STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXV, nr 1 (2021), s. 35–54
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxv1.03
Article

BARBARA TOMCZYK *

KNOWER AT RISK: UPDATING EPISTEMOLOGY IN THE LIGHT OF ENHANCED REPRESENTATIONS

SUMMARY: The epistemological consequences of the increasing popularity of artificial cognitive enhancements are still confined to the margins of philosophical exploration, with priority given instead to ethical problems requiring urgent practical solutions. In this paper, I examine the less popular, yet still important, problem of the threats to which the very knowledge-forming process is exposed when its subject uses artificial cognitive enhancers. The theory of knowledge I call upon is borrowed from virtue epistemologists who, together with proponents of active externalism, seek to define the conditions that will protect artificially enhanced agents from a loss of epistemic agency. I invoke three such conditions (authenticity, integration and reciprocal causation), rejecting the last one. Incorporating active externalism into virtue epistemology points to the possibility of treating extended systems, composed of humans and artifacts, as extended subjects of knowledge. In the final part, however, I present two arguments against such an extension of epistemic agency.

KEYWORDS: cognitive enhancement, virtue epistemology, active externalism, extended cognitive system, epistemic agency.

Introduction

Knowledge is surely one of the most desirable of goods. It is considered a source of power and prosperity; its possession is rewarded and the lack of it rebuked. Modern technological developments have enabled the production of

* Maria Curie-Skłodowska University, Faculty of Philosophy and Sociology. E-mail: barbara.tomczyk@mail.umcs.pl. ORCID: 0000-0001-8145-7755.

artifacts supporting the acquisition of knowledge on a previously unprecedented scale—something that has inspired bold ideas about future possibilities. The most enticing of these is that of learning effortlessly and immediately, as seen in the 1999 movie *The Matrix*, where the acquisition of knowledge (of how to practice kung fu and operate a helicopter) by attaching electrodes to the agent's head seems to bypass all natural cognitive mechanisms. However, that which arouses fascination and envy among viewers of this movie raises concerns amongst epistemologists. Can beliefs or skills gained without effort embody the highest epistemic value? Can someone involved in this scenario be considered a genuine subject of knowledge? These doubts are of particular concern to virtue epistemologists, who define knowledge in terms of the cognitive achievement that the agent has attained using his or her own cognitive faculties, and for which he or she deserves credit. The need to reconcile the solutions of virtue epistemology with current and anticipated technological challenges has prompted precise analyses of the conditions that the agent's belief (or skill) must fulfill in order to be considered an instance of their knowing something. Although I will not be presenting any detailed analysis of these conditions, I will draw attention to them in order to answer the questions I myself find most pressing in the context of enhanced epistemic agency: Why, how, and from what should we protect the subject of knowledge that uses artificial cognitive enhancements? These questions are based on certain assumptions that I will be analyzing in subsequent parts of the article. To begin with, I indicate the particular understanding of epistemic agency that I intend to adopt. I borrow this from John Sosa, Ernest Greco and Duncan Pritchard—the founders of virtue epistemology. I also explicate the very need to protect agency against such threats. Next, I will substantiate the assumption about the existence of threats posed by the use of cognitive artifacts, and will point to the strategies for defense employed by virtue epistemologists. In the final part, I will consider the proposal of extending epistemic agency to include the entire extended cognitive system (consisting in each case of a human being and an artifact) that could potentially protect the agent's representations from the negative impact of using artifacts. Such a strategy seems to be an obvious consequence of including Clark and Chalmers' thesis of active externalism within the theory of knowledge under discussion. Nevertheless, I will show that this proposal does not achieve its intended goal, and therefore does not justify the introduction of the concept of an extended agent into epistemology.

Knowledge as a Cognitive Achievement and Its Subject

Knowledge is a mental representation of a special kind. Knowledge-that, which is of special interest to epistemologists, is a kind of belief: namely, an assertive attitude towards a given judgment. To count as knowledge, belief must meet additional requirements, of a kind that have been the focus of lively discussion amongst epistemologists ever since antiquity. For present purposes, I shall accept the conditions imposed on belief by proponents of virtue epistemology in

the form of an externalist and reliabilistic theory of knowledge: one that introduces the concept of cognitive ability to reliabilism. 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).[1] Virtue epistemologists have pointed out that this is an insufficient condition for knowledge, and have illustrated their point with many counterexamples.[2] Here, I present the most popular of them, which will also prove useful later on:

> TRUETEMP: Suppose a person, whom we shall name Mr. Truetemp, undergoes brain surgery by an experimental surgeon who invents a small device which is both a very accurate thermometer and a computational device capable of generating thoughts. The device, call it a tempucomp, is implanted in Truetemp's head so that the very tip of the device, no larger than the head of a pin, sits unnoticed on his scalp and acts as a sensor to transmit information about the temperature to the computational system in his brain. This device, in turn, sends a message to his brain causing him to think of the temperature recorded by the external sensor. Assume that the tempucomp is very reliable, and so his thoughts are correct temperature thoughts. All told, this is a reliable belief-forming process. Now imagine, finally, that he has no idea that the tempucomp has been inserted in his brain, is only slightly puzzled about why he thinks so obsessively about the temperature, but never checks a thermometer to determine whether these thoughts about the temperature are correct. He accepts them unreflectively, another effect of the tempucomp. Thus, he thinks and accepts that the temperature is 104 degrees. It is. Does he know that it is? (Lehrer, 1990, pp. 162–163)

The intuitive answer to the above question is "no". Thus, not every reliable belief-forming process leads to knowledge. According to virtue epistemologists, knowledge must derive from cognitive ability: i.e., the correctness of a known true belief must be due to the manifestation of a cognitive ability. Truetemp's belief derives from a reliable process, but not from any cognitive ability of his, and therefore he does not know (Greco, 2010; Pritchard, 2010). In order to assign Truetemp knowledge, the belief-forming process would have to have been appropriately integrated into his cognitive architecture, so that the belief would be the result of his cognitive abilities. Only after this condition is met can Truetemp, or any other agent, be considered a subject of epistemic credit and responsibility. The agent's cognitive character consists of all his cognitive abilities, both innate and acquired. What should be especially emphasized when discussing this theory of knowledge is the importance that its proponents attach to the properties of the belief-formation process itself. That is to say, this process cannot be truth-conducive through sheer luck, and cannot consist solely in the use of other people's cognitive abilities. Knowledge must be a product of the

[1] For detailed discussion of Goldman's theory of knowledge, see the present author's book-length study (Trybulec, 2012).

[2] Among the most influential virtue epistemologists, we should mention Ernest Sosa (1988; 2007), John Greco (1999), and Duncan Pritchard (2006).

cognitive abilities of its subject: only then does he or she own this special kind of representation—thus being responsible for it. What is important, moreover, is that the subject of knowledge does not have to be aware of the reliability of the processes resulting in this special kind of representation, or the extent to which they are integrated with his or her cognitive character. Virtue epistemologists are epistemic externalists, so they accept that the subject of knowledge need not know the way in which this epistemically valuable belief is formed. In the following, however, I intend to point out that this externalism has to be suspended when the agent, in order to solve a cognitive task, decides to go beyond his or her natural abilities and employ some artifact.[3]

There are two reasons why I refer to virtue epistemology when examining the influence of artificial cognitive enhancements on the process of acquiring knowledge. First, I recognize that its proponents have proposed an extremely insightful analysis of knowledge, presenting convincing solutions to many classic problems relating to this. Secondly, it is a theory that is constantly developing, whose proponents are actively engaged in upgrading previous solutions in the light of new cognitive phenomena and the philosophical concepts needed to explain them. Among such phenomena are artifacts that not only improve the natural cognitive processes, but also may, in the near future, enable the achievement of a cognitive goal that completely bypasses them. Yet the enthusiasm generated by such a vision is overshadowed by doubts as to whether such a process could be considered to represent a success on the part of the agent, such that he or she could be given credit for it. The growing popularity of artificial cognitive enhancements risks a blurring of epistemic responsibility and a decline in the value of knowledge—in which the latter may eventually cease to be a desirable achievement. Below, I will indicate in which cases of the use of artifacts the threat to cognitive achievement is the most real.

Cognitive Enhancements and Artificial Representations

The purpose of using cognitive enhancements is to quickly and effectively acquire knowledge, both propositional and procedural. Such enhancements include, in the broadest sense, any method that has the effect of improving the functioning of the human cognitive system. They can be divided into natural ones, such as learning, meditation and mnemonics, and artificial ones, which include the use of pharmacology, artificial intelligence and genetic modifications. In the narrow sense I am referring to in this paper, enhancement—as opposed to therapy such as is used to combat the effects of a neurological disease or injury—aims at improving the cognitive abilities of a healthy person. The improvement in question concerns both the receptivity of the human sensory apparatus and intellectual efficiency as this relates to memory, intelligence and creativity,

[3] For a more detailed discussion of virtue epistemology and its application to the study of extended cognitive systems, see the author's book (Trybulec, 2017).

and even to control over emotion, mood and desire (Sandberg, Bostrom, 1993). The use of artifacts that are external to the human body does not raise as many doubts as direct stimulation of the neural system. Thus, the ethical and epistemological discussion focuses mainly on cases of the second type, even though external enhancements also represent an important area of epistemological research.

Direct stimulation of the neuronal processes responsible for specific cognitive states usually takes the form of psychoactive substances or implants placed in appropriate areas of the brain. Such enhancements, due to their immediate effect, are much more effective than external artifacts but, on the other hand, they can lead to unforeseen, long-lasting and not always desirable side effects. As for psychoactive substances, many of these are obtained from plants commonly used to enhance attention, memory and creativity. Such effects are caused by, among other things, caffeine, theine, guaranine and nicotine, yet it is doubtful whether their use can be considered an instance of the enhancement of cognitive processes through artifacts. Meanwhile, there is no such doubt in the case of such chemicals as nootropic and precognitive drugs. These pharmaceuticals are mainly used therapeutically to slow down the cognitive damage caused by Alzheimer's and Parkinson's, and to prevent attention-deficit hyperactivity disorder (ADHD). However, they are also applied as cognitive enhancers in healthy people, because they improve the functioning of neurotransmitters and neurons, and ensure better blood circulation in the brain. A popular cognitive enhancer with a therapeutic purpose is, for example, Modafinil. Above all, this is used in the treatment of narcolepsy and sleep apnea, but it also has properties sought after by healthy people, as it accelerates the learning process by strengthening memory and engendering increased concentration (Gunia, 2015). Among the known psychoactive drugs that show a capacity for the enhancement of creativity, self-esteem and the desire for self-improvement, Prozak, an antidepressant, should also be mentioned. A much more dangerous group of enhancers are narcotic substances such as amphetamines and their derivatives, which stimulate and increase concentration, but are also highly addictive.

Alongside chemical substances, the largest group of artificial cognitive enhancers are IT artifacts created as a result of the development of artificial intelligence. As far as external artifacts are concerned, most of these function as memory stores, data-mining analysis and visualization programs aimed at supporting processes of reasoning, imagining and decision making (Kisielnicki, 2008). Devices connected to the human body, or implemented inside it, enter into more proximate and often reciprocal causal relations with brain processes, and a person usually does not have as much control over their operation as in the case of external artifacts. An example of the feedback that occurs directly between brain neural activity and such an artifact would be 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). One case of such an interface is furnished by the project presented in 2019 by the company Neuralink, which, although designed to help people with neurological

injuries, is ultimately intended to provide cognitive enhancement of unimaginable power by directly connecting the human brain with artificial intelligence. The connection consists in installing sensors in the brain in the form of thin threads that read neuronal activity and transmit the signal to an implant placed behind the ear. The implant, in turn, should decode this signal and send it to the computer running the appropriate program. As a result, it would be possible to send commands to artificial intelligence and receive information from the latter directly just via thought. The question that arises in the context of the discussion about agency is that of how much control a person would have over the representations directly produced in their mind by such an enhancement. It is the degree of this control that determines whether beliefs implemented in this artificial way can be considered knowledge understood as an achievement. Admittedly, not every representation that acquires the status of knowledge arises as a result of a human being's conscious decisions: that is not the case, for example, where sensory representations are concerned. All such mental states should, nevertheless, be produced by the cognitive abilities that belong to the person in question. Only then does he or she own these mental states and constitute their subject. In the case of the brain-computer interface described above, it seems that the representation can be created artificially, bypassing the natural cognitive process (or at least a significant part of it that is running in the perceptual apparatus). Yet is there really something wrong with that? In the next section, I will seek to justify a positive answer to that question by spelling out what I take to be the most serious threats to epistemic agency that are related to the use of artificial cognitive enhancers.

Enhanced, Yet Autonomous?

A necessary condition for assigning any kind of (moral, legal, epistemic) responsibility for some action undertaken, and thus for the possibility of its evaluation in terms of whatever value it realizes, is the intellectual autonomy of its subject. An agent is intellectually autonomous if he or she is able to make decisions according to his or her own will, and exercises control over the actions to which they lead. Obviously, the use of cognitive aids does not, as such, pose a threat to cognitive autonomy. Indeed, relying on other people's knowledge and obtaining information from reliable sources are essential for cognitive success. There is, however, a threshold beyond which this success ceases to be creditable to the agent: the agent must rely on others and other sources of knowledge "up to the point that doing so would be at the expense of her own capacity for self-direction. And this makes intellectual autonomy, essentially, a virtue of self-regulation in the acquisition and maintenance of our beliefs" (Carter, 2020a, p. 2940). The boundary of autonomous agency is not determined arbitrarily, but results from analyses of cases such as THRUTEMP, which have led virtue epistemologists to formulate the already-mentioned necessary condition for counting as knowledge. To recall, they require true belief to be the result of the agent's use of his or her own cognitive abilities. Adam Carter, one of the leading contemporary

virtue epistemologists, analyzes this condition in detail in the context of developing technological cognitive enhancements and when considering their impact on the knowledge-forming process (Carter, 2020c; forthcoming). Ultimately, he formulates a definition of autonomous belief, proposing a condition that is supposed to protect epistemic agency against possible threats from the use of the latest—or even just anticipated—technology. Before presenting this proposal, I will specify exactly what it is intended to protect the subject of knowledge against.

The discussion concerning the risk of using cognitive enhancements has mainly unfolded in the field of ethics, and has raised many important issues that call urgently for both solutions and appropriate regulative responses.[4] The problem of knowledge as addressed by those dealing specifically with ethical issues is most strongly related to the issue of agent autonomy. I will devote some attention to it, as it is the ground from which epistemological doubts have arisen.

It seems that supporting natural cognitive abilities through artifacts can only be beneficial. An agent is able to perform a given task faster or better, and sometimes its execution is simply impossible without the use of the relevant artifact. Intuitively, when it comes to identifying the agent *qua* initiator of the enhanced cognitive activity, the situation seems clear: it is a human being. Yet deeper reflection reveals a basis for doubt. If cognitive success depends on the use of an artifact without which it would not have happened, is it still the agent's achievement? The person using the artifact still remains the agent, as he or she is the initiator of the activity, but the resulting success does not seem to be entirely creditable to him or her. The intuition underlying the problem of authenticity can be expressed in the following question: To whom do we ascribe the greater cognitive achievement—the person who solves mathematical problems aided by nothing but their own memory, or the one who uses a calculator for this purpose? Everyday life shows that innate talent and skills that have been developed are valued more highly than the use of a cognitive enhancement, even where the persons involved achieve the same goal at the same time. It might seem that what we appreciate is the effort that a person using his or her own cognitive ability has to make to solve a given task, but this is not always the case. A genius can multiply three-digit numbers effortlessly, yet this does not earn him any less credit. What seems decisive for the decision to attribute achieved success is the agent's use of his or her own cognitive abilities, whether innate or developed. This is a kind of capital that is difficult to trade, and therefore has a special value. Naturally, by paying for a prestigious education one can acquire highly valued cognitive abilities, but the process is long and tedious compared to the immediate effects of some psychoactive substances or intracerebral implants.

The question about the agent's autonomy in the context of cognitive enhancement is therefore as follows: In a scenario where an agent uses an artifact to perform a given task, to whom should the achievement, and thus the epistemic

[4] Ethical considerations pertaining to cognitive enhancements have been explored by, among others, Jan-Christoph Bublitz (2013) and Walter Veit (2018).

responsibility, be ascribed? Can a person who checks the result of performing addition on a calculator be credited with adding numbers? It seems that in this latter situation no one can be credited with any achievement: the activity of adding numbers together simply did not occur, and there is no subject to which the success of the calculation can be attributed. The only action in this scenario is that of a human being checking the result in a calculator without calculating it. The agent who calculates the sum is the person who carries out addition in his or her head, or on a piece of paper—although the latter activity counts for less, as if the mere fact of aiding oneself in one's task with anything reduces the level of success. Imagine, though, a situation in which, after using Modafinil, a person performs a calculation in his or her head that he or she would not have been able to do without this enhancement. Thus, the agent does not exploit some process executed by an artifact (calculator), but rather employs an artifact (Modafinil) in order to perform a cognitive process that, if he or she had been more gifted or better educated, would have been achievable naturally, using just his or her own cognitive resources. Does such an enhancement raise similar doubts as the use of a calculator? From an ethical point of view, the use of this drug may still be questionable, in that it results in the playing field ceasing to be a level one between enhanced and unenhanced individuals. Epistemically, however, as I will show below, the situation is clear: the subject of calculation is the human being, and Modafinil does not shift the responsibility away from him or her—nor does it diminish his or her cognitive success.

Using a calculator or the Internet to perform cognitive tasks, while raising some questions about who should take the epistemic credit, does not undermine human agency. The person is still the initiator of cognitive activity, and chooses the method of achieving the goal. The real challenge epistemologists have to face is when cognitive enhancement disrupts the agent's identity, rendering the mental states that cause the action inauthentic. Here I am not referring to numerical identity, but rather, so to speak, to "being the same person" as before the enhancement: to the maintenance of psychological continuity with oneself—i.e., with one's own character. Only when the condition is met of identifying with one's enhanced mental states, feeling in control of them and having them as one's own, can the agent take epistemic responsibility for the actions they cause. If the cognitive enhancement is strong enough to disturb the sense of identity with one's "former self", if a person loses their sense of decision-making and exercising control over the actions in question, then their agency will be put in question, as will be the possibility of praising or blaming them for any possible success or failure. Such a situation may happen when, by directly affecting brain structure, the enhancement modifies representations that guide the agent in their actions, or the general dispositions and talents that define their personality. Changes of personality while retaining agent identity are of course possible, but they must be introduced in an appropriate manner over the course of a process of education, so as to allow for gradual assimilation. Rapid pharmacological or IT modifications are not properly coupled to the natural human cognitive mecha-

nism, making it difficult to identify the subject of the enhanced actions (Fischer, 2000). An additional complication is introduced by those enhancements—currently mainly pharmacological ones—that result in emotional states that are positively evaluated by the agent and mistakenly assessed by the latter as forming a part of his or her psychological character. The agent, guided in his or her action by such enhanced emotions, has a sense of agency, decision-making and preservation of identity, but the mental states responsible for determining action are not authentic, and this suffices to undermine his or her epistemic agency.

The problem of the agent's autonomy and the authenticity of their mental states in the context of cognitive enhancement has been analyzed in great detail from the perspective of ethics and the philosophy of law by Bublitz and Merkel (2009). These authors point out that 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 in the brain that takes over some of the natural cognitive functions. As for the pharmacological enhancers in current use, these do not constitute such replacements, as their operation consists in the optimization or modification of already existing structures and neural connections. Hence, the actions that result from these changes are still effects of the functioning of the mechanism owned by the agent in question, allowing the latter to retain full-blooded agency. On the other hand, such enhancements may well be regarded by those committed to the use of traditional, longer-lasting methods, such as involve an element of self-denial, as effortless shortcuts that cannot count as genuine cases of achievement. Nevertheless, these intuitions, motivated by a sense of unfairness, do not affect the epistemic status of enhanced representations, which, after meeting the appropriate conditions, can constitute full-fledged cognitive achievements. According to Bublitz and Merkel, the most important of these conditions is a conscious decision to utilize the enhancement made by an agent who knows the expected results of its application or, if unfamiliar with them, is aware of the risk being taken. In other words, a person, in order to be a responsible subject of his or her mental states and actions, cannot be manipulated in a way that is completely beyond his or her conscious control. When this happens, he or she ceases to be the subject of the actions performed, and the resulting belief cannot be regarded as their own cognitive achievement.

Even if the above condition is met, and the agent's identity is secure, those focused primarily on ethical issues remain concerned by the fact that, in the near future, cognitive enhancements may well remove certain obstacles in the absence of which it no longer makes sense to speak of something having been achieved (Kass, 2004). By depriving a human being of the need to make an effort, they will erase an important aspect of his or her life: one that relates to pride, praise, winning and admiration, but also to failure, shame and humiliation. When a goal comes effortlessly, it ceases to be an achievement and becomes an emotionless, trivial action that is hard to praise or criticize. The essence of this problem is accurately presented by Michael Sandel:

> [A]s the role of the enhancement increases, our admiration for the achievement fades. Or rather, our admiration for the achievement shifts from the player to his pharmacist… This suggests that our moral response to enhancement is a response to the diminished agency of the person whose achievement is enhanced. The more the athlete relies on drugs or genetic fixes, the less his performance represents his achievement. (Sandel, 2012, pp. 25–26)

When there is no possibility of losing, when one knows the "cheat code" for a given game, it loses its sense, as winning ceases to be satisfying in that it no longer delivers the same thrill. In most of the tasks that a person undertakes, effort is a necessary condition for considering its completion an achievement. Moreover, systematic artifactual support of a kind that frees the agent from the necessity of making any cognitive effort threatens him or her with an increasing level of dependence that may subsequently lead to complete cognitive impotence in situations where this enhancement is unavailable. This is what drivers who make constant use of car satellite navigation experience when their device fails or is fully discharged. Their employment of the enhancement causes their ability to orient themselves effectively in relation to their surroundings to decline drastically, resulting in a loss of epistemic agency. To counteract this threat, virtue epistemologists seek to precisely pinpoint those situations in which the use of cognitive enhancements contributes to a loss of control and agency, and how to avoid this.

Autonomous Belief, Reciprocal Causation, and Integration as Conditions for Epistemic Agency

Epistemologists, and those working in the area of ethics, agree that the most serious threat to epistemic agency is related to the possibility of manipulating the agent's cognitive processes and mental states in a way that is beyond his or her control. When this happens, the right to freedom of thought may be violated. More specifically, such a situation occurs when the agent is supplied, without his or her knowledge, with representations in a way that completely bypasses his or her natural cognitive process (*acquisition manipulation*), or when autonomous representations are, without his or her knowledge, eradicated from his or her mind (*eradication manipulation*) (Carter, 2020b). As long as the mind was reduced to a Cartesian thinking substance, and the content of mental states was available only to the subject, the threat of thought manipulation amounted to mere theoretical speculation. Yet technological developments 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 increasingly influential idea that it may extend beyond the skull, and even beyond the agent's organism, in a way that involves processes occurring, and information states obtaining, in artifacts themselves. This idea, proposed by Andy Clark and David Chalmers under the label of "active externalism", indicates that in some

cases of cognitive activity, a person is coupled with an external artifact in such a strong causal relationship (*continuous reciprocal causation*) that they co-constitute a single cognitive system (Clark, Chalmers, 1998). The physical realization base of cognitive processes, dispositional beliefs, or perceptual states may therefore extend beyond the safe, Cartesian "theater of the mind" into a publicly accessible world. Hence, given that some thoughts can be realized outside of the brain, they also need to be protected from the two types of manipulation mentioned above.

Adam Carter has carried out a highly detailed and insightful analysis of the condition that must be satisfied where autonomous belief is concerned, in order to serve as a protection in respect of artificially enhanced representations purporting to constitute knowledge. Here, I will only seek to the general contours of its overall outcome. In short, a belief will count as autonomous if, and only if, it has a compulsion-free history. This, in turn, will be the case if and only if the agent has not acquired the belief in a way that so bypasses or preempts his or her cognitive competences as to leave the agent improperly incapable of dispensing with that belief (Carter, 2020c). The subject of knowledge should, in other words, acquire a true belief as a result of using their own, unmanipulated cognitive abilities. If, however, these abilities are enhanced by some artifact, it should be properly integrated with the agent's cognitive character. This integration will be of a different nature to that which occurs in the process of acquiring new cognitive abilities or improving existing ones by methods of natural development. In the latter case, new dispositions do not have to be consciously accepted by the agent as is required in the scenario of an artifact being utilized. Virtue epistemologists, in collaboration with proponents of active externalism, have sought to explain how artificial enhancement can be integrated into the agent's cognitive character so that its use does not undermine their epistemic credit, and thus their knowledge.[5] In particular, they indicate two conditions that must be met for this to happen. First, according to the guidelines of Clark and Chalmers, the processes taking place within the agent and inside the artifact must be continuously linked via feedback loops. When this happens, the human being and the artifact form one system, in which the boundaries between organic and external processes are blurred, so that their separate study becomes futile. This kind of feedback only occurs if the enhancement is constantly present in the agent's life, easily and directly accessible, and applied uncritically, in a manner analogous to biological cognitive processes (Clark, Chalmers, 1998). Second, at some point in their life, the agent must have consciously incorporated the external enhancement into their cognitive abilities by accepting it as reliable (Pritchard, 2010).[6]

[5] Notably, the "Extended Knowledge Project", led by Duncan Pritchard, was undertaken at the University of Edinburgh from 2013 to 2015. As a part of this, virtue epistemologists (Pritchard, Carter) collaborated with proponents of active externalism (Clark, Palermos). The results obtained were published in book form (Carter, Clark, Kallestrup, Palermos, Pritchard, 2018).

[6] This condition is also present in Clark and Chalmers (1998), and in Rowlands (2010).

This requirement is illustrated by the Truetemp case described in the first paragraph of the present article. To remind readers, Truetemp, although he can determine the temperature in the room, has no knowledge of it, because this belief did not arise as a result of his cognitive abilities, but rather due to the operation of a device inserted into his brain. In order to attribute knowledge to Truetemp, it must be at least assumed that he knows the source of his true beliefs and has accepted them as reliable. Now let us consider another case. We may imagine that a scientist has installed a sensory substitution system in the body of a blind person without his or her knowledge. It is a device that converts information specific to the damaged sense modality into stimuli received by a working one. Would we consider the cognitive success caused by the operation of such a system to be the result of this person's use of his or her extended cognitive abilities? It seems not. Such a person is in the same epistemic scenario as Truetemp, because he or she has never consciously included a new competence into the framework of his or her cognitive system. They do not know the source of their reliability, and so could not be credited for the success in question.

The doubts that pertain to the influence of cognitive enhancements on epistemic agency do not therefore concern the sheer fact of their application, but rather their proper integration with the agent's cognitive system. At this point, it is worth emphasizing once again the difference between biological (natural) and extended (enhanced) cognitive processes. In the case of the former, the condition of consciously endorsing them as reliable and making a decision to use them does not have to be met in order for them to count as constitutive of the agent's cognitive character. This condition concerns only artificial cognitive enhancements used to improve biological processes. However, we should keep in mind that virtue epistemology typically adopts a reliabilistic stance towards knowledge. To remind readers, the epistemic status of a belief is determined, according to its proponents, by properties of the belief-forming process. Moreover, the agent need not be aware of these properties, and need not know whether the process is reliable or whether it meets other conditions proposed by virtue epistemologists. In this respect it is tantamount to an externalist theory of knowledge. Yet the above considerations pertaining to the need for conscious integration of cognitive enhancements with biological processes on the part of agents are internalistic in nature. In order to incorporate the process of manipulating an artifact into the framework of their cognitive systems, agents must, at some point in their lives, consciously acknowledge this enhancement as reliable, and embrace its continuous and unreflective utilization. In other words, an agent must know, or have known at some point in their life, the reasons underpinning the belief they now have as a result of using the relevant artifact. Hence, while working out the conditions governing knowledge for artificially enhanced agents, the virtue epistemologist must part company with the externalists and take up instead the position of some kind of internalist-reliabilistic hybrid.[7]

[7] I also develop this line of reflection in my book-length study (Trybulec, 2017).

Upgrading Epistemology With Active Externalism: Some Problems

All the conditions for the safe use of cognitive enhancement indicated in the previous paragraph focused on its integration with the agent's cognitive character. The most important challenge for epistemologists is to explain what, exactly, this integration is supposed to amount to. One answer, as I have already shown, is suggested by proponents of active externalism. Let me recall that, according to Clark and Chalmers, proper integration should consist of a continuous and reciprocal causal link between the agent's natural cognitive abilities and the processes taking place in the artifact itself. This means that the functioning of the former changes the operation of the latter, which in turn affects the former, and so on.[8] It is worth pausing for a moment here to reflect carefully on this. It will not take long before one realizes that the condition of reciprocal causal coupling appears too strong and difficult to fulfill when using some artificial enhancements. Is it possible, for example, to constitute such a dynamic system out of the conjunction of a human being with a psychoactive substance such as Modafinil? It seems that in the scenario of taking a pill, the causal relationship is one-sided and consists solely in the effect of the substance on the human nervous system, without feedback. The human being can, at most, monitor the changes taking place in his or her cognitive functioning and control the dose of the substance, but is not able, consciously or not, to change its impact on his or her cognitive character. Nevertheless, the condition of exerting control and retaining a sense of agency in the face of such an enhanced cognitive character is fulfilled, and it would be implausible to claim of such an artifact that it had taken epistemic responsibility away from the agent. The belief generated with the support of Modafinil is autonomous, and the agent has deliberately decided to incorporate this substance into the cognitive abilities responsible for this mental state. It seems, therefore, that the condition of reciprocal causal coupling, considered necessary by Clark and Chalmers for the existence of an extended system, is too strong when it comes to determining what counts as an epistemically safe utilization of a pharmacological artifact. In short, not every coupling between a human and an artifact that results in knowledge constitutes an extended cognitive system.

Active externalism seems to fall short of the hopes invested in it by epistemologists: the condition that it specifies, of an enhancement's having to be integrated with the agent's cognitive character, is not necessary for knowledge to be obtained through manipulation of the artifact. There is, moreover, a tension between active externalism and virtue epistemology, due to the internalist condition that requires the agent to consciously embrace the extended cognitive process as being reliable. That is to say, it does not favor the functionalist attitude that marks out supporters of active externalism in their dispute with those seeking to assert the importance of biologically determined prejudices ("bio-prejudices").

[8] The idea of mutual feedback as a necessary condition for epistemic subjectivity being enhanced by an artifact is analyzed in detail by Palermos (2014).

According to functionalists, the nature of the cognitive process (be it 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 processes are not epistemically equivalent. As has already been noted, in order to incorporate the manipulation of artificial cognitive enhancement into the agent's cognitive character, the agent must consciously and freely decide about it, which he or she need not do in the case of such biological processes as we see manifested in our ordinary perceptual or rational faculties.

To maintain epistemic agency, the agent supporting himself or herself with some artifact should be concerned about its proper integration with their cognitive character. When deciding to use an artificial cognitive enhancement, they ought to be vigilant and attentive. The more thoroughly agents have familiarized themselves with how an artifact works, and how it affects their natural cognitive processes, the better protected they will be against manipulation or loss of control over the corresponding artificially enhanced process of belief-formation. Active externalism defines the conditions for an extended cognitive system whose cognitive processes are distributed and impossible to divide into the biological and the artificial. Yet, as was shown, not every use of an artifact in the knowledge-forming process constitutes such a system. Any such use, however, requires a person to consciously and freely accept the coupling between artificial enhancement and his or her natural cognitive processes, regardless of whether it be one-way or reciprocal. Hence, even in the case of very radical cognitive enhancement of the sort that is, at present, only part of a boldly anticipated future, maintaining the agent's cognitive autonomy is possible. Human cognitive dependence on technology is inevitable, but so long as epistemic vigilance is maintained it need not be detrimental to our epistemic agency. To reiterate, the possibility of assigning a cognitive achievement of sorts to the agent will be determined not so much by the type of enhancement utilized by the latter, but rather by the kind of influence it exerts on the agent and the degree of control the agent exercises over it.

Beyond Control

Even when all conditions for knowledge acquired with the support of artificial enhancement are met, epistemologists still have reservations. By way of concluding these considerations, I will point to the two areas of doubt that I consider the most serious. The internalist condition requiring the agent to consciously accept the impact of artificial enhancement on natural cognitive abilities significantly limits the technological possibilities for generating knowledge. That is, one cannot produce it by secretly installing a belief-forming implant, or administering a psychoactive substance to the agent. Of course, it is possible—albeit only theoretically, for the time being—to artificially and discreetly create in the agent's mind a representation with some appropriate content, but this will not

count as knowledge from an epistemological point of view. Hence, the subject of knowledge seems to be protected from cognitive manipulation, yet the question arises as to how realistic and effective this protection is in practice. The consequences of using an enhancement, such as a psychoactive substance, can be somewhat unpredictable not only for the agent, but also even for specialists charged with controlling its use. Even if we assume that the agent is familiar with the nature of the influence exerted by a given substance, he or she may not be able to distinguish between his or her natural mental states and those produced by the enhancement itself. As a consequence, the agent loses control over the artifact and becomes susceptible to manipulation by other people, which leads to a loss of ownership of the resulting mental state. On the other hand, as was already indicated, after consciously incorporating enhancement into his or her cognitive character, the agent no longer needs to constantly control it. The artifact can become a part of the agent's cognitive system that works beyond the bounds of his or her consciousness. Were it not for the problematic internalistic condition that speaks in favor of "bio-prejudices", this would be an ideal scenario for adherents of active externalism. The extended cognitive system would function as a natural one and would not require any special treatment. The bad news, however, is that special treatment is indeed necessary—a point emphasized not only by virtue epistemologists, but also by the proponents of active externalism themselves. Clark and Chalmers give expression to this necessity by formulating four conditions for having a mental state (a belief) partly realized by an artifact (Clark, Chalmers, 1998), thus lending support—surely against their own intentions—to the thesis propounding the cognitive advantageousness of biological processes.

Another weak point when it comes to defending epistemic agency against the negative influence of cognitive enhancement concerns the authenticity of the mental states responsible for its control and the sense of ownership of the cognitive character that results. Carter's account of what is required in order to preserve the authenticity of belief draws attention to the necessity of using only the agent's cognitive abilities in the knowledge-forming process. Imagine, however, that the enhancement (be it a pharmaceutical one or an implant), though applied by the agent voluntarily, shapes his or her mental states responsible for the sense of control and agency. Assume, moreover, that the agent has agreed to such an influence, and—even more—that he or she has agreed to the artifact changing his or her identity (desires, emotions and beliefs). Are his or her mental states still authentic? It seems not, since they did not result from the agent's cognitive abilities. On the other hand, the agent has consciously and voluntarily incorporated the artifact into his or her cognitive character, making its processes his or her own. Actually, if there is reciprocal causation between natural and artificial processes, it is difficult to distinguish one from the other because they shape each other. Hence, it becomes impossible to determine whether a given mental state was triggered by the agent's cognitive abilities or by artificial processes that bypass his or her cognitive character. Even if this scenario represents no more

than an audacious imagining of future possibilities, epistemologists surely need to prepare for it and be aware of any doubts about, or threats to, epistemic agency that it may bring on, even if they do not have ready solutions yet. Maybe, in the scenario just described, it would make sense to accept the proposal that, together with a human being, the artifact constitutes the agency of an extended system, or even that it comes to partly make up the subject of knowledge produced within such a system.[9] Having said this, while tempting, I myself do not consider this solution satisfactory.

Extending the realization base of epistemic agency to an artifact, while it may seem theoretically possible, does not, in my opinion, compel us to accept the thesis of an extended subject of knowledge. I would like to point out two reasons for such a verdict. Firstly, the subject of cognition bears epistemic responsibility for success or failure. Yet only a reflective system can be thus responsible. Such a system is distinguished by the ability to assess one's own mental states in terms of rationality and compliance with some adopted hierarchy of values. It also has the ability to make a free choice based on consideration of its possible consequences. In order to do that, an agent must have access to the contents of his or her mental states, be aware that they belong to himself or herself, realize that they derive from his or her own cognitive abilities, and be of the conviction that he or she controls them. Even if these conditions are met when aided by some cognitive enhancement, epistemic responsibility, which is associated with the apportioning of credit and blame, rests with the human and not with the human-plus-artifact. Surely, though, this is not the case when the sense of agency is created without one's knowledge or will, as it is when one's natural cognitive abilities are completely bypassed or manipulated, so that the condition of autonomy and cognitive integration is not met. In any other (non-pathological) case, the subject of knowledge will be the human being, because only he or she can be the object of epistemic evaluation—and of any reward or punishment associated with this.

One may wonder, nevertheless, whether it is at all possible for the realization base of epistemic agency to be extended without the agent itself also being so. Since mental states determining agency would be co-realized by artificial processes linked via reciprocal causation with natural ones, why not consider them states of the extended agent taken as a whole, and not just of one of its parts (i.e., the human being)? The first reason for not doing so has been outlined above, and concerns our intuitions and practices relating to the attribution of agency. At the same time, a theoretical grounding for this has been provided by Lynne R. Baker (2009), and this may itself be regarded as furnishing our second reason for confining epistemic agency to the human being within an extended cognitive system. While Baker's proposal concerns our understanding of the self in extended cognitive systems, it is not too much of a distortion to apply it to mental states in general. She refers to the division of reality into levels introduced by proponents

[9] The thesis of extended agency has been developed by, among others, Malafouris (2008). For its analysis and evaluation by the present author (Trybulec, 2020).

of nonreductive physicalism. Mental states, according to this stance, belong to the properties of a higher level of the cognitive system, and arise from a lower level, that of physical properties. The two types of properties have different characteristics. Physical properties, as opposed to mental ones, manifest themselves in space—within the agent's organism, or outside it. Higher-level systemic properties, such as agency, do not occupy space, so it is impossible to determine whether they are inside or outside the agent's body. As Baker argues, the fact that the social, linguistic, and physical environment plays a vital role in shaping the agent's mental states, and even partly realizes some of them, does not mean that the agent himself or herself is extended in any way. In other words, the agent's subpersonal states may consist in part of extra-biological elements that, by entering into complex causal relationships with one another, produce higher-level systemic properties such as beliefs, desires, and other mental states. The physical realization base of agency is in this case extended, but the agent itself is not, as the term "extended" applies only to physical properties. This observation seems to undermine the very thesis of the extended mind as put forward by Clark and Chalmers. However, I will not address that problem here. At this point, I would like instead to just focus on drawing the conclusion that artificial cognitive enhancements cannot take over some of the epistemic credit and responsibility from human beings, and therefore cannot share epistemic agency with them.

Concluding Remarks

The goal I set myself in this paper was to consider the epistemological consequences of the increasing popularity of artificial cognitive enhancements. Technological developments that are such as to allow for reasonable predictions as to their future mode of operation are of legitimate concern to philosophers studying the conditions of agency. The alarm has been raised primarily by those dealing with ethics, as the consequences of the increasing influence of artifacts on the human mind are linked to practical issues of social justice, and so demand urgent regulation. In the present article, though, I have sought to address another dimension of this phenomenon—the epistemological worries, which are less popular and therefore less frequently raised. I have pointed to scenarios in which the use of an artifact may deprive the agent of cognitive achievement, making him or her lose epistemic agency. I have also looked at three conditions for enhanced belief and knowledge (authenticity, reciprocal causation, and integration) that are suggested in the literature on this issue, and have dismissed one of them (reciprocal causation) as unnecessary. Despite my rejection of this condition, I find the collaboration of virtue epistemologists with supporters of active externalism to be most fruitful. The latter have certainly enriched epistemological considerations with their explanation of the relationship that unites a human being and an artifact into a single knowledge-forming (epistemic) system, and this suggests the possibility of treating such an object as an extended agent, where mental states such as knowledge, intentions and desires belong not just to

the human being, but rather to the entire system. Such an approach would make it possible to solve the problem of what it means for human identity and agency to be distorted by an artifact that has nevertheless been correctly incorporated into the framework of the agent's cognitive character: in such cases, an artifact would co-constitute an instance of extended identity and agency—i.e., it would share these with the human being involved. On the other hand, in the final part of this paper, I have presented two arguments against such an extension of epistemic agency. Of these, the former refers to the close connection of agency with responsibility, while the latter invokes the concept of systemic properties that have different characteristics from their physical realization base.

As a consequence of the considerations pursued here, a doubt may arise as to why we should care about protecting epistemic agency at all. Is the dissolution of the subject of knowledge really something we should fight against? Well, yes! The decline of the epistemic agent entails a fading away of epistemic responsibility. That is to say, if there is no one to attribute a given achievement to, then no one can be responsible for either the cognitive success in question or its absence. Epistemologists are resolutely engaged in searching out the conditions for knowledge that will serve as its touchstone in every—even the most fantastic—scenario. These efforts, though, are not driven solely by theoretical ambitions. Doubts about the subject of knowledge resulting from enhanced cognitive abilities may already, in the near future, cause practical problems relating to the need to determine who should be praised or blamed for a given result. In this paper, I have also pointed to the problem of the reduction of cognitive effort, which becomes ever more serious, the more frequently and systematically people use enhancements. Both the lack of a need to demonstrate one's own skills and the lack of any risk of failure contribute to lowered self-esteem, as well as to a diminution in the sense of satisfaction associated with success and of anger connected with failure—both emotions that motivate self-improvement and development. All these doubts and concerns are sufficient reasons to care about the authenticity of our mental representations, and for taking seriously appeals for epistemic control and vigilance in the face of the rapid technological developments surrounding cognitive enhancements.

REFERENCES

Baker, L. (2009). Persons and the Extended Mind Thesis. *Zygon, 44*(3), 642–658.

Bublitz, J.-C., Merkel, R. (2009). Autonomy and Authenticity of Enhanced Personality Traits. *Bioethics, 23*(6), 360–374.

Bublitz, J.-C. (2013). My Mind is Mine!? Cognitive Liberty as a Legal Concept. In: E. Hildt, A. Francke (Eds.), *Cognitive Enhancement* (pp. 233–264). Dordrecht, New York: Springer.

Carter, A. (2013). Extended Cognition and Epistemic Luck. *Synthese, 190*(19), 4201–4214.

Carter, A., Clark, A., Kallestrup, J., Palermos, S. O., Pritchard, D. (Eds.). (2018). *Extended Epistemology*. Oxford: OUP.

Carter, A. (2020a). Intellectual Autonomy, Epistemic Dependence and Cognitive Enhancement. *Synthese, 197*, 2937–2961.

Carter, A. (2020b). Varieties of (Extended) Thought Manipulation. In: M. Blitz, C. Bublitz (Eds.), *The Future of Freedom of Thought: Liberty, Technology, and Neuroscience*. London: Palgrave Macmillan. Manuscript submitted for publication.

Carter, A. (2020c). Epistemic Autonomy and Externalism. In: K. Lougheed, J. Matheson (Eds.), *Epistemic Autonomy.* London: Routledge.

Clark, A., Chalmers, D. (1998). The Extended Mind. *Analysis, 58*(1), 7–19.

Fischer, J. (2000). Responsibility, History and Manipulation. *Journal of Ethics, 4*, 385–391.

Goldman, A. (1979). What is Justified Belief? In: G. S. Pappas, (Ed.), *Justification and Knowledge* (pp. 1–25). Dordrecht: Reidel.

Greco, J. (1999). Agent Reliabilism, *Philosophical Perspectives*, *13*, 273–296.

Greco, J. (2010). *Achieving Knowledge: A Virtue-Theoretic Account of Epistemic Normativity*. Cambridge: CUP.

Gunia, A. T. (2015). Koncepcje wzmocnienia poznawczego. Próba definicji oraz przegląd metod. *Avant, VI*(2), 35–56.

Kass, L. R. (2004). *Life, Liberty and the Defense of Dignity: The Challenge for Bioethics*. San Francisco: Encounter Books.

Kisielnicki, J. (2008). *MIS – Systemy Informatyczne Zarządzania*. Warsaw: Placet.

Lehrer, K. (1990). *Theory of Knowledge*. London: Routledge.

Malafouris, L. (2008). At the Potter's Wheel: An Argument for Material Agency. In: C. Knappet, L. Malafouris (Eds.), *Material Agency. Towards a Non-Anthropocentric Approach* (pp. 19–36). New York: Springer.

Palermos, O. S. (2014). Knowledge and Cognitive Integration. *Synthese, 191*, 1931–1951.

Pritchard, D. (2006). *What is This Thing Called Knowledge?* New York: Routledge.

Pritchard, D. (2010). Cognitive Ability and the Extended Cognition Thesis. *Synthese, 175*, 133–151.

Rowlands, M. (2010). *The New Science of Mind*. Cambridge MA: The MIT Press (A Bradford Book).

Sandberg, A., Bostrom, N. (2006). Converging Cognitive Enhancements. In: W. S. Bainbridge, M. C. Roco (Eds.), *Annals of the New York Academy of Sciences* (1093, pp. 201–227). Oxford: Blackwell.

Sandel, M. J. (2012). *The Case against Perfection: What's Wrong with Designer Children, Bionic Athletes, and Genetic Engineering?* In: S. Holland (Ed.), *Arguing About Bioethics* (pp. 25–26). London: Routledge.

Sosa, E. (1988). Beyond Skepticism, to the Best of our Knowledge. *Mind, 97*, 153–89.

Sosa, E. (2007). *A Virtue Epistemology: Apt Belief and Reflective Knowledge*. Oxford: Clarendon Press.

Trybulec, B. (2012). *Epistemologia znaturalizowana a normatywność*. Lublin: Wydawnictwo UMCS.

Trybulec, B. (2017). *Wiedza i jej podmiot w szerokich systemach poznawczych*. Warsaw: IFiS PAN.

Trybulec, B. (2020). Podmiot czy agent? Rozumienie podmiotowości w erze artefaktów poznawczych. *Filozofia i Nauka. Studia filozoficzne i interdyscyplinarne, 8*(2), 89–113.

Vallabhaneni, A., Wang, T., He, B. (2005). Brain-Computer Interface. In: B. He (Ed.), *Neural Engineering* (pp. 85–121). New York: Springer US.

Veit, W. (2018). Cognitive Enhancement and the Threat of Inequality. *Journal of Cognitive Enhancement, 2*, 404–410.