Journal for General Philosophy of Science
https://doi.org/10.1007/s10838-023-09666-1

ARTICLE

The Locus of Agency in Extended Cognitive Systems

Barbara Tomczyk[1]

Accepted: 28 September 2023

Abstract
The increasing popularity of artificial cognitive enhancements raises the issue of their impact on the agent's personal autonomy, and issues pertaining to how the latter is to be secured. The *extended mind* thesis implies that mental states responsible for autonomous action can be partly constituted by the workings of cognitive artifacts themselves, and the question then arises of whether this commits one to embracing an *extended agent* thesis. My answer is negative. After briefly presenting the main accounts on the conditions for autonomous agency, and analyzing how the latter can be protected from threats posed by the use of cognitive artifacts, I argue that autonomous agency is essentially tied to conscious experience and intentionality, which in turn can only be attributed to the human part of any extended cognitive system. I present both theoretical (conceptual) and practical arguments against recognizing the entire extended system, composed of one human and an artifact, as an autonomous agent.

Keywords Autonomous agency · Extended mind · Coupled cognitive system · Artificial cognitive enhancement · Epistemic and moral responsibility · Conscious experience

1 Introduction

The idea of the extended mind, which has received support from the tremendous progress made in the field of artificial cognitive enhancements, has come to inspire animated discussion in many areas of philosophy and cognitive science. Its consequences are far-reaching, in that it challenges the standard, internalist understanding of autonomous agency that makes reference to conscious mental states, epistemic and moral responsibility, creditability and blameworthiness, sense of effort and other properties that seem exclusively attributable to human beings in the context of human–machine interaction (Frankfurt 1971; Chisholm 1976; Taylor 1966).[1]

[1] To be clear, I do not exclude the possibility that some of these features can be attributed to non-human animals. I do not undertake this issue in this article. By the phrase "exclusively attributable to human beings" I mean not attributable to artifacts.

✉ Barbara Tomczyk
barbara.tomczyk@mail.umcs.pl

[1] Faculty of Philosophy and Sociology, Maria Curie-Sklodowska University in Lublin, Lublin, Poland

Published online: 09 December 2023

The extended mind thesis was proposed by Andy Clark and David Chalmers who argue that in some cases of cognitive activity a person is coupled with an external artifact with such a dense, reciprocal, causal interaction (*continuous reciprocal causation*), that they together constitute one cognitive system. Both, a human agent and an artifact, insofar as they remain in such a relation, play an active causal role in the cognitive process that could result in an extended mental state, such as dispositional belief that is realized partly beyond the human's organism (Clark, Chalmers 1998). Clark and Chalmers originally justified their thesis via the example of Otto, a person suffering from Alzheimer disease, who is replacing his internal memory with information in a notebook. The authors of "Extended Mind" argue that items in the notebook play the same functional role for Otto as biological memory plays for a healthy agent. Assuming the "parity principle" (Clark, Chalmers 1998, 8),[2] that brings the idea that cognitive processes and mental states are identified by what they do, rather than by the material events that realize them, an item in Otto's notebook carries the content of his dispositional belief.

Clark and Chalmers' functionalist interpretation of the extended mind thesis was seriously objected. Critics pointed namely on essential functional differences between Otto's notebook and internally represented information (Adams, Aizawa 2001, Rupert 2009).[3] Responding to these objections, proponents of the extended mind thesis proposed new understandings of an extended cognitive system, independent of functionalism, the most influential of which are complementarity approaches, also known as second-wave views (Menary 2010; Sutton 2010; Farina 2021; Farina, Lavazza 2022a, 2022b), third-wave views (Sutton 2010) and arguments grounded on predictive processing theory (Kirchhof, Kiverstein 2019).[4] Critics of Clark and Chalmers' functionalist interpretation argue, for example, that a cognitive artifact constitutes with a human organism one cognitive system not because it plays the same function as internal elements, but rather because it complements existing internal functions. Functional difference between internal and external processes is not an obstacle to the formation of an extended cognitive system, rather it makes the external element valuable for the system's cognitive efficacy (Sutton 2010; Menary 2010). Hence, an extended cognitive system should be understood as a single cognitive unit of analysis in which neural and external resources make complementary contribution to bringing about intelligent behavior. Moreover, advocates of second and third-wave accounts, broaden the set of elements that constitute human cognitive processes to include social and cultural factors. Philosophers such as Richard Menary (Menary 2010), Edwin Hutchins (2011), Lambros Malafouris (2008) among others, argue that every cognitive activity that a given person undertakes is constituted by cognitive practices shaped by cultural norms and cognitive institutions like legal systems (Gallagher, Crisafi 2009). Such normative cognitive practices can be, as they argue, vehicles of cognition even though they do not satisfy the parity principle, for they cannot, even in principle, be done in the head. Such an argumentation goes much further than Clark and Chalmers' toward embedding cognitive processes in the social and cultural environment.

[2] "If, as we confront some task, a part of the world functions as a process which, were it done in the head, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world is (so we claim) part of the cognitive process. Cognitive processes ain't (all) in the head!" (Clark, Chalmers 1998, 8).

[3] The most frequently cited differences are that the notebook is subject to other-party manipulation, external representations are assessed though perceptual system, external representations do not possess non-derived content and they do not exhibit the same effects as internal memory (e.g. recency, primacy, chunking effects).

[4] The terminology of 'first-wave' and 'second-wave' arguments is suggested by John Sutton (2010).

Advocates of various versions of the extended mind thesis refer to continuous reciprocal causation between internal and external resources as the relation that underlines an extended cognitive system. A cognitive artifact, that composes such a system, is a vehicle of external representations that affect, and often change, natural human cognitive abilities. The human agent, in turn, manipulates the workings of the artifact, and in this way such causal reciprocity comes to be realized. There are plenty of such devices around us, the most common of which are smartphones with their diverse applications, and computers with software programs. When using these, the agent offloads a certain amount of information onto the artifact and it, in turn, determines the actions they will undertake next. There is a cumulative information flow between the agent and the cognitive artifact, and each step in the extended cognitive process depends on the previous ones. It is almost impossible to distinguish internal and external parts of the process, and so makes more sense to conceive of the agent and an artifact as one cognitive system (Clark, Chalmers 1998; Clark 2010; Menary 2010; Sutton 2010; Hutchins 2011; Kiverstein, Farina 2011; Heersmink 2012; Carter 2021).

Admitting the existence of extended cognitive systems, however, does not necessarily entail the extended mind thesis. This second view is much more radical, as it enables not only partial realization of cognitive processes outside the human organism, but also extracranial realization of mental states, which poses the risk of cognitive bloat (Adams, Aizawa 2001). In order to protect a mind from the threat of uncontrollable spreading into the world, Clark and Chalmers impose four constraints on extended dispositional belief known as 'glue and trust' conditions that include: constancy of use, facility of access, trust and prior endorsement (the fact that the information has been consciously endorsed in the past) (Clark, Chalmers 1998). Not everyone agrees however that these conditions are sufficient to recognize the artifact as part of the agent's mind. Kim Sterelny, for example, proposes an additional condition of entrenchment and personalization, which says that an artifact must be customized to an agent's individual usage and moreover the agent's cognitive routines have to be altered to incorporate the resulting personalized artifact (Sterelny 2010). Thus, an agent plays an active role in making the external resource part of their own mind, which results in the experience of *mineness* towards it. Another condition—epistemic possession—is proposed by Robert Clowes. It indicates that a cognitive artifact should only be considered as a part of an agent mind when it is minimally cognitively penetrable, policable and revisable by the agent (Clowes 2015). Hence, an artifact should be open to some sort of scrutiny when it is needed, it cannot be completely out of an agent's control. Strengthened with these six conditions, the extended mind thesis sounds much more convincing as far as highly trusted, individualised and entrenched human-artifact system is concerned. The crucial question is however, what is the influence of such deeply incorporated artefact on human's autonomous agency. Do they together constitute one extended agent, or is agency restricted to the human component? In other words, what is the locus of autonomous agency in the extended cognitive system?

Human use of cognitive enhancement does not always result in the constitution of an extended system. In the broadest sense, any method that has the effect of improving the functioning of the human cognitive system could be recognized as a cognitive enhancement. They can be divided into natural, such as learning, meditation, and mnemonics, and artificial, which include the use of pharmacology, artificial intelligence and genetic modifications. In this paper, I am referring only to artificial cognitive enhancements. Such artifacts are designed to improve both the sensitivity of human senses and the intellectual efficiency related to the memory, intelligence and creativity and even to the control over emotions, mood and desires (Sandberg, Bostrom 2006). Artificial cognitive enhancements

can be divided into those that directly stimulate neuronal processes responsible for specific cognitive states, such as psychoactive substances or implants placed in appropriate areas of the brain,[5] and those that are external to the human body. The latter usually function as memory stores, data mining analysis and visualization programs, which support the process of reasoning, imagining and decision making. Artifacts connected with the human body or implemented inside it, enter into close and often reciprocal causal relations with the brain processes and as such they constitute with a human organism one cognitive system. Such systems are for example sensory substitution devices that provide access through one sensory modality to features of the perceived object that are generally experienced through another sensory modality (Farina 2013). One of the most common and effective among them are visual-to-tactile substitution devices that convert images into tactile stimuli (Kaczmarek, Bach-y-Rita 1995).

The most controversial and debated aspect of the phenomenon of cognitive enhancement lies in the fact of the artifact's being directed at the human mind itself, and thus at someone's personality, emotions and agency. How strong can this influence be? Can it affect one's sense of autonomy, and if so, can the agent still be considered autonomous in respect of their actions? How can one prevent agent autonomy being subject to manipulation?[6] Advocates of the extended mind thesis seek to formulate conditions for the autonomy-securing use of cognitive enhancements. The most important of these is appropriate integration of the artifact with natural human cognitive abilities, so that together they form one cognitive system. As long as the mind was reduced to the Cartesian thinking substance, and the content of mental states was available only to the subject, the threat of thought manipulation was only a theoretical speculation. Yet, technological development, that may in the near future lead to an avalanche of artificial cognitive enhancements, have made it a practical possibility that urgently needs to be counteracted. Additionally, the mind has been 'weakened' in its defense against manipulation by the spread of the idea lying behind the extended mind thesis, namely that the mind may extend beyond the skull and even beyond the agent's organism in a way that it involves processes and information states occurring in artifacts. The physical realization base of cognitive processes, dispositional beliefs or perceptual states may therefore extend beyond the safe Cartesian theater into a widely accessible world.

The increasing scope and significance of the impact of such enhancements on the human mind in our own time are such that no theorist genuinely interested in cognition can be left indifferent to this phenomenon. In the present paper I discuss two main issues. First, does cognitive enhancement pose a threat to agent autonomy, and if so, how can we secure the latter? Second, can an extended cognitive system, composed of one human and a cognitive artifact, in its entirety, be considered an autonomous agent? An answer to the first of these questions is bound to be more specific and technical, as it requires the identification of conditions for both autonomous agency and the protection of the agent's autonomy

[5] An example of a system in which feedback occurs directly between brain neural activity and the artifact is the brain-computer interface. It can be initiated using an electroencephalogram, or, more invasively, by attaching electrodes to the cortex of the brain (Vallabhaneni, Wang, He 2005). A project of such a system was presented in 2019 by the Neuralink company and it was intended to provide cognitive enhancement of unimaginable power by directly connecting the human brain with artificial intelligence. Namely, the connection consists in installing sensors in the brain in the form of thin threads that read neuronal activity and transmit the signal to the implant placed behind the ear. The implant, in turn, decodes this signal and sends it to the computer running the appropriate program. As a consequence, it would be possible to send commands to artificial intelligence and receive information from it directly with just thought (Jawad 2021).

[6] I have undertaken these issues in the contexts of epistemic agency in the article: Tomczyk 2021.

from any detrimental influence eventually exerted by such artifacts. I argue that in seeking to determine what autonomous agency implies, one cannot neglect its phenomenal aspect: namely, the feeling of being an agent, and what this consists in. This experience is essential for an agent to act freely and responsively. I devote the first two Sections below to exploring different views on the conditions for autonomous agency, while in the third I consider the impact that artificial cognitive aids have on agent autonomy, and how this might be prevented from taking on a negative character. The issue of what it means for extended cognitive agency to be attributed to an entire cognitive system composed of both a human being and an artifact will then be addressed in the last part of the paper. I present the arguments for and against such an extension, concluding that the latter are more convincing. Extended cognitive processes could in some cases partly constitute human mental states responsible for autonomous agency, yet it is the human being who ultimately makes the decision, even though it may have been strongly influenced by external and unconscious factors. Consequently, only the human part of the extended cognitive system takes the credit or blame for a given action, and it is hard to imagine that future technological solutions will change anything in this matter. What I will try to show is that the thesis of extended autonomous agency, in the context of one-person extended cognitive systems, is not theoretically well-grounded and in practical terms it is empty. As long as intentionality, and phenomenal and access consciousness are attributed only to the human component of an extended cognitive system, autonomous agency, which is itself essentially tied to them, should also be treated this way.

2 Personal Autonomy, its Conditions, and its Role in Agency

While philosophers agree that free action can only be undertaken by an autonomous agent, there is no such consensus as regards what such autonomy amounts to, or what its conditions might be. At the most general level, autonomy implies self-government and the power to initiate action. The autonomous agent has authority over their decisions to act, such that they can be regarded as their own. Undoubtedly, however, free decisions are affected by external influences that are not themselves subject to the agent's authority and over which they have no control. The crucial question that all accounts of autonomy try to answer concerns the extent to which the agent's decisions with regard to acting can be affected by these forces in such a way that the action in question, and responsibility for it, remain attributable to their own agency. In other words, what is the criterion for distinguishing autonomy-undermining influences on an agent's mental states from those whose effects are harmless in this respect. It seems that philosophers are unable to agree about the precise nature of the threat posed in this regard.

The discussion on autonomous agency should be distinguished from the one that concerns the definition of action as such, namely the question under what conditions a being performs any action at all. According to the leading naturalistic conception of agency—Causal Theory of Action (CTA), acting consist in the performance of intentional actions, and the question of the conditions for their autonomy is a separate and further issue.[7] The theory states "that behavior counts as action if and only if it is caused in the right kind of way by mental antecedents which constitute the agent's own reasons for the action (Bishop

[7] Proponents of Causal Theory of Action include, among others: Davidson (1971), Goldman (1970), Brand (1984), Bishop (1989), Enç (2003).

1997, 251). The proponents of this conception dispute, among other things, whether these mental antecedents consist of relevant desire-belief pairs (Goldman 1970; Davidson 1971), or relevant intentions (Brand 1984; Bishop 1989; Enç 2003). All of them, however have to face to the most serious problem for this theory—the problem of deviant causal chains, that occurs when the causal link between person's mental states and their rational behavior obtains, yet intuitively it is not sufficient for action (Bishop 1997). To solve this problem, proponents of CTA analyse various cases of deviant causal chains, including those that involve intermediate intentional actions performed by a second agent. I will refer to this problem shortly at the end of this paragraph, for the solution proposed by one of the CTA representatives—Myles Brand, can be used to indicate situations in which the workings of the artifact undermines not only human autonomy, but their agency itself. What I would like to make clear however is that in this article I refer only to unintentional artifacts, hence the deviant situations that I consider are not the cases of *heteromesial* or *prosthetics* deviance that include a second agent (Enc 2003; Peacocke 1979; Bishop 1989). The cooperation of human and intentional artifact is another issue that poses a serious challenge not only for philosophers of action but above all to ethicists, yet it is a discussion for another article.

Analyses of the conditions for intentional agency conducted by CTA supporters are not without significance for the discussion on the conditions of autonomous agency, yet the difference between the subject matter of these two philosophical endeavors should not be underestimated.[8] Hence, to be clear, my concern in this article is with two issues related to autonomous agency. Firstly, whether artificial cognitive enhancements constitute a threat to human autonomy and if yes, how to prevent it. Secondly, whether the entire human-artifact system could be recognized as an autonomous agent. For this reason, I bring up the discussion below regarding autonomous action, not action per se.

In the literature that explores the conditions for autonomous action, one can distinguish two main lines of thought: one internalist, the other externalist. The most prominent internalist conception is characterized as 'coherentist', and states that an agent's action is autonomous if and only if their motivation to act coheres with their higher-order attitudes which represent their point of view on the action (Dworkin 1988; Frankfurt 1971; Bratman 1979). An important feature here is that both the origin and the content of these mental states are irrelevant in this respect. The agent need not care about the belief-forming process, or about the relationship of such attitudes to reality. All they need to do is occupy a point of view from which they control and endorse their own motives, intentions and beliefs leading to action. Only then can the agent be said to govern their actions, in that they cannot occur without their permission or consent.

Within the externalist approach, two main currents can be distinguished: accounts that focus on *responsiveness-to-reasons*, and those which appeal to *responsiveness-to-reasoning*. According to the first, an agent is only autonomous if their mental states responsible for a given action are responsive to reasons for behaving in the way they do.[9] An agent, in other words, must understand why they have reasons to act so. A *responsiveness-to-reasoning* account, meanwhile, will stress the importance of the very reasoning process itself that is such as to result in the mental states that initiate a course of action. Hence, the agent can be said to govern their actions, on condition that they evaluate their motives in relation to other attitudes they possess and adjust those motives to these evaluations (Christman

[8] I am grateful to an anonymous reviewer for bringing this to my attention.

[9] Among the representatives of this account of autonomous action are Fischer and Ravizza (1993) and Nelkin (2007).

1991). What differs this account from *responsiveness-to-reasons* is an observation that it is not enough for an autonomous agent to be aware of the reasons that guide their behavior. Responsiveness to one's own reasoning prevents the agent from acting blindly on the basis of possessed reasons, which can be imposed on them, without calling their attitudes into question. *Responsiveness-to-reasoning* account allows therefore to exclude actions undertaken as a result of indoctrination from the class of those counting as autonomous, as the reasoning process in such cases will have been so heavily manipulated that any resultant action could not possibly be recognized as agent-governed. Indeed, such a conception of autonomous agency has a much more internalist character than the *responsiveness-to-reasons* account. Specifically, internalists focus on the relations between attitudes possessed by an agent, their ability to draw inferences from them and rationally reflect on them (Frankfurt 1971). The relations between the agent's mental states and external reality, namely the processes that formed these attitudes, are less important for the evaluation as long as the action they caused is autonomous. In this respect, *responsiveness-to-reasoning* account resembles that of the coherentists. What differentiates this account from a coherentist one is the thesis that an agent can be mistaken about their own reasoning processes, and consequently about the authority they have over their own actions (Buss and Westlund 2018). The addict, for example, is so devoted to the act of taking drugs that whatever reasoning process they undertake, its conclusions cannot be attributed to their agency: we would hardly wish to consider this a case of genuine (i.e. rational) reasoning on their part, given that it is governed by external forces.[10] Hence, to be autonomous, an agent must be able to reject the process of reasoning for reasons they possess, and this an externalist condition absent from coherentists accounts.

Still, what is it for an agent to *have* the power that initiates action? All of the mental abilities mentioned above—a reflective point of view consisting in higher-order attitudes towards motivational states, responsiveness to reasons, and the ability to engage in the appropriate sort of reasoning process—could be influenced by an external force to an extent that would undermine the agent's autonomy. It is intuitively clear that to count as governing their action, the agent's reasons and motives such as serve to initiate it cannot themselves be determined by events over which they have no control. This intuition is developed by *incompatibilists*, who state that if an action could be fully explained as the effect of causal powers independent of the agent, it would not be autonomous, as the agent would not have authority over it. So even if they are responsive to reasons, and even if their motivational states are the outcome of appropriate reasoning, the action will not be their own unless they fully control the external factors that influence their attitudes. Consequently, to secure agent autonomy, the first cause of an action should be the agent himself. This does not mean that their decisions cannot be influenced or motivated by antecedent events: it only means that they cannot be determined by them—in other words, the power of agency cannot be reduced to the power of external forces (Chisholm 1976; Clarke and Reed 2015; O'Connor 2009). Hence, the conditions that give rise to an agent's autonomous

[10] The very promising proposal of fitting reasons-responsive approach and mesh theory (which could be classified as coherentist) into a comprehensive theory of agency was presented by Michael McKenna and Chad Van Schoelandt. According to the authors, their proposal can be deployed to solve the difficulties that both mesh and reason-responsive theories face (including a willing addict). The free action in defied by them as follows: Hybrid-Mesh-RR: "A person acts freely in the strongest sense necessary for moral responsibility only if (1) she possesses the ability to act from a suitably integrated and harmoniously functioning psychic mesh; (2) in acting as she does she is appropriately reasons-responsive; and (3) the reasons-responsive resources from which she acts must be accessible to consideration from within the framework of the agent's mesh" (McKenna, Van Schoelandt 2016, 56).

actions, and which are beyond their control, will not be sufficient to produce them: there must be some sort of additional power that comes from the agent theirself. Moreover, the cause of autonomous action will not be itself some mental state or other event arising within the agent: rather, it must be the agent theirself as an enduring substance. Without this, according to incompatibilists, autonomous agency is an illusion, in that freedom and determinism are simply incompatible. Yet the lack of clarity surrounding attempts to make sense of this special sort of agent causation means that this sort of account has rarely been proposed as a conception of autonomous agency.[11]

Despite the differences, there is considerable agreement amongst the various conceptions proposed where certain features of autonomous action are concerned. At the most general level, autonomy is related to agents' cognitive capacities, such as reasoning, awareness and evaluation of the motives and reasons for a given action. The greater these cognitive capacities are, the wider the scope of the agent's autonomy, where this then opens the door to a discussion of the impact cognitive enhancement has on agency.

Before focusing on this, however, it is worth recalling two positions on autonomy which stress the importance of external factors that determine mental states responsible for agent's autonomy. Such accounts, much more then internalist ones, provide a good starting point for the discussion on the extended agency. First, represented by John Christman, focuses on the agent's acceptance of a process of desire and intention formation rather than the agent's awareness and evaluation of the mental states they possess. Christman draws attention to the very process of acquiring beliefs and desires—which, as I show below, can be significantly influenced by artifacts. The condition for autonomy that this author proposes can serve as a criterion that will be indicative of the situations in which this impact is detrimental to agential self-government (Christman 1991). On the coherentist view mentioned above, the agent is autonomous if they reflect critically on their intention to act and, at the level of their own higher-order attitudes, approves of their own entertaining of such an intention. These processes of identifying with an intention cannot be manipulated or constrained if the resulting action is to be considered truly the agent's own. Christman, however, points out that coherentists allow for a situation in which the agent is autonomous even though they are unaware of the process of forming the relevant intention or desire, or even if this process is totally artificial and external to their cognitive character (because, for example, it has been imported using an implant into their brain by a mischievous scientist).[12] According to him, what is crucial in assessing their autonomy as an agent is not their evaluation of the desire they possess, but their evaluation of the process of forming the latter, and their ability to resist it given the chance. There are many factors that could undercut an agent's capacity for proper evaluation and resistance. Hence, what is most important for the agent is that they be self-aware as regards any changes to their cognitive character and their origins. Obviously, to be autonomous, an agent must satisfy other conditions as well, such as consistency across the attitudes and values that guide them in

[11] Attempts have been made to reconcile agent autonomy with determinism by compatibilists. Ned Markosian, for example, claims that it is possible for an agent to be autonomous and responsible for their action, even if some factor beyond their control causes it. He calls this situation 'double causation', and it takes place when two independent events cause a third one. In the case of an autonomous action, it is enough if it is caused partly by an agent and partly by an external cause (Markosian 1999).

[12] Another serious problem facing this position is that it generates an infinite regress. Coherentists require that the higher-order attitudes of identification with a given desire be autonomous themselves, so there must be a further level of higher-order attitudes where one's identification with this state of identification is formed, and so on (Christman 1991). Frankfurt (1987) offers a response to the infinite-regress problem, but according to Christman his proposed solution is not a promising one.

the direction of the activity in question, or not being engaged in self-deception (Christman 1991). This second requirement assumes a capacity for self-government, meaning that the agent is aware of their beliefs and desires and of the process that has formed them. Consequently, people with severe psychopathologies, such as delusions, paranoia and other neuroses, are excluded from being autonomous agents, as they are not able to evaluate the origins or consistency of the attitudes that move them to act. Moreover, the desire-forming process is beyond their control, so they cannot resist it even if they want to.

The second position on autonomy, which emphasizes, even more than Christman's, the influence of external factors on agent's intentions and decisions to act, is relationism. Supporters of this position indicate that every individual is deeply socially constituted. This means that the values and desires that guide their actions are defined in terms of interpersonal relations and mutual dependencies. As Mason Cash explains: "[they] are grounded in shared, intersubjective norms of the social and linguistic practice of ascribing intentional states to one another as reasons for actions" (Cash 2010, 648). One's sense of autonomy is thus decentralized and socially constituted by external relations with others and by complex social determinants, such as race, ethnicity, gender, and class. To be autonomous it is not enough for an agent to reflect on their own, isolated mental states, they need to be aware that all of their values and desires, everything that motivates their choices is constituted by social factors. Personal autonomy is hence a property of human interactions that comprise individual conditions of agency. In a word, a person is autonomous only when their position in these complex interactions reflects the authentic values and standards of the free person (Christman 2004).

I recall relationism towards agency for it parallels to the extended agency thesis, which focuses on the way in which various environmental factors determine mental states responsible for autonomous action, especially the feeling of agency which is one of the most essential among them. Although I focus in this paper only on one type of these factors—artificial cognitive enhancements, relationism seems to provide support for advocates of extending agency beyond a human being to encompass the entirety of the extended system. In the last part of the paper, In will present arguments that such an extension is too far-reaching, and even assuming relationism, autonomous agency is specifically human as far as extended cognitive systems are concerned.

Theorists who define and analyse the conditions for autonomous agency in the context of artificial enhancements, focus mostly on the cases when the process that forms the intention to act is manipulated artificially, with or without the agent's awareness and permission (Carter 2021; Bublitz, Merkel 2009; Sandberg, Bostrom 2006; Fisher 2000). However, the influence of an artifact can be detrimental not only to the autonomy of action, but also to the very agency of the person in question. This is indicated by supporters of the Causal Theory of Action, struggling with the problem of deviant causal chains. Recalling, the deviant situation occurs when a relevant mental states cause the relevant event in such a way that clearly and intuitively, this event is not an action at all. Philosophers of action presented various scenarios of such a deviance.[13] Yet in the context of this article it is worth to bring up real situations that were studied by the researchers working in the MAIA project (Mental Augmentation through determination of Intended Action) (Vanacker et al. 2007). The situations concern brain-computer interfacing (BCI) that aims at directly capturing

[13] One of the most famous scenario was presented by Donald Davidson. It involves a climber who intends to rid himself of the danger of holding another man on the rope by loosening his grip. This intention makes him so nervous that as a consequence of this emotion he loses his hold on the rope. Although he had an intention to do it, intuitively it is not his intentional action (Davidson 1973).

brain activity in order to enable a user to drive a wheelchair without using peripheral neural or motor systems. In these cases it is still the person's intention that is the cause of the wheelchair's move. There occur however deviant situations caused by the errors in the interaction between a human and an intelligent device that include mismatches between the user's intentions and the activity of the device. This could lead to the mistakes in the user's sense of agency which prompt the researchers to recognize such BCI as shared control systems. Moreover, errors in the human-device interaction could result in deviant causal chain. Imagine that user's intention to move the wheelchair causes the workings of the device, but as a consequence of an error the device takes over control (the control of the behavior switches implicitly from user to intelligent device) and causes the wheelchair's move which happens to be in accordance with the user's intention. The intuition that moving a wheelchair is not the user's action in this case can be explained by the solution to the problem of deviant causal chains proposed by Myles Brand. He argues, that action should be proximately caused by a mental event of intending, so that there is no causal space for intervening wayward events (Brand 1989). Hence, in every case of human-artifact interaction, when the proximate cause of human behavior is the workings of an artifact and not human intending, there is no action at all and human is not an agent. Cognitive enhancement could thus be detrimental not only to human epistemic autonomy, but also to their intentional agency as such.

When an artifact replaces human mental states as a proximate cause of a given behavior, human surely cannot be considered an agent. What happens though, if an artifact strongly influences or even constitutes such mental states? Since relevant mental event, for example intending (albeit artificially enhanced or created), is now a proximate cause of a behavior, human being can be considered an agent, at least due to the Brand's conception. Yet, it is not obvious that the agency that they exhibit is autonomous. The crucial question to answer, when assessing threats to the agent's autonomy, is to what extent these influences affect those of their reflective abilities, such as minimal rationality and self-awareness, that could protect them from self-deception. What is of overriding importance for the agent to act autonomously when enhanced by an artifact is to be aware of the origins of their beliefs, desires and intentions, that cause these actions, and to have control over them. As I show in the third section below, cognitive artifacts can enhance or disrupt such abilities and thus also the agent's autonomy itself. Even more, in conjunction with a human being they sometimes, under appropriate conditions, constitute a single extended cognitive system. The influence that such an artifact may exert on human mental states leading to a given action can, in such cases, be so essential as to raise the question of whether such an extended system should be treated as an extended agent. So, what are the grounds for attributing autonomous agency exclusively to human beings? My response will be to argue that the main reason for so doing is the ability to *feel* agency, which can only be attributed to the human part of such an extended system. Before explaining why this is so crucial, though, I shall briefly present the main issues that arise, and that have been explored, in connection with the phenomenology of agency.

3 How Does it Feel to be an Agent, and why is this Important for Autonomous Agency?

The sense of agency is one of the most common experiences in our lives. We often experience our actions as purposive, and an intention to act as being our own. We feel the authorship and the effort associated with a given activity. There are various positions that seek to explain what these experiences involve: i.e. what, exactly, it means that an action is felt to be done on purpose. Those who endorse the *mental causation* thesis state that experiencing oneself as the author of a given movement involves an experience of a mental state—namely, one's intention—as causing it (Wegner 2002; Hohwy 2004). Others, adopting the *agent causation* thesis, argue that it is the feeling one has of oneself as the source that is essential for the experience of agency (Chisholm 1976; Taylor 1966, Horgan, Tienson and Graham 2003). The reasoning behind the latter is that there are cases of volitional disorders where an addict, or a person located on the obsessive–compulsive spectrum, experiences their actions as caused not by theirself but by desires they have that are beyond their control. Yet the question still remains, of what this feeling of being a source consists in. Should it be understood in causal terms? Those who adopt the *agent causation* thesis argue that this is exactly the way an agent experiences their own activity: namely, that they feel that their decisions are caused by theirself as the ultimate source of their action, in light of the reasons they possess (O'Connor 2009). An agent, in other words, experiences the feeling of *mineness*, that is the feeling that the activity is intended, initiated and controlled by their own self. Importantly, the causal role of the self is not understood, by the proponents of this account, as reducible to the causal role of the self's mental states. Acting agent experiences theirself as a substance, not as a bundle of introspectively accessible mental states. The feeling of initiative and control excludes the possibility that the behavior is causally determined by any events, whether mental or physical that realize the former (Bayne 2008; Nida-Rümelin 2018). Hence, the phenomenology of agency seems to be in tension with the influential causal-state theory, since it reveals that it is the agent theirself that is the cause of an action, and not their mental states.[14] The feeling of authorship supports the intuition that the agent's decisions are not causally determined, but free—something which, as many philosophers point out, is required for moral responsibility (Kant 1785/1996, Taylor 1966; Chisholm 1995). This intuition underlines incompatibilist and libertarian accounts on autonomous action according to which free action is inconsistent with being caused by

[14] An interesting solution to this problem is proposed by John Bishop. He incorporates the idea of agent causation into Causal Theory of Action. Specifically, he combines the *mental causation* thesis with volitionism whose proponents claim that what is essential for significantly free action is the agent's own exercise of certain mental capacities, such as the capacity to form the intention to satisfy a particular desire. Bishop argues that such mental exercises of control constitute mental actions that belong essentially to the causal history of significantly free actions. He appeals to higher-order intentions, which belong to the agent as a practical reasoner. In other words, an agent has the general constitutive practical intention, not derived from more fundamental desires and values, to settle the conflict between their desires. This setting the priorities is not done for a reason, so it is a kind of a mental action which is irreducible. This is the way to bring an ontologically irreducible mental action, which is necessary for significant free action, within Causal Theory of Action. Bishop points however that to act autonomously and freely an agent does not have to exercise intentional control over the formation of these high-order intentions, the capacity to form them is a part of an agent's nature and could be realized beyond their consciousness. It does not contradict an agent's feeling that the operation of choosing between alternative desires constitutes their own mental action. Practical intelligence and genuine freedom of action emerges from the functional concatenation of basic and automatic mental and bodily actions (Bishop 1997).

agent's mental or physical properties.[15] Whether it is at all possible to experience oneself as such an unmoved mover, undetermined by one's mental states, is another problem calling for careful analysis that lies beyond the scope of this paper (Bayne, Neil 2006). Here, it will be enough to assume that a feeling of authorship is essential to the experience of agency, and that it involves an experience of oneself as the cause of an action.[16] This is closely related to another important component of the experience of agency: namely, the sense of effort. To undertake an action one has to invest energy and will power, and this involves a sense of oneself as the source of that force that brings it about (Bayne 2008).

The crucial question here to ask is whether the experience of agency is necessary for autonomous action, namely how it contributes to the agent's cognitive economy and whether this contribution is essential. Should systems that do not feel the agentive control over their activities (e.g. artifacts) be excluded from the group of autonomous agents? Should genuine agency be reduced to the human part of an extended cognitive system, the part that is able to feel it? Tim Bayne indicates two putative functions of agential experience. Firstly, it seems that the feeling of authorship and control is necessary for making a free decision between alternative actions. High-level plans and willed intentions cannot be formed by a system which does not have an experience of creating them, and keeping tack on executing them. Secondly, agentive experience enables an agent to reflect on their own actions, namely what they exactly desire and intend, whether they are rational in their choices and whether they are successful in their execution. Hence, a function of agentive experience could be to provide knowledge of one's own agency (Bayne 2008). If the agent did not feel the control over their actions, they would not have beliefs that they are in control, and without them, their autonomy would be questioned and that would lead, for example, to problems with assigning them responsibility for the action taken. In other words, an agent understands what it is to be active, because they experience it, if those experiences are illusionary, they are not autonomous agents (Nida-Rümelin 2018).[17]

What interest me most in considering the phenomenal aspect of agency, is the question whether the experience of being an agent could be constituted by cognitive artifacts and, if so, whether one should conclude that the entire extended system is an autonomous agent. To try to answer this, I will now look more closely at how cognitive artifacts could affect the experience of being one and the same person over time. The feeling of acting

[15] An interesting proposal of compromise between libertarianism and compatibilism is presented by Martin Nida-Rümelin. Namely, he introduces the distinction between causal and metaphysical determination, and argues that free action may be metaphysically determined and yet not causally determined by previous events. It means that for a given freely acting person there could be no metaphysically possible counterfactual situation in which the same relevant preconditions are fulfilled, but the person acts otherwise. At the same time, action of this person is not determined by microphysical properties. According to Nida-Rümelin, there is a specific kind of causation, as far as free human action is concerned, which relates not events as causes and effects, but persons (agents) to events. Hence, this account integrates the main idea of the compatibilist theory (that the free action is compatible with determination in the sense of there being no metaphysically possible alternative to the way the agent acts in a given case) with the true of the libertarian and incompatibilist theory (that free action is incompatible with microphysical determination) (Nida-Rümelin 2018).

[16] This feeling is accurately described by Terry Horgan through the example of clenching one's fist: "You experience your arm, hand, and fingers as being moved by *you yourself*—rather than as experiencing their motion either as fortuitously moving just as you want them to move, or passively experiencing them as being caused by your own mental states. You experience the bodily motion as generated by *yourself*" (Horgan 2007, 187).

[17] The cognitive tasks related with the sense of the self are analyse also by Jacob Hohwy. He argues, that the experience of the self plays an important role in agency and bodily movement in perception and in planning and attention (Hohwy 2007).

is strongly related to the experience of being a self, of having a personal identity. Most philosophers agree that memory plays the central role in constituting this experience. Proponents of psychological-continuity views argue that it is crucial both for maintaining a person's psychological continuity over time and for shaping the experience of who they are (Locke 1689/1997; Garrett 1998; Parfit 1971; Shoemaker 1970). Specifically, a person experiences theirself as the same agent through the passage of time if they remember an experience they had in the past as their own. Empirical studies conducted by Shaun Nichols and Michael Bruno show that it is fairly common intuition (Nichols, Bruno 2010).[18] If memory and other psychological facts that determine personal identity could be constituted by external factors, then these would be constitutive of the agent's self and have to be recognized as a part of the latter. A frequently mentioned example of this situation is encountered in the form of Otto and his notebook (Clark, Chalmers 1998). As a reminder, Otto is a person suffering from Alzheimer's disease, who uses his notebook as a substitute for biological memory. The notebook is essential to every action he undertakes, so no matter what he is doing, he carries it with him and constantly updates the information it contains. Andy Clark and David Chalmers, who brought Otto into fictional existence, assume that information in his notebook plays a very similar functional role to information stored in the biological memory system of a healthy person. It is easily accessible, trustworthy, and has been endorsed at some point in time. The moral of this thought experiment is that it is the function of the artifact, and how its integration into the agent's cognitive system, that matters, not the fact that it is external to the organism. Thus, this extended system of Otto and his notebook is seen by Clark and Chalmers as the agent of the described action.

The information in Otto's notebook is crucial for his experience of personal identity. Without it he would have no access to his beliefs, desires or the other mental states that determine who he is. He would thus be unable to identify himself with the person he was in the past, and would not experience his own continuity over time (Heersmink 2017). Moreover, the notebook seems to be responsible for Otto's experience of possessing agency in respect of the actions he has undertaken in the past. If this is right, shouldn't it be considered a part of an extended agent that includes both, Otto and the notebook? In my view, the consequences of such a contention are so far-reaching that, in practice, agency is never recognized as extended in this kind of way, and indeed should not be. In the last part of this paper, I will present both practical and theoretical arguments for why that is so.

The increasing importance of cognitive artifacts in our lives, and the popularity of the extended mind thesis, have given rise to a new problem that concerns the locus of agency and personhood within extended systems. Our memory is supplemented and supported by a variety of cognitive artifacts designed to externalize cognitive work so that a part of it, or even all of it, is performed by external representations and other structures in our environment. These includes computer systems, calculators, maps, diagrams, models, timetables and many other cognitive aids that help us perform such cognitive tasks as remembering, planning, learning or calculating. Proponents of different versions of the extended mind thesis impose different conditions on the artifacts that are supposed to be parts of extended

[18] Not everyone agrees with psychological-continuity views based on memory criterion. David Behan presents critical arguments directed by several philosophers against this account of personal identity (Behan 1979). The critics point, for example, on situations where an old person does not remember some events from their youth. According to the memory criterion, the old person is not identical with a young one although they are one and the same human being (impossibility result). Anticriterialists also object psychological-continuity account arguing that psychological continuity is not always required for a person to persist (Merricks 1998). Advocates of brute-physical view point, on the other hand, that person identity is based not on the agent's memory of past experiences but rather on identity of their body (Ayer 1936).

cognitive systems. They all agree, however, that deep integration with a cognitive artifact strongly shapes the agent's cognitive abilities, which themselves are essential elements of their experience of who they are and what they are capable of doing. This is particularly the case for people suffering from a decline in memory-related capacities (Clowes 2015; Sandberg, Bostrom 2006; Rhodes, Starner 1996). Cognitive artifacts may to some extent substitute for dysfunctional aspects of someone's biological memory (as in Otto's case), and delay its further disintegration. At the same time, though, their essential role in shaping the agent's personal identity is also vividly manifested where healthy agents are concerned. This is the reason why some extended-mind theorists are willing to treat these assistive technologies as constitutive elements of an agent's personhood. Richard Heersmink, for example, argues that "[p]ersonal identity can thus neither be reduced to psychological structures instantiated by the brain nor to biological structures instantiated by our biological organism. […] [W]e should broaden our concepts of the self so as to include social and artifactual structures, focus on external memory systems in the (empirical) study of personal identity, and not interfere with people's distributed minds and selves" (Heersmink 2017, 3149). If this is right, should we not also broaden our conception of autonomous agents? I think that if one takes Heersmink's position as a basis, the answer will be positive, yet there are strong reasons not to do so. Before focusing on these, however, I will devote the next section to showing how, and under what conditions, cognitive artifacts can have a constitutive impact on those mental structures of agents responsible for their autonomy. The presence of such an influence in many cognitive situations supports the extended agent thesis, which I nevertheless seek to challenge.[19]

4 The Influence of Artificial Cognitive Enhancements on Personal Autonomy: Hopes and Challenges

Generally, cognitive artifacts are designed to enhance the human mind, so why can't they have a positive impact on agent autonomy itself?[20] For those who associate autonomy with the ability to reason, its enhancement is an everyday phenomenon (Schaefer, Kahane, Savulescu 2014). Freedom and self-determination, which are at the core of most conceptions of autonomy, are shaped by deliberative capacities that lie within the scope of potential cognitive enhancement. Autonomous agents should be able to evaluate different

[19] The possibility of an extended experience of agency is also supported by the idea of extended consciousness advocated by some enactivists (Noe 2004; O'Regan 2011) and advocates of predictive processing theory (Kirchhof, Kiverstein 2019). They emphasize the strong (constitutive) dependence of an agent's conscious experience on environmental factors, especially on an agent's motor activity in a specific biological and cultural environment. Yet, analyses conducted by those researchers concern mainly perceptual consciousness, and especially visual conscious experience. Applying their arguments to explain the sense of agency poses a more serious challenge. I do not claim that consciousness, be it visual or related to agency, does not depend on social and environmental factors. It obviously does. The question is however, what this dependence means. According to advocates of extended consciousness, the dependence is not only causal, but constitutive. The realizers of perceptual experience can extend beyond the brain to include bodily and worldly elements. While causal dependence is evident, the thesis that conscious experience is constituted by external factors along with internal mental states is much more radical claim and is being challenged as not well grounded (Prinz 2009; Chalmers 2019).

[20] An insightful discussion with reductive, neuroscientific accounts on human agency is presented in the work of Andrea Lavazza and Mario De Caro (Lavazza, De Caro 2010). The authors present objections to the optimistic view that complete explanatory reduction of the human mind to the electro-chemical functioning of the brain will bring about positive consequences at the social, cultural and political level.

options, infer and weigh up the consequences of alternative courses of action, assess potential goals and methods of achieving them, solve problems on their own, and so on. The most common examples of autonomy violation mentioned in the literature—psychological manipulation, deception and lack of self-awareness—affect the agent's ability to reason and deliberate properly. Every method of preventing these by improving the agent's logical competence, pattern recognition, linguistic abilities, memory, etc., serves to enhance their agential autonomy, in that one needs to possess all these cognitive capacities if one is to choose freely between various options and undertake one's chosen actions intentionally. Such capacities enable one to control the accuracy and coherence of one's mental states, identify fallacious arguments, and recognize alternative options with regard to action. Even genetic manipulation, which is among the most controversial means of cognitive enhancement,[21] could be justified and considered valuable, as it promises to improve the child's ability to reason and deliberate, making the latter less susceptible to external violations of its autonomy (Schaefer, Kahane and Savulescu 2014).

Despite the undeniably positive impact of cognitive artifacts on agent autonomy in many situations involving rational evaluation, the threats and downsides of artificial enhancement have tended to be raised more frequently by philosophers of mind, epistemologists and ethicists (Biblitz, Merkel 2009; Sandberg, Bostrom 2006; Carter 2021). In particular, they emphasize the risk of diminished authenticity, social inequity, threats to human nature and dignity, automatization of the decision to act and lack of responsibility. Putting social issues to one side, and focusing on the suppression of individual agent autonomy, the most crucial question to answer concerns the extent to which the agent ought to be aware of the workings of an artifact and its impact on their mental states. On the one hand, to constitute an extended system together with an artifact a human being should deploy it automatically, and unreflectively place trust in what it delivers. Only then can it be considered functionally equivalent to innate human mental resources (Clark, Chalmers 1998). Reminding, the parity principle introduced by Clark and Chalmers implies that to count as functionally equivalent to internal processes, an artifact should not be subjected to the agent's considered attention and evaluation, as internal processes are often not themselves objects of conscious attention. The relation of continuous reciprocal causation, which constitutes an external cognitive system, occurs when the enhancement is easily and directly accessible and applied uncritically, analogous to biological cognitive processes. Unfortunately, this opens the way to every form of manipulation seeking to target the agent's mental states, a problem that has been noticed and deeply analyzed by representatives of the *extended knowledge*. Generally, this approach aims at developing the conditions for the epistemically valuable beliefs, i.e. knowledge, that arise from the operation of an extended cognitive system. In other words, its proponents examine the influence of the extended mind thesis on analyses specifically related to the concept of knowledge. Nevertheless, they are often viewed as adopting a stance in tension with the *extended mind* thesis: they would like to impose internalist conditions on knowledge in order to protect its subject from all possible external influences that could undermine the agent's epistemic autonomy. When referring to the advocates of the *extended knowledge*, I mean the authors who defend this approach from the perspective of virtue epistemology (Pritchard 2010; Carter 2021). Hence, they define knowledge in terms of the cognitive achievement that an agent has attained using their own cognitive faculties and for which they deserve credit (Sosa 1988; Greco 1999; Pritchard 2010). Virtue epistemology is an externalist and reliabilistic theory of knowledge,

[21] One of the philosophers who has objected to biological enhancement on the grounds that it potentially undermines agent autonomy is Jürgen Habermas (2003).

which introduces the concept of cognitive ability to reliabilism.[22] Namely, the belief-forming process resulting in knowledge cannot by luckily truth-conducive and it cannot consist solely of the use of other people's cognitive abilities. Virtue reliabilism stresses the need for an agent to be creditable with having achieved cognitive success in respect of arriving at true beliefs. Not only must a cognitive process be reliable if it is to culminate in knowledge, it also needs to be truly the agent's own—i.e. creditable to *their* agency.

In a situation where a cognitive process resulting in knowledge is artificially enhanced, an artifact should be properly integrated with the agent's cognitive character. It means that the agent must, at some point in their life, consciously incorporate external enhancement into their cognitive abilities by accepting it as reliable (Clark, Chalmers 1998; Pritchard 2010). There is however a tension between this condition and the extended mind thesis, namely it does not favor the functionalist attitude specific to the first-wave supporters of this thesis fighting against bio-prejudices. According to functionalists, the nature of the cognitive process (biological or artificial) is irrelevant to its knowledge-conducive function. Yet, the intuitions extracted by virtue epistemologists by means of many thought experiments indicate the weakness of this position (Carter 2013). Biological and artificially enhanced cognitive processes are not epistemically equivalent. As has already been said, in order to incorporate the manipulation of artificial cognitive enhancement into agent's cognitive character, the agent must consciously and freely decide about it, which they do not have to do in the case of biological processes such as perceptual or rational faculties.[23] The dilemma as regards the extended mind versus extended knowledge theses could be solved by indicating the difference between biological (natural) and extended (enhanced) cognitive processes—something already suggested by Clark and Chalmers in the form of their fourth criterion for extended beliefs, which is past endorsement.[24] That is to say, where internal cognitive processes are concerned, the condition of consciously endorsing them as reliable and making a decision to utilize them need not be met for them to count as knowledge-conducive. This condition only pertains to artificial cognitive enhancements used to improve biological processes. Interaction with a device could become automatic and unreflective over time, but to result in authentic mental states, responsible action and genuine knowledge it must be consciously accepted at the beginning and monitored from time to time for its reliability.

Furthermore, the dilemma of functional parity versus conscious acceptance of cognitive enhancement can be extended from the issue of knowledge to the wider problem of autonomous agency as such. In the context of most views concerning agency, the autonomous agent, just like the subject of knowledge, is treated as being aware of the reasons that determine their actions. This is essential for monitoring their source and rationality, and for reacting in the event that any signs of the unreliability of the relevant process are noticed, or that any kind of external manipulation being performed on their mental states is detected. The real threat to agency arises in situations where the natural cognitive process has been replaced by a completely different mechanism: for example, by an implant placed

[22] Reliabilism itself states that the subject has a justified belief if and only if it is the product of a reliable cognitive process, i.e. one that in most cases leads to true belief (Goldman 1979). Virtue epistemologists point out that this is an insufficient condition for knowledge, and illustrate it with many counterexamples (Greco 1999, Pritchard 2010).

[23] For more detailed discussion of the conditions that must be met for extended knowledge to be consistent with the requirements of epistemic security, see my article: (Tomczyk 2021).

[24] "Fourth, the information in the notebook has been consciously endorsed at some point in the past, and indeed is there as a consequence of this endorsement" (Clark, Chalmers 1998, 17).

in the brain that takes over some of our natural cognitive functions. The agent's autonomy will only have been preserved if they consciously decide to utilize the enhancement, and are aware of the expected results of its application—or, if unfamiliar with them, are at least aware of the risk being taken. Hence, the agent, if they are to be autonomous, cannot be manipulated in a way that is completely beyond their conscious control (Bublitz, Merkel 2009). When this happens, they ceases to be the subject of the actions performed: simply, they cannot be regarded as their achievement, and they cannot be credited with or blamed for them. Adam Carter specifies this condition by pointing out that autonomous mental states must have a compulsion-free history—a requirement that is only satisfied if the agent has not acquired them in a manner that bypasses or preempts their cognitive competences in such a way as to deny them a proper capacity for dispensing with that belief (Carter 2021). Only after this condition has been met can the mental states that motivate actions be said to be truly the agent's own, such that they can be given credit and assigned complete responsibility for the latter. If the relevant mental states are the results of an enhanced cognitive process, then this should be properly integrated with the agent's cognitive character so as to co-constitute a single extended cognitive system. Still, the question remains of whether such an extended system will itself count as an autonomous agent: i.e. whether it can be said to possess its own personal autonomy and agency.

So far, I have mainly referred to the first-wave of arguments for the extended mind thesis, for they focus directly on the one-person, extended cognitive system, which interests me the most. It should be noted, however, that proponents of a second and a third-wave arguments observe that Clark and Chalmers do not take into account the strong influence of social and cultural factors on humans cognitive activity seriously enough (Menary 2010; Gallagher, Crisafi 2009).[25] Cognitive practices shaped by cultural norms and cognitive institutions constitute human's cognitive activities to such an extent that the individual agent seems to dissolve into the surrounding environment that constitutes them, so that it is difficult to define their boundaries. However, the agent is undoubtedly still there, as they make decisions, act and take responsibility for their actions. This is admitted by the proponents of relational autonomy and distributed cognition themselves. Diana Meyers, for example, defines the autonomous agent as one who critically reflects on social forces, which determine their decisions, endorsing some of them and rejecting others. It is an individual who constructs their autonomous self being embedded in social and cultural context (Meyers 2005). Mason Cash on the other hand argues, that an autonomous agent is not an objective fact, but it is a characteristic that community members assigns to a given person, because they satisfy certain conditions for autonomous and rational action set by shared social practice and its norms. Our sense of agency and the experience of being ourselves develop as a result of such assignment (Cash 2010).

Analyses conducted by relationists and by representatives of the 'social' waves of the extended mind thesis cannot by overestimated, for they show how strongly mental states responsible for autonomous agency, and the conscious experience of being an agent are embedded and dependent on social and cultural factors (Farina, Lavazza 2021). Every individual mind is made up of them. Yet, eventually there is an individual person who gets the credit and the blame for a particular action, because the decision to act is their own even if it is shaped by external factors. For this reason, and for those that I present below, I am inclined to adopt individual-centered conception of autonomous agency. I would like to make it clear, that I refer in this article only to a single—person cognitive systems; as far as

[25] There are however representatives of a second-wave arguments, who notice Clark's interest in cultural and social factors, expressed, among others, in Clark 2004 (Sutton 2010; Farina 2021).

group systems are concerned agency and responsibility often cannot be attributed to a single individual. Yet, though extremely interesting, this is a topic I cannot take up here. In the following, I return thus to the analysis concerning the agency of a single-person artifact-enhanced cognitive system.

The extension of the boundaries of the agent beyond the human organism as such is something that carries far-reaching ethical consequences. If the entire extended cognitive system is considered to be a subject of the mental states that determine personal autonomy and identity, then why not assign it the status of full-blooded personhood? Why should a person be confined to an organism? And yet, if an artifact conjoined with a human organism constitutes an extended person, then, on the basis of the right to self-ownership and personal autonomy, it should be protected against assault and violation, just as the biological parts of that system are (Carter, Palermos 2016). The rationale behind this argument is that those material realizers of the agent's mental states and cognitive capacities that constitute their personal identity could include factors from beyond the human body—providing just that they be appropriately integrated with the biological cognitive processes themselves. Such dense feedback loops between the human organism and technology are nowadays constituted when use is made of smartphones, telescopes, hearing aids, smart glasses and watches, and many other cognitive aids. These artifacts become transparent to their users, in other words they become means through which the environment is experienced and acted on (Brey 2000). It is often the case that technology is incorporated into the agent's body schema and it becomes a part of their bodily space (Merleau-Ponty 1962/1945). As a result, it constitutes not only their motor and perceptual abilities but it could also shape agent's personal autonomy and identity. If so, then intentionally inflicting damage on an integrated epistemic artifact should qualify as a case of personal assault. This idea, which is a consequence of both, the extended mind thesis and a widely embraced assumption about what should count as personal assault, seems radical and counterintuitive. Nevertheless, the more artifacts become subtle, intimate and discreet as a result of technological progress, the less controversial the thesis of 'extended assault' appears. Accepting the extended mind thesis has the effect of blurring the strict division between artefacts and biological parts of human body as far as ethical issues are concerned. It means that interventions into the artifact, that is claimed to be a part of human's agency, counts as interventions into their mind (Levy 2007). A suggestive example of such an intimate human-artifact integration is furnished by Neil Harbisson and his 'eyeborg'. Harbisson suffers from achromatopsia, and as a consequence sees things only in monochrome. Thanks to a device implanted in his occipital bone that converts visible colors into sound waves, he is able to distinguish colors by hearing them. He feels the "eyeborg" to be—and treats it as—a part of his body: one that yields new sensory content. So how should the intentional damaging of an external part of such a device be regarded? Should it be viewed as an instance of personal assault or just damage to property? In this case our intuitions are not so clear-cut.

Both enthusiasts regarding artificial cognitive enhancement, and those who are inclined to take a more cautious approach, would agree that in some cases it may partly constitute the mental states responsible for autonomous action and for the feeling of being an agent. In the final Sect. here, though, I shall argue that assuming that this is so does not furnish a sufficient basis for adopting the thesis of the extended agent.

5 Non-Extended Autonomous Agency

Accepting the possibility of extended cognition conducted by an extended system does not mean agreeing on the existence of an extended agent. My argument for treating agency as specifically human, as far as human-artifact cognitive system is concerned, follows from the thesis of the non-extended character of phenomenal consciousness and intentionality and is supported by unacceptable practical consequences to which the adoption of the extended agency thesis leads.

To be an agent, the cognitive system must be able to experience itself as a cause of its own actions. The feeling of possessing beliefs, desires and intentions as agent's own, that is not imposed by anyone and anything from beyond their conscious control, would seem to be the primary condition in play when attributing action and responsibility for it. Intentionality is no less important: standing in an intentional relation to experienced phenomena enables the agent to access the content of their mental states, where these determine the action in question, and to understand their meaning. In what follows below, I shall make reference to Chalmers argument that lend support to the thesis that consciousness, which is necessary for agency, is confined to the human biological organism, as far as a human-artifact extended system is concerned. I will then present theoretical and the practical argument against the extended agency thesis and I will conclude my considerations with a polemic with Farina and Lavazza's extended agency thesis (Farina, Lavazza 2022a, 2022b).

While Clark and Chalmers argue in favor of an extended self, they view this as outstripping the boundaries of consciousness. Dispositional beliefs, for example, are not conscious, yet they do help constitute the agent's personal identity (Clark, Chalmers 1998). Phenomenal consciousness correlates, according to Chalmers, with the physical processes that enable this or that given information to be directly available for global control. Only internal brain processes can be regarded as furnishing such correlates, as processes extended via perception or action only provide indirect access, for purposes of global control, to the information they carry. To be precise, such information, in order to be used in that way, must pass through three stages: from object to eye, from eye to visual cortex, and from visual cortex to loci of control. Meanwhile, the internal, neuronal correlates of consciousness only have to travel some portion of the final stage. Assuming that phenomenal consciousness requires information to be directly available for purposes of global control, extended consciousness is therefore impossible (Chalmers 2019).

As I have already noted, Chalmers refers in his argumentation to both, phenomenal and access consciousness, where the latter is construed as access to the content of one's propositional attitudes. Otto's extended belief is dispositional, so it does not require direct access for purposes of global control to the information it carries. Thus, perception and action constitute a boundary for consciousness, although they do not do so when it comes to cognition. If, as I am arguing, agency requires phenomenal consciousness and intentionality, then it is necessarily confined to the human part of any extended cognitive system. At the same time, and *contra* Clark and Chalmers, I am also inclined to reject externalism about the self and this is the theoretical part of the argument against the extended agency thesis with which I would like to close my considerations.

The division of the self into its internal and external parts seems to be metaphysically suspect, and only possible on the grounds of a theory of identity. If we can agree on some kind of non-reductive physicalism, then it is reasonable to think that a self, understood as the personal identity of an agent, is a higher-level, emergent property of a physical system.

As such, it does not occupy any space, in which case talk of its 'internal' or 'external' character makes little sense.[26] Hence, even assuming that the system is extended, such a thing cannot be asserted of its mental properties. The fact that the social, linguistic and physical environment plays a crucial role in shaping how an agent thinks about the world and theirself does not mean that they, as a person, extend to these external factors. An agent is a subject of phenomenal consciousness and not an extended cognitive system. The cognitive processes that they engage in may be extended, yet they theirself, as their subject, are not. Consequently, even if Otto's memory extends to his notebook, he himself does not. He retains a special status in the extended cognitive system: namely, that of a conscious agent with a first-person point of view. This is not to say that physical realizers of cognitive processes cannot be partially extended: they will be so where there is an adequate causal coupling between internal and external physical processes involved in realizing such a cognitive process. Nevertheless, such coupling, and thus such cognitive extendedness, takes place only at the physical and neurological level, not the personal or mental one. To repeat, the subject of mental properties is the intentional self, and not the extended system as such. In fact, the whole discussion about the extended mind is concerned with its physical realizers, not the extended mind itself. A common view of realization is that the realized properties are located where the realizer properties are located, yet this leads to multiple ambiguities, absurd scenarios, and troublesome practical and theoretical consequences.

Treating mental states as extended leads to the problem of their subject having to be thought of as extended too, yet agreeing on this, one surely also has to face up to its counterintuitive consequences as regards moral and epistemic responsibility and creditability. An example of just such a troublesome situation is Otto committing a crime (Swallow 2013). According to the original scenario, he remembers this fact only by virtue of having access to his notebook. Without it he is not the same person, for his diachronic identity—i.e. his life narrative—is constituted by the information in that notebook. In other words, Otto's self includes his notebook. If this is right, the person who commits the crime is Otto-plus-his-notebook, since the notebook is essential for Otto to make the decision to commit the crime and to remember the experience of committing it. The conclusion implied by this scenario is hard to accept: one would only be entitled to punish Otto-plus-his-notebook, since Otto on his own counts as a different person. The entire extended system would thus have to be held responsible and, in consequence, be granted rights and obligations on a par with the individual human being. This, then, is the implication of the extended mind thesis: if memory is extended, and if personal identity is partly constituted by memory, then the self must also be extended. However, when it comes to punishing such an extended individual for a crime, the question will always then arise of whether we are in fact dealing with the very same person who committed it. The most crucial question, though, is that of who we would actually want to reward or punish: doing so only seems to make sense when dealing with an agent who is aware of her actions and their consequences. Even if Otto needs to have access to his notebook to be blamed for a crime, it is him, not Otto-plus-his-notebook, who is to be punished. His agency is confined to him, for within this extended system only he has an experience of being an agent committing a crime. At this point it is worth noticing, that external mental states with affective and motivational elements, such as wants, desires, hopes and longings, are much more difficult to imagine than those referring to such facts as the address of the museum (Sterelny 2010). What would Otto's daily activities look

[26] This line of thought about the self, consciousness and agency has been further pursued by Lynne Rudder Baker (Baker 2009).

like if all his preferences and emotions, such as his favorite music, food, sexual preferences, were located in the notebook? It seems that these mental states are internal, for they are essentially related to conscious experience of one's own personality. There are other thought experiments in the literature that point to practical problems with attributing credit and blame to the entirety of an extended system. Yet, it should be stressed, that they cannot constitute the argument against the extended agency thesis, they could only be a motivation to try to argue against such an extension. If there were strong theoretical arguments against limiting agency to human being as it comes to extended human-artifact systems, these practical consequences would have to be accepted and dealt with. Yet, I do not recognize strong reasons towards such a struggle, on the contrary, I find strong reasons against them.

Before concluding, I would like to recall and reflect on the argument in favor of the extended mind thesis which I find very convincing although I do not agree with its implications regarding extended agency. Advocates of the extended mind thesis tend to perceive the usage of artifacts, that are deeply integrated with human organism, as enhancements in the first place, not as a threat to human's autonomy. This is because they view the bounds of the cognitive agent as including both, human organism and cognitive technologies. Extended cognitive systems are simply cognitively more effective than those limited to the organism. Moreover, there is often the case that the artifact becomes transparent for the agent, so that they experience it as a part of their body. This is exactly the case of Neil Harbisson, mentioned in the previous section. Long-term and constant use of the 'eyeborg' has shaped his personality and his experience of being an autonomous agent. A central question is whether such technologies should be treated as parts of a human agent or rather as external elements only causally affecting their mental states. Mirko Farina and Andrea Lavazza offer moral arguments in favor of the former (Farina, Lavazza 2022a, 2022b). In a nutshell, they argue that treating an artifact as external to the agent makes it a potential threat to their autonomy and it could only be dismissed by meeting special conditions, such as consciously incorporating this technology into agent's cognitive character, as virtue epistemologists put it. All the worries about technology as violating an agent's freedom of decision and sense of control arise from such an organism-bound account of an agent. A very different situation occurs when an artifact is considered as part of an agent, for then it does not pose an external threat to their autonomy. Such an extended agent possesses greater cognitive capabilities, makes more rational decisions and is more effective in their implementation. Surely, what Farina and Lavazza point out, for such an extended system to arise an artifact should be consciously incorporated by the human into their cognitive repertoire. After this happens however, it becomes a constitutive part of a human's cognitive character, so it is no longer a threat to their self-control, autonomous agency and responsibility. Moreover such an artifact should be legally protected against any violation just as biological parts of human body are.

What is crucial for Farina and Lavazza's defense of the extended agency thesis is that it does not assume that the individual's identity is constitutionally distinct from the device, but rather their self and personality is co-realized by it (Farina, Lavazza 2022a). Hence, there is no specific part of the system on which technology could intervene without agent's permission, because the whole extended system is perceived as an autonomous agent. This, according to Farina and Lavazza, is morally a more preferable approach to artificial enhancements than the conception that refer to causal relation between human mind and technology. The deeply integrated artifact is coextensive and constitutive of an agent's cognition and their experience of agency. In such cases, when devoid of their device an individual will no longer be the same moral agent than they were before. Their options for

action are drastically limited and they cannot be held fully responsible for the failure of the activity they always perform using the artifact beforehand, for this would be highly unjust (Farina, Lavazza 2022a; 2022b).[27]

I agree with Farina and Lavazza that an individual suddenly devoid of their deeply integrated device is not the same agent as they were using it and they should be legally and morally treated accordingly. I also agree that the extended mind thesis provides a better justification for such an attitude than the internalistic accounts. Yet, I do not consider their argument as a challenge to my thesis of non-extended agency. Most importantly, a human agent has to consciously endorse the enhancement into their cognitive character in order to make it its integral part. If the autonomy of a human being is to be respected, they have freely to decide to use a device, only after that they become a part of an extended cognitive system. This situation, especially if it is long-term and constant, changes dramatically agent's mental states responsible for actions they undertake and I agree with Farina and Lavazza that within the extended cognitive system they become different cognitive and moral agent. Their agency however cannot be transferred partially to the artifact so that the extended system becomes an autonomous agent as a whole. It is still a human being who decides about their actions, is aware of their intentions and acts accordingly.

It is without a doubt that as a result of using cognitive technologies human beings have undergone profound changes as cognitive and moral agents. Not only cognitive processes and mental states are constituted by the workings of deeply integrated and increasingly transparent cognitive enhancements, but also conscious experience of being an autonomous agent and, more generally, of being oneself. More and more we rely on technology in our day-to-day decisions. The crucial question however is who makes these decisions, is it a human being or an extended cognitive system as such. If we assume this latter scenario, we should also accept that it is the extended cognitive system who is an autonomous agent in the cognitive, epistemic and moral aspect. In this article I defend the view that this is not the right way to consider agency within the extended cognitive systems. Only the conscious individual can be autonomous in the sense of making free decisions based on beliefs and desires that are truly their own. The autonomous agent is aware of their value hierarchy, and chooses an action that is or is not in accordance with it, anticipating the consequences of each choice. Agency, then, cannot be extended to artifacts, on the grounds of the experience that constitutes it and which can only be attributed to the human part of an extended system. The autonomous agent is the one who feels the agency, is aware of being the source of their actions, and feels in control of it. Every position regarding autonomous agency invoked in this paper has assumed the consciousness of the agent, along with their reflective first-person perspective. Acceptance of the process of intention formation, self-government, critical reflection, self-awareness as regards changes to the agent's own cognitive character and their sources, the feeling of authorship, and the sense of effort—none of these cognitive processes can be extended to artifacts. For this reason, even if we assume that intentions and decisions to act are partly realized by external factors, they cannot be attributed to the extended system as a whole. From a theoretical standpoint, they can be considered as higher-order properties of the overall system, but the internal part of

[27] The authors refer to the case of a person who use a necklace with a chemical sensor that makes an annoying sound when this person drinks alcohol. This device is supposed to prevent the person from drinking excessively, hence it is considered as their moral resource, for it helps them to peruse morally good behavior. Farina and Lavazza argue, that if the necklace fails and the person gets drunk, we cannot judge them as an individual independent of this device, for they are now a different moral agent from the one they were with the necklace. It is much better from a moral point of view to treat this tool as part of the agent rather than just as external aid to the person's willpower (Farina and Lavazza 2022a, 2022b).

their physical realization, responsible for conscious reflection, is so crucial that we have to acknowledge it as determining the boundary of what counts as autonomous agency. It is Otto who remembers and decides, even if his memory is constituted by the contents of his notebook.

6 Concluding Remarks

What I was trying to show in this article is that the acceptance of the existence of extended cognitive systems conducting extended cognitive processes does not necessarily lead to the adoption of the extended agency thesis. I presented arguments against the second thesis accepting the first. In order to do this, I first brought up a discussion about the conditions of autonomous action, which reveals that the essential properties for such action are specifically human. Especially significant for autonomy is its phenomenal aspect, namely, the feeling of being an agent, and what this consists in. Having considered this issue, I focused on the impact that artificial cognitive aids have on agent autonomy, and how this might be prevented from taking on a negative character. The arguments I have presented are grounded in the discussion about the extended mind thesis and the consequences it leads to when it comes to understanding knowledge, consciousness, self and responsibility. In doing so, I invoked not only the first wave of arguments, represented by Clark and Chalmers, but also the next two waves that root the human mind much more firmly in the social and cultural environment. I concluded, however, that the arguments from representatives of these later waves do not support the thesis of extended agency more strongly than the Clark's original idea. In the last part, I backed up the argument for the non-extended nature of consciousness and intentionality with Chalmers' argumentation and with a critique of the extended self thesis. Next, I pointed on the difficulty to accept practical implications of the extended agency thesis, related mainly to the problem of assigning responsibility for the action taken by the extended agent. I ended with a reflection on Farina and Lavazza's argument regarding the superiority of the extended mind thesis over internalist positions in moral and ethical discussions and I concluded that it does not imply the need to extend autonomous agency to the entire extended system.

Can an extended cognitive system, composed of one human and a cognitive artifact, in its entirety, be considered an autonomous agent? Taking into account everything said above, I answer to this question in the negative. When it comes to a single-person extended cognitive system, agency is specifically human, for such are its essential features, namely intentionality, consciousness and autopoiesis. Perhaps artificial systems will someday reach such complexity that they will exhibit these characteristics, yet if this were to happen, the cooperation of humans with them will constitute more-then-one-subject system and the analyses concerning its agency and mentality in general will resemble those concerning a group system. This interesting and challenging topic requires however separate analysis.

References

Adams, Frederick, and Kenneth Aizawa. 2001. 'The bounds of cognition': *Philosophical Psychology* 14 (1): 43–64. https://doi.org/10.1080/09515080120033571.

Ayer, Jules Alfred. 1936. *Language, truth, and logic*. London: Gollancz.

Baker, Lynne Rudder. 2009. 'Persons and the extended mind thesis': *Zygon* 44 (3): 642–658. https://doi.org/10.1111/zygo.2009.44.issue-3.

Bayne, Tim. 2008. 'The phenomenology of agency': *Philosophy Compass* 1 (3): 182–202. https://doi.org/10.1111/j.1747-9991.2007.00122.x.

Bayne, Tim, and Neil Levy. 2006. The feeling of doing: deconstructing the phenomenology of agency. In *Disorders of volition*, ed. Natalie Sebanz and Wolfgang Prinz, 49–68. Cambridge, MA: MIT Press.

Behan, David. 1979. Locke on persons and personal identity. *Canadian Journal of Philosophy* 9: 53–75.

Bishop, John. 1989. *Natural agency: An essay on the causal theory of action*. Cambridge: Cambridge University Press.

Bishop, John. 1997. Naturalising mental action. In *Contemporary action theory*, vol. 266, ed. Ghita Holmström-Hintikka and Raimo Tuomela. Dordrecht: Springer.

Brand, Myles. 1984. *Intending and acting: Toward a naturalized action theory*. Cambridge, MA: MIT Press.

Brand, Myles. 1989. Proximate causation of action. *Philosophical Perspectives.* 3: 423–442. https://doi.org/10.2307/2214276.

Bratman, Michael. 1979. Practical reasoning and weakness of the will. *Noûs* 13: 131–151. https://doi.org/10.2307/2214395.

Brey, Philip. 2000. Technology and embodiment in Ihde and Merleau-Ponty. In *Metaphysics, epistemology, and technology. Research in philosophy and technology*, ed. C. Mitcham. London: Elsevier/JAI Press.

Bublitz, Jan Christoph, and Reinhard Merkel. 2009. Autonomy and authenticity of enhanced personality traits. *Bioethics* 23 (6): 360–374. https://doi.org/10.1111/j.1467-8519.2009.01725.x.

Buss, Sarah, and Andrea Westlund. 2018. Personal autonomy. In *The stanford encyclopedia of philosophy*, ed. Edward Zalta. Cambridge: Academic Press.

Carter, Adam. 2013. Extended cognition and epistemic luck. *Synthese* 190 (19): 4201–4214.

Carter, Adam. 2021. Epistemic autonomy and externalism. In *Epistemic autonomy*, ed. Kirk Lougheed and Jonathan Matheson. London: Routledge.

Carter, Adam, and Spyridon Orestis Palermos. 2016. Is having your computer compromised a personal assault? The ethics of extended cognition. *Journal of the American Philosophical Association* 2 (4): 542–560. https://doi.org/10.1017/apa.2016.28.

Cash, Mason. 2010. Extended cognition, personal responsibility, and relational autonomy. *Phenomenology and the Cognitive Sciences* 9: 645–671. https://doi.org/10.1007/s11097-010-9177-8.

Chalmers, David. 2019. Extended cognition and extended consciousness. In *Andy clark and his critics*, ed. Matteo Colombo, Elizabeth Irvine, and Mog Stapleton, 9–20. Wiley-Blackwell.

Chisholm, Roderick. 1976. The agent as cause. In *Action theory*, ed. Mules Brand and Douglas Walton, 199–211. Dordrecht: D. Reidel Publishing Co.

Chisholm, Roderick. 1995. Agents, causes, and events: The problem of free will. In *Agents, causes and events: Essays on indeterminism and free will*, ed. Timothy O'Connor, 95–100. New York: Oxford University Press.

Christman, John. 1991. Autonomy and personal history. *Canadian Journal of Philosophy* 21: 1–24. https://doi.org/10.1080/00455091.1991.10717234.

Christman, John. 2004. Relational autonomy, liberal individualism, and the social constitution of selves. *Philosophical Studies* 117: 143–164.

Clark, Andy. 2004. *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford University Press.

Clark, Andy. 2010. *Review of adams and aizawa and Rupert*. http://manwithoutqualities.com/2010/11/06/clark-review-of-adams-aizawa-and-rupert/.

Clark, Andy, and David Chalmers. 1998. The extended mind. *Analysis* 58 (1): 7–19.

Clarke, Randolf, and Thomas Reed. 2015. Free will and agential powers. *Oxford Studies in Agency and Responsibility* 3 (1): 6–33. https://doi.org/10.1093/acprof:oso/9780198744832.003.0002.

Clowes, Robert. 2015. Thinking in the cloud: The cognitive incorporation of cloud-based technology. *Philos. Technol.* 28: 261–296. https://doi.org/10.1007/s13347-014-0153-z.

Davidson, Donald. 1980. Agency. In *Essays on action and events*, ed. Donald Davidson, 43–61. Oxford: Clarendon Press.

Davidson, Donald. 1980. Freedom to act. In *Essays on action and events*, ed. Donald Davidson, 63–81. Oxford: Clarendon Press.

Dworkin, Gerald. 1988. *The theory and practice of autonomy*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511625206.

Enç, Berent. 2003. *How we act: Causes, reasons, and intentions*. Oxford: Oxford University Press.
Farina, Mirko. 2013. Neither touch nor vision. Sensory substitution as artificial synaesthesia. *Biology and Philosophy* 28 (4): 639–655. https://doi.org/10.1007/s10539-013-9377-z.
Farina, Mirko. 2021. Embodiment: Dimensions, domains, and applications. *Adaptive Behavior* 29 (1): 73–99. https://doi.org/10.1177/105971232091296.
Farina, Mirko, and Andrea Lavazza. 2021. Knowledge prior to belief: Is extended better than enacted? *Behavioral Brain Sciences* 44: e152. https://doi.org/10.1017/S0140525X2000076X.
Farina, Mirko, and Andrea Lavazza. 2022a. Incorporation, transparency and cognitive extension: Why the distinction between embedded and extended might be more important to ethics than to metaphysics. *Philosophy & Technology* 35 (1): 1–21. https://doi.org/10.1007/s13347-022-00508-4.
Farina, Mirko, and Andrea Lavazza. 2022b. Mind embedded or extended: Transhumanist and posthumanist reflections in support of the extended mind thesis. *Synthese* 200: 507. https://doi.org/10.1007/s11229-022-03963-w.
Fischer, John Martin. 2000. Responsibility, history and manipulation. *Journal of Ethics* 4 (4): 385–391.
Fischer, John Martin, and Mark Ravizza. 1993. *Perspectives on moral responsibility*. Ithaca: Cornell University Press.
Frankfurt, Harry. 1971. Freedom of the will and the concept of a person. *The Journal of Philosophy* 68: 5–20. https://doi.org/10.2307/2024717.
Frankfurt, Harry. 1987. Identification and wholeheartedness. In *Responsibility, character, and the emotions*, ed. Ferdinand Schoeman, 27–45. Cambridge: Cambridge University Press.
Gallagher, Shaun, and Anthony Crisafi. 2009. Mental institutions. *Topoi* 28 (1): 45–51. https://doi.org/10.1007/s11245-008-9045-0.
Garrett, Brian. 1998. *Personal Identity and self-consciousness*. London: Routledge.
Goldman, Alvin. 1970. *A theory of human action*. Englewood Cliffs, NJ: Prentice-Hall.
Goldman, Alvin. 1979. What is justified belief? In *Justification and knowledge*, ed. G.S. Pappas, 1–25. Dordrecht: Reidel.
Greco, John. 1999. Agent reliabilism. *Philosophical Perspectives* 13: 273–296.
Habermas, Jürgen. 2003. *The future of human nature*. Cambridge: Polity Press.
Heersmink, Richard. 2017. Distributed selves: Personal identity and extended memory systems. *Synthese* 194: 3135–3151. https://doi.org/10.1007/s11229-016-1102-4.
Heersmink, Richard. 2012. Mind and artifact: A multidimensional matrix for exploring cognition-artifact relations. In *Proceedings of the 5th AISB symposium on computing and philosophy*, ed. Mark Bishop, Yasemin Erden, 54–61.
Hohwy, Jacob. 2004. The experience of mental causation. *Behavior and Philosophy* 32: 377–400.
Hohwy, Jacob. 2007. The Sense of self in the phenomenology of agency and perception. *Psyche: An International Journal of Research On Conscousness* 13: 1–20.
Horgan, Terrence. 2007. Mental causation and the agent-exclusion problem. *Erkenntnis* 67: 183–200. https://doi.org/10.1007/s10670-007-9067-9.
Horgan, Terence, John Tienson, and George Graham. 2003. The phenomenology of first-person agency. In *Physicalism and mental causation: The metaphysics of mind and action*, ed. Sven Walter and Heinz-Dieter. Heckmann, 323–40. Exeter, UK: Imprint Academic.
Hutchins, Edwin. 2011. Enculturating the supersized mind. *Philosophical Studies* 152 (3): 437–446. https://doi.org/10.1007/s11098-010-9599-8.
Jawad, Akram J. 2021. Engineering ethics of neuralink brain computer interfaces devices. *Annals of Bioethics and Clinical Applications* 4 (1): 1600. https://doi.org/10.23880/abca-16000160.
Kaczmarek, Kurt A., and Paul Bach-Y-Rita. 1995. Tactile displays. In *Virtual environments and advanced interface design*, ed. Woodrow Barfield and Thomas Furness, 393–414. New York: Oxford University Press.
Kant, Immanuel. 1996. The groundwork of the metaphysics of morals in practical philosophy. In *The cambridge edition of the works of immanuel kant*, ed. Paul Guyer and Allen Wood, 37–108. Cambridge: Cambridge University Press.
Kiverstein, Julian, and Mirko Farina. 2011. Embraining culture: Leaky minds and spongy brains. *Teorema* 32: 35–53.
Lambros, Malafouris. 2008. At the Potter's wheel: An argument for material agency. In *Material agency towards a non-anthropocentric approach*, ed. Carl Knappet and Lambros Malafouris, 19–36. Cham: Springer.
Lavazza, Andrea, and Mario De Caro. 2010. 'Not so fast. On some bold neuroscientific claims concerning human agency': *Neuroethics* 3: 23–41. https://doi.org/10.1007/s12152-009-9053-9.
Levy, Neil. 2007. Rethinking neuroethics in the light of the extended mind thesis. *American Journal of Bioethics* 7 (9): 3–11. https://doi.org/10.1080/15265160701518466.
Locke, John. 1689/1975. *An essay concerning human understanding*. Oxford: Clarendon Press.

Markosian, Ned. 1999. A compatibilist version of the theory of agent causation. *Pacific Philosophical Quarterly* 80: 257–277. https://doi.org/10.1111/papq.1999.80.issue-3.
Mckenna, Michael, and Chad van Schoelandt. 2016. Crossing a mesh theory with a reasons-responsive theory. Unholy spawn of an impending apocalypse or love child of a new dawn? In *Agency, freedom, and morall responsibility*, ed. Andrei Buckareff, Carlos Moya, and Sergi Rosell. London: Palgrave Macmillan.
Menary, Richard. 2010. Cognitive integration and the extended mind. In *The extended mind*, ed. Richard Menary, 227–243. Cambridge: MIT Press.
Merleau-Ponty, Maurice. 1962. *Phenomenology of perception*. New York and London: Routledge.
Merricks, Trenton. 1998. There are no criteria of identity over time. *Noûs* 32: 106–124.
Meyers, Diana T. 2005. Decentralizing autonomy: Five faces of selfhood. In *Autonomy and the challenges to liberalism*, ed. John Christman and Joel Anderson, 27–55. New York: Cambridge University Press.
Michael, Kirchhof D., and Julian Kiverstein. 2019. *Extended consciousness and predictive processing: A third wave view*. Cambridge: Routledge.
Nelkin, Dana Key. 2007. Do we have a coherent set of intuitions about moral responsibility? *Midwest Studies in Philosophy* 31: 243–259. https://doi.org/10.1111/j.1475-4975.2007.00159.x.
Nida-Rümelin, Martin. 2018. Freedom and the phenomenology of agency. *Erkenn* 83: 61–87. https://doi.org/10.1007/s10670-016-9872-0.
Noe, Alva. 2004. *Action in perception*. Cambridge MA: MIT Press.
O'Connor, Timothy. 2009. Agent-causal power. In *Dispositions and causes*, ed. Toby Handfield, 184–214. Oxford: Oxford University Press.
O'Regan, John Kevin. 2011. *Why red doesn't sound like a bell. Understanding the fell of consciousness*. Oxford University Press.
Schaefer, G. Owen, Guy Kahane, and Julian Savulescu. 2014. Autonomy and enhancement. *Neuroethics* 7: 123–136. https://doi.org/10.1007/s12152-013-9189-5.
Parfit, Derek. 1971. Personal identity. *Philosophical Review* 80: 3–27. https://doi.org/10.2307/2184309.
Peacocke, Christopher. 1979. *Holistic explanation: Action space, interpretation*. New York: Oxford University Press.
Prinz, Jesse. 2009. Is consciousness embodied. In *The cambridge handbook of situated cognition*, ed. Murat Aydede and Philip Robbins, 419–437. Cambridge: Cambridge University Press.
Pritchard, Duncan. 2010. Cognitive ability and the extended cognition thesis. *Synthese* 175: 133–151. https://doi.org/10.1007/s11229-010-9738-y.
Rhodes, Bradley J. 1997. The wearable remembrance agent: A system for augmented memory. *Personal Technologies* 1: 218–224.
Rupert, Robert. 2009. *Cognitive systems and the extended mind*. New York: Oxford University Press.
Sandberg, Anders, and Nick Bostrom. 2006. Converging cognitive enhancements. In *Progress in convergence: Technologies for human wellbeing*, ed. William S. Bainbridge and Mihail C. Roco, 201–227. New Jersey: Blackwell Publishing.
Shaun, Nichols, and Michael Bruno. 2010. Intuitions about personal identity: An empirical study. *Philosophical Psychology* 23: 293–312. https://doi.org/10.1080/09515089.2010.490939.
Shoemaker, Sydney. 1970. Persons and their pasts. *American Philosophical Quarterly* 7: 269–285.
Sosa, Ernest. 1988. Beyond skepticism, to the best of our knowledge. *Mind* 97: 153–189.
Sterelny, Kim. 2010. Minds: Extended or scaffolded? *Phenom Cogn Sci* 9: 465–481. https://doi.org/10.1007/s11097-010-9174-y.
Sutton, John. 2010. Exograms and interdiscilinarity: History, the extended mind and the civilizing process. In *The extended mind*, ed. Richard Menary, 189–225. MIT Press.
Swallow, Jessica. 2013. *Sharing the blame: Implications of the hypothesis of extended cognition for personal identity and ethics*. University of Exeter.
Taylor, Richard. 1966. *Action and purpose*. Englewood Cliffs, NJ: Prentice-Hall.
Tomczyk, Barbara. 2021. Knower at risk: Updating epistemology in the light of enhanced representation. *Studia Semiotyczne* 35(1): 35–54. https://doi.org/10.26333/sts.xxxv1.03.
Vallabhaneni, Anirudh, Tao Wang, and Bin He. 2005. Brain-computer interface. In *Neural engineering*, ed. Bin He, 85–121. Cham: Springer.
Vanacker, Gerolf, José R. del Millán, Eileen Lew, Pierre W. Ferrez, FerranGalán Moles, Johan Philips, Hendrik Van Brussel, and Marnix Nuttin. 2007. Contextbased filtering for assisted brain-actuated wheelchair driving. *Computational Intelligence and Neuroscience* 45: 123. https://doi.org/10.1155/2007/25130.
Wegner, Daniel. 2002. *The illusion of conscious will*. Cambridge, MA: MIT Press.