Heil, J. (2019). Mental Causation. In *The Stanford encyclopedia of philosophy* (Summer 2019 Edition), Edward N. Zalta (ed.), accessed November 7, 2020 from https://plato.stanford.edu/archives/sum2019/entries/mental-causation/.
Scanlon, T. (2008). *Moral dimensions: Permissibility, meaning, blame*. Belknap, Harvard University Press.
Schulzke, M. (2013). Autonomous weapons and distributed responsibility. *Philosophy & Technology, 26*(2), 203–219.
Shields, G. S. (2014). Neuroscience and conscious causation: Has neuroscience shown that we cannot control our own actions? *Review of Philosophy and Psychology, 5*(4), 565–582.
Smart, J. J. (1961). Free-will, praise and blame. *Mind, 70*(279), 291–306.
Smilansky, S. (2000). *Free will and illusion*. Oxford University Press.
Stahl, B. C. (2006). Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. *Ethics and Information Technology, 8*, 205–213. https://doi.org/10.1007/s10676-006-9112-4.
Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77.
Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy, 48*, 1–25.
Smith, A. M. (2008). Control, responsibility, and moral assessment. *Philosophical Studies, 138*(3), 367–392. https://doi.org/10.1007/s11098-006-9048-x.
Sullins, J. P. (2006). When is a robot a moral agent. *International Review of Information Ethics, 6*(12), 23–30.
Swinburne, R. (2018). Cartesian substance dualism. *The Blackwell Companion to Substance Dualism,* J. J. Loose, A. J. L. Menuge & J. P. Moreland (eds.), Oxford: Wiley-Blackwell, 133–151.
Tye, M. (2018). Qualia. In *The Stanford encyclopedia of philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), Retrieved November 7, 2020, from https://plato.stanford.edu/archives/sum2018/entries/qualia/.
Vargas, M. (2013). *Building better beings: A theory of moral responsibility*. Oxford University Press.
Wallace, R. J. (1994). *Responsibility and the moral sentiments*. Harvard University Press.
Wallach, W., Allen, C., & Smit, I. (2008). Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *AI & Society, 4*(22), 565–582.
Waller, B. N. (2011). *Against moral responsibility*. MIT Press.
Walter, H. (2011). Contributions of neuroscience to the free will debate: From random movement to intelligible action. In *The Oxford Handbook of Free Will,* 2nd edition, R. Kane (ed.), Oxford University Press