Synthese (2020) 197:4065–4081
https://doi.org/10.1007/s11229-018-01920-0

How to stay safe while extending the mind

Jaakko Hirvelä[1]

Received: 18 December 2017 / Accepted: 20 August 2018 / Published online: 5 September 2018

Abstract
According to the extended mind thesis, cognitive processes are not confined to the nervous system but can extend beyond skin and skull to notebooks, iPhones, computers and such. The extended mind thesis is a metaphysical thesis about the material basis of our cognition. As such, whether the thesis is true can have implications for epistemological issues. Carter has recently argued that safety-based theories of knowledge are in tension with the extended mind hypothesis, since the safety condition implies that there is an epistemic difference between subjects who form their beliefs via their biological capacities and between subjects who have extended their cognition. Kelp, on the other hand, has argued that a safety-based theory of knowledge can be correct only if the extended mind thesis is true. While these claims are not logically inconsistent, they do leave the safety theorist in an uncomfortable position. I will argue that safety-based theories of knowledge are not hostage to the truth of the extended mind thesis, and that once the safety condition is properly understood it is not in tension with the extended mind thesis.

Keywords Extended mind thesis · Safety condition · Virtue epistemology · Luck · J. Adam Carter · Christoph Kelp

1 Introduction

According to the extended mind thesis, laid out by Clark and Chalmers (1998), cognitive processes can extend beyond the boundaries of skin and skull. The proponents of the extended mind thesis subscribe to the following principle:

> M-PARITY PRINCIPLE: If, as we confront some task, a part of the world functions as a process which, were it to go on in the head, we would have no

✉ Jaakko Hirvelä
jaakko.hirvela@helsinki.fi

[1] Department of Philosophy, History, Culture and Art Studies, University of Helsinki, Unioninkatu 38 PL 24, 00014 Helsinki, Finland

> hesitation in accepting as part of the cognitive process, then that part of the world is part of the cognitive process. (Clark & Chalmers, 1998, p. 8)

The m-parity principle guards against the metaphysical prejudice of giving a privileged role to the processes that occur within the boundaries of our bodies when explaining our cognition. Those who accept the m-parity principle think that we should give no special weight to the processes that occur within our bodies, since processes that are external to our bodies can function in the same way from a common-sense functionalist point of view and serve the same roles as intracranial processes. Given that the extended mind thesis is a metaphysical thesis it can have epistemological consequences. Carter argues that the m-parity principle motivates an epistemic parity principle:

> E-PARITY PRINCIPLE: For agent S and belief p, if S comes to believe p by a process which, were it to go on in the head, we would have no hesitation in ascribing knowledge of p to S, then S knows p. (Carter, 2013, p. 4203)

The e-parity principle is supposed to guide epistemic theorizing in a similar way as the m-parity principle guides metaphysical theorizing by guarding us against unwanted prejudice. The mere locality of a process is not an epistemically relevant factor. In the classic example of cognitive extension that Clark and Chalmers (1998) present, the process of consulting the notebook is the extracranial analogue of consulting one's biological memory[1]:

> OTTO: Otto suffers from Alzheimer's disease and like many Alzheimer's patients, he relies on information in the environment to help to structure his life. Otto carries a notebook around with him everywhere he goes. When he learns new information, he writes it down. When he needs some old information, he looks it up. For Otto, the notebook plays the role usually played by a biological memory. (Carter, 2013, p. 4202)

It seems that whatever Otto writes in the notebook he dispositionally believes. Accordingly, his notebook entries are a part of the physical basis for his dispositional knowledge. To hold otherwise would be to commit bioprejudice. Several theories of knowledge are able to deliver the correct verdict regarding this case. For instance, virtue epistemological theories of knowledge are able to accommodate this case of knowledge since Otto exhibits a great deal of cognitive virtue in his actions (Pritchard 2010, p. 145).[2] Otto updates the notebook meticulously and the way in which he retrieves information from the notebook is reliable. Moreover, there is no reason to suppose that Otto does not satisfy the safety condition when he forms true beliefs as a result of consulting the notebook or that the beliefs stored in the notebook would not be safe. According to a very rough formulation of the safety condition, a subject's belief that p is safe just in case it could not easily have been false given the way in which it

[1] See Wikforss (2014, pp. 470–472) for the argument that the cognitive process that Otto undergoes while consulting his notebook is not functionally similar to the cognitive process that he would go through if he consulted his biological memory.

[2] Pritchard (2010), Vaesen (2011) and Kelp (2014) have argued that other cases of extended cognition are problematic for robust virtue epistemology. Greco (2012) defends robust virtue epistemology from the argument raised by Vaesen.

was formed. If the notebook contains mostly true information which is gathered in a reliable way, it is entirely possible that Otto could not easily have formed a false belief by consulting the notebook.

In what follows I will argue that safety-based theories of knowledge are compatible both with the truth and with the falsity of the extended mind thesis once the safety condition is properly understood. In Sect. 2 I will lay out the safety condition and the kind of luck that it seeks to eliminate from the realm of knowledge. In Sect. 3 I will present Carter's argument for the tension between the safety condition and the extended mind thesis. In Sect. 4 I will argue that by recognizing that the safety condition must be globalized to a set of propositions we are able to dismiss Carter's argument for the apparent tension between safety-based theories of knowledge and the m-parity principle. In Sect. 5 I will argue that in order to dissolve the tension between safety-based theories and the e-parity principle we need to relativize the safety condition to virtuous methods of belief formation that the subject uses in the actual world. In Sect. 6 I will argue contra Kelp that safety-based theories of knowledge are not hostage to the truth of the extended mind thesis.

2 Safety and epistemic luck

Before moving on, it is useful to have a more precise account of the safety condition and the kind of luck that it aims to eliminate from the realm of knowledge. The safety condition is often put forward as an anti-luck condition for knowledge. As such, beliefs that are safe are supposed to be non-luckily true. In particular, the safety condition is motivated by the modal account of luck as developed by Pritchard. According to the modal account of luck, the fact that event E occurred is a matter of luck for subject S only if:

(i) E occurs in the actual world but fails to occur in most nearby possible worlds where the relevant initial conditions for E are the same, and,
(ii) E is a significant event for S (Pritchard 2005, pp. 129, 132).[3]

On this account, a belief is true as a matter of luck only if it is true in the actual world but false in most nearby possible worlds where formed on the same basis. Notice that when considering whether a belief is true as a matter of luck condition (ii) is automatically satisfied, since all true beliefs are (at least somewhat) significant. Even though the modal account of luck delivers correct verdicts regarding a wide range of cases, there is reason to believe that its focus is too narrow in that it asks us to consider only the modal profile of the event that occurred in the actual world when considering whether the occurrence of that event is a matter of luck for the subject. To see why it is necessary to consider the modal profile of other similar events that are equally significant, consider the following case offered by Coffman (2007, pp. 395–396):

GAME SHOW: Suppose that S is on a game show, and that there was just before t no chance S would neither receive the prize concealed

[3] Note that Pritchard (2015) has in his later work abandoned condition (ii). I think this is a mistake, but I will not defend that claim here.

> by Door 1 nor receive the prize concealed by Door 2. Now, let E1 be S's receiving the prize concealed by Door 1, and let E2 be S's receiving the prize concealed by Door 2, where there is just before t only a small chance that E1 will occur at t. Further, suppose the prizes concealed by Doors 1 and 2 are equally good for S. Finally, suppose E1 occurs at t.

Intuitively, S is not lucky in receiving the prize concealed by Door 1 (Coffman 2007, p. 396). This is because in the vast majority of nearby possible worlds S receives the prize concealed by Door 2, and that prize would have been just as good for S. In order to deal with this case we need to reformulate condition (i) as follows:

> (i)* E occurs in the actual world and neither it nor any event E* that is of the same type and at least as significant to S, occurs in most nearby possible worlds where the relevant initial conditions for E are the same.[4]

The reformulation of the modal account of luck can also be motivated by considering the metaphysics of events. It is plausible to think that events are partially individuated in terms of the time when they occur. For example, on the property exemplification account of events an "event (or state) is a structure consisting of a substance (an *n*-tuple of substances), a property (an *n*-adic relational attribute), and a time" (Kim 1976, p. 160). On this account two events are the same just in case they are identical with respect to the substance, property and time.

To see how this account of events motivates condition (i)*, assume that an event E^1 which is significant to S occurs in the actual world at t^1, but does not occur in most nearby possible worlds where the relevant initial conditions for E^1 are the same. Assume also that in most nearby possible worlds (where the initial conditions for E^1 stay the same) an event E^2 that is equally significant to S as E^1 is and identical with respect to substance and property to E^1, occurs at t^2. On the original modal account of luck, the fact that E^1 occurred in the actual world is a matter of luck for S. Given, however, that the events E^1 and E^2 are very similar, and that there is only a small temporal interval between the E^1 and E^2, the modal profile of E^2 is relevant when determining whether the fact that E^1 occurred is a matter of luck for S. Condition (i)* delivers the correct verdict that E^1 is not a matter of luck for S, since an equally significant event of the same type could very easily have occurred in its stead.

The reformulation of the modal account of luck motivates a globalized safety condition, according to which a subject S's true belief that *p*, formed via method M, is safe only if S could not easily have formed a false belief via M. More precisely:

SAFETY: S's belief that *p*, which belongs to a set of propositions *Q*, is safe if and only if:

[4] Coffman (2007, p. 396) argues that the events in question need to be similar to each other in order to count as relevant. What events count as relevantly similar varies from case to case. This is, of course, rather vague, but given that we are not trying to provide a reductive analysis of luck, but a helpful elucidation of it, this is not a fatal problem. The modal account of luck, as it is stated by Pritchard, is already quite vague, but still useful. I would like to thank an anonymous reviewer of *Synthese* for making me consider this issue.

(i) in most nearby possible worlds, and in all of the very closest possible worlds, where S believes in a proposition that belongs to Q via the same method M that S uses in the actual world, S's belief is true.

The globalized version of the safety condition can be motivated by reflecting on the nature of luck. Therefore, globalizing the safety condition to a set of propositions is not ad hoc. Moreover, the motivation for globalizing the safety condition to a set of propositions does not stem only from considerations that have to do with epistemic luck.[5] Knowledge of propositions with singular content requires that the safety condition must be globalized to a set of propositions (Gendler and Hawthorne 2005, pp. 333–334; Hawthorne 2004, p. 56). Rabinowitz, for example, writes that:

> Knowledge of propositions with singular content requires safety to be formulated in a globally reliable way. Consider the case in which Jones, looking at a real barn surrounded by fake barns, forms the true belief that "*that* is a barn." The intuition is to deny Jones knowledge despite the fact that there is no close world in which that very barn is not a barn (assuming that a barn is essentially a barn). Since Jones could easily have falsely believed of a fake barn that "*that* is a barn," which expresses a different and false proposition, Jones is denied knowledge. (Rabinowitz 2018)

The intuition that the subject in the fake barns case lacks knowledge does not stem from the fact that the belief he actually formed could easily have been false (since it could not easily have been false given its content), but rather from the fact that the subject could easily have looked at another barn-like structure, and formed a false, albeit different, belief.[6] Finally, the globalization of the safety condition can be motivated by considering non-epistemic cases. If two assassins are after me and one of them is caught, I am not properly safe from being stabbed to death. For me to be properly safe both of the assassins need to be caught.[7]

The globalized version of the safety condition has several advantages over its simpler predecessor, the chief of them being that it is not trivially satisfied if the subject believes in a necessary truth. However, the globalization of the safety condition raises problems of its own. Most importantly, it generates a new kind of generality problem because the extension of safe beliefs will vary greatly depending on how the set of propositions Q is restricted. If the set of relevant propositions is very large safe beliefs will be hard to come by, whereas if the set is very small the condition might be too easy to satisfy.[8]

[5] See Williamson (2000, p. 101) for an argument why the safety condition needs to be of the globalized kind.

[6] I would like to thank an anonymous reviewer of *Synthese* for highlighting this fact.

[7] An anonymous reviewer of *Synthese* encouraged me to motivate the globalized safety condition in greater detail. This section was substantially improved as a result.

[8] According to Pritchard (2012, pp. 256–257), the relevant set of propositions is adequately restricted by the basis of belief-formation that the subject has in the actual world. This feature of his view makes it even more important to provide an answer to the generality problem as it inflicts the safety condition. Crucially, Pritchard does not provide an account of how to individuate bases of belief formation. Williamson (2009, p. 325) agrees that the relevant set of propositions cannot be adequately restricted solely in terms of the basis

I have argued elsewhere (2017, 2018a) that the set of propositions should be restricted in terms of the subject's subject matter of inquiry. In order for one's belief that p to be safe from error it must be the case that one could not easily have ended up with a false belief in one's inquiry whether p. Or alternatively: If S's belief that p is safe then there is a question Q to which p is a correct answer and S could not easily have formed a belief in a false answer to Q.

For our present purposes it does not matter whether the relevant set of propositions is restricted in terms of the method of belief formation, the closeness of propositions, or in terms of subject matters of inquiry. What is important to recognize at this juncture is that a properly formulated safety condition is globalized to a set of propositions. Now that we have a properly formulated safety condition at our disposal, it is time to examine whether it is in tension with the extended mind hypothesis.

3 Carter's argument

Carter argues that safety-based theories of knowledge run into trouble with the m-parity principle and with the e-parity principle. He starts by noting that the following pair of cases is structurally similar, yet the safety condition is satisfied only in one of them if the extended mind thesis is true.

FAKE BARNS: Barney is driving through the countryside and is identifying objects to amuse his son. Barney sees a barn ahead, points towards it, and utters "That's a barn." His corresponding belief is true and justified. Unbeknownst to Barney, he is driving through barn façade county, where almost every object that looks like a barn is in fact a barn façade, which he would not be able tell apart from the real thing.[9]

> JOKESTER: Otto consults his notebook to determine when his doctor's appointment was today, and finds the correct time, noon, written in the book. Unbeknownst to Otto, his notebook had been stolen by a jokester, who fudged with the times of Otto's other appointments that day, changing them all back an hour. The jokester, however, overlooked the doctor's appointment, leaving the original and correct time intact. (Carter, 2013, p. 4024)

The cases are taken to be structurally similar because Otto and Barney are in a similar situation; both are presented with convincing fakes (barn façades and fake memories) and one non-fake, and happen to form a belief on the basis of the non-fake. In a very clear sense both Otto and Barney suffer from environmental epistemic luck, in that the epistemic environment in which they happen to be in is epistemically inhospitable, but yet in their peculiar circumstances they manage to form a true belief by a fluke. Given

Footnote 8 continued
of belief formation that the subject has in the actual world. He claims that all of the relevant propositions have to be "close" to each other. For a critique of Williamson's proposal see Hirvelä (2017). Sosa (2015) advances also a globalized version of the safety condition, though he does not engage with the problem of how to restrict the relevant set of propositions in detail. For a critique of Sosa's formulation of the safety condition, see Hirvelä (2018b).

[9] This case appears originally in Goldman (1976), though he credits Carl Ginet for it.

that both FAKE BARNS and JOKESTER involve environmental epistemic luck, both cases should be cases of ignorance rather than knowledge. In fact, the vast majority of epistemologists consider FAKE BARNS to be a clear case of ignorance. Moreover, it seems that JOKESTER is also a case of ignorance. Carter argues, however, that the safety condition is satisfied in JOKESTER while it is not satisfied in FAKE BARNS. If this is so, then it would seem that safety-based theories of knowledge are committed to bioprejudice.

Carter's argument for this conclusion is the following. When evaluating whether a true belief that p is safe we need to check whether p is true in nearby possible worlds where the belief that p is formed via the same cognitive process that was used in the actual world to form the belief that p. In FAKE BARNS the cognitive process that Barney uses is "pointing to one of the barn-looking objects in the facade-littered countryside" (Carter 2013, p. 4204). Given that this is Barney's method, his belief is clearly unsafe. In most nearby possible worlds his belief will end up being false.

If the extended mind thesis is true, then the cognitive process that Otto uses in the actual world is partially determined by the notebook that he consults in the actual world, and therefore the notebook should be kept fixed in the relevant possible worlds. Crucially, this will involve keeping fixed the entries of the notebook, and this will entail that the notebook will contain the correct time for the doctor's appointment. But if that is the case, then Otto's belief about the time of the appointment will be true in all relevant possible worlds and hence his belief is safe! Given that JOKESTER is a clear case of ignorance, this is a bad thing for the safety condition. In order to deliver the correct verdict regarding the case the proponent of the safety condition needs to *exclude* the fact that the notebook contains the correct time for the doctor's appointment. If JOKESTER and FAKE BARNS are structurally similar, the proponent of the safety condition must reject the m-parity principle. After all, the safety condition will deliver the verdict of ignorance regarding both cases only if the notebook is not treated as part of the cognitive process that Otto uses in the actual world, since in all possible worlds where Otto forms a belief by consulting it about the doctor's appointment he will end up with a true belief. Therefore safety theorists are committed to metaphysical bioprejudice. The notebook is not a part of Otto's cognition.

Carter also argues that the safety condition is in tension with the e-parity principle. He does this by presenting an intracranial analogue of JOKESTER where knowledge is intuitively present:

> FORGETFULNESS: Otto* (without Alzheimer's) has a normally functioning biological memory, which he relies on to organize his world. Atypically for Otto*, he forgets the time of his other appointments today – believing they were earlier than they actually are – though he does remember that his doctor's appointment is at noon. (Carter, 2013, p. 4207)

FORGETFULNESS seems to be an intracranial analogue of JOKESTER. If that is true, then the cases should be epistemically on a par. Given, however, that in FORGETFULNESS Otto* forms his belief "*by consulting a clear memory of the time of the appointment*—he could not easily have been wrong" (Carter 2013, p. 4207). Therefore Otto* does satisfy the safety condition, which is problematic given that it is not satisfied in the extracranial analogue (at least if we exclude the notebook). Moreover,

it seems that these are intuitive verdicts. Intuitively, Otto does not have knowledge while Otto* has. In the next section I will argue that both of Carter's arguments fail. The m-parity principle is not violated by the safety condition since Otto could easily have formed *a* false belief by consulting his notebook. The e-parity principle is not violated since the cases are in fact not analogous. In JOKESTER Otto's memory is not a virtuous faculty, whereas in FORGETFULNESS it is.

4 No tension with m-parity principle

First of all, it should be noted that Carter's argument rests on an overly simplified version of the safety condition that is not globalized to a set of propositions. Similarly, he does not recognize the need to globalize the modal account of luck to a set of events (Carter 2013, p. 4205). These mistakes are of course understandable, but they are mistakes nonetheless. Crucially, Carter's entire argument for the claim that safety-inspired anti-luck epistemology is in tension with the m-parity principle rests on the fact that he does not have a globalized version of the modal account of luck and of the safety condition at his disposal.

It is quite easy to see that the safety condition, as we have formulated it, is not satisfied in JOKESTER, even if we keep the notebook fixed (as it is constituted in the actual world) in all relevant possible worlds. After all, Otto could very easily have formed a false belief by consulting his notebook. Therefore his true belief is not safe. True, the belief that he formed in the actual world could not have been false if we keep the notebook fixed, but this does not mean that he is safe from error, which is what the safety condition requires.

One might object that FAKE BARNS and JOKESTER are not analogous from an epistemic point of view since there is an epistemic difference between the cases, namely that the non-globalized version of the safety condition is satisfied in JOKESTER while it is not satisfied in FAKE BARNS, and that this fact alone suffices to show that the globalized safety condition is in tension with the m-parity principle. However, at this juncture it is important to recall that the local version of the safety condition is satisfied in FAKE BARNS since the proposition Barney believes in has singular content. Assuming that barns are essentially barns, the proposition Barney believes in the actual world could not easily have been false since it is constitutive of that singular proposition that it is about a real barn. The intuition that Barney lacks knowledge stems from the fact that he could very easily have pointed at a fake barn, and believed of it that [that is a barn].

Therefore, the local safety condition that Carter has in mind is, despite first impressions, satisfied in both FAKE BARNS and in JOKESTER. Thus, the above line of argument does not succeed in demonstrating that safety theories treat FAKE BARNS and JOKESTER differently. Moreover, the globalized safety condition is not satisfied in FAKE BARNS, since Barney could easily have formed another similar belief which would have been false. Given that the globalized version of the safety condition is not satisfied in FAKE BARNS, or in JOKESTER, it delivers the intuitive verdict that the subjects of these cases lack knowledge. Therefore, I conclude that Carter's argument for the tension of anti-luck epistemology and the m-parity principle is unsuccessful,

once we recognize that the safety condition must be globalized to a set of propositions in the first place.

5 No tension with the e-parity principle

Carter's argument for the tension between the safety condition and the e-parity principle hinges on the claim that FORGETFULNESS is an intracranial analogue of JOKESTER and that knowledge is present in the former but not in the latter. It is hardly surprising that it is intuitive to think that knowledge is present in FORGETFULNESS. After all, many hold that to remember that p is simply to know that p. According to knowledge-first epistemologists, for example, remembering that p is a factive mental state and given that knowledge is the most general factive mental state remembering that p is just to know that p (Williamson 2000). One might argue then that the way in which FORGETFULNESS is spelled out forces us to conceive it as a case of knowledge, since this is already said in the case description. This feature of the case is unfair given the dialectic situation, and therefore the case should be described in a more neutral manner. Perhaps we could substitute "does remember that" with "does have a true belief that". If the case was reformulated along the suggested lines it would cease to be an intuitive case of knowledge, or at least the intuition of knowledge would diminish, since it would no longer be stipulated in the case description that Otto* knows when the appointment is. Of course, if Otto* does not know when the appointment is, the safety condition will not treat the pair of analogous cases differently, and hence will not violate the e-parity principle.

This response is, however, somewhat disappointing, since we have not engaged with the original formulation of FORGETFULNESS. Crucially, the safety theorist does have the means to deal with the original formulation of the case as well, though by offering principled reasons for thinking that FORGETFULNESS is not an intracranial analogue of JOKESTER.

A crucial difference between the cases is that in FORGETFULNESS Otto*'s belief is formed through a cognitive virtue, whereas in JOKESTER Otto's belief is formed via a non-virtuous belief-forming process. This explains why the cases are dis-analogous. Moreover, safety theorists can tap into this feature, because they can maintain that the safety condition should be relativized to the virtuous methods of belief-formation that the subject uses in the actual world.

The safety condition can be restricted to virtuous methods of belief formation because knowledge is always gained through epistemic virtues or competences. This is something that virtue epistemologists, such as Sosa (2007, 2009, 2015), Zagzebski (1996), Greco (2010), Pritchard (2012), Miracchi (2015) and Carter (2016) readily accept. For example, according to Pritchard (2012, pp. 247–249), epistemic theorizing is guided by two master intuitions, the *anti-luck intuition* and the *ability intuition*. The former intuition dictates that knowledge is incompatible with epistemic luck, while the latter states that knowledge is always gained through the exercise of one's cognitive abilities. Assuming that knowledge is always gained through the exercise of epistemic competences, the safety theorist can relativize the safety condition to virtuous methods of belief formation without fear of focusing on a too narrow class of methods of belief

formation. I have argued elsewhere (2017) that by relativizing the safety condition to virtuous methods of belief-formation the safety-theorist is able to offer an elegant solution to the generality problem as it inflicts the safety condition. Note that by relativizing the safety condition to virtuous methods of belief formation used by the subject in the actual world, we do not subscribe to the more demanding idea, often endorsed by virtue epistemologists, according to which in cases of knowledge the truth of one's belief has to be creditable or attributable to one's cognitive virtues.

If the safety condition is to be relativized to virtuous methods of belief formation, we need an account of when a belief is virtuously formed. Following Sosa (1991, p. 284), epistemic virtues, or competences, can be understood as stable dispositions seated in the subject to acquire or maintain true beliefs and avoid false beliefs within a certain field of propositions, while in certain environments and conditions. According to Sosa (2010, pp. 465, 467), dispositions have a three-part structure. They involve (i) constitution, (ii) condition and (iii) situation. The constitution of a perceptual competence includes rods and cones and the visual cortex, the condition includes being awake and sober, and the situation includes being in adequate lighting conditions. A disposition can be lost by undermining its constitution, condition or situation. For example, by manipulating a subject's visual cortex with magnetic pulses in order to cause temporary lesions one will destroy the constitution of the visual competence.

The field of propositions consists of propositions in which the relevant virtue can produce beliefs in. In the case of an olfactory virtue the field of propositions will consists of propositions such as [this smells like lilac and gooseberries]. The environment and conditions specify in what kind of environment and conditions one must be disposed to attain true beliefs and avoid false beliefs. The fact that I am disposed to form false beliefs about the colour of objects at night or after having ingested powerful hallucinogens does not entail that my vision would not be an epistemic virtue while in conditions that are suitable for the use of vision. External conditions can either prohibit or enable the exercise of epistemic virtues.

Finally, a feature that separates epistemic virtues from merely reliable dispositions is that epistemic virtues have to be integrated into one's cognitive character (Pritchard 2012, p. 262). A recently developed brain lesion that causes one to believe that one has a brain lesion will not count as an epistemic virtue, even though it will dispose one to believe what is true. In order for that brain lesion to count as a virtue the subject would have to integrate it into her cognitive character. Perhaps, if one went to a doctor who explained that one suffers from a rare brain lesion which causes one to form the belief that one has a brain lesion, the brain lesion would be integrated into one's cognitive character. However, without such integration the reliably true beliefs caused by the brain lesion would not count as knowledge. In fact, many authors have argued on the basis of brain lesion-type cases for the insufficiency of process reliabilism, which does not require that reliable processes be integrated into one's cognitive character in order for them to be knowledge conducive (Bonjour 1980; Greco 2010; Lehrer 1990; Palermos 2014; Plantinga 1993). By claiming that the knowledge relevant dispositions have to be integrated into the cognitive character of the subject, virtue epistemologists are able to deal with such cases.

How strongly cognitive abilities have to be integrated into one's cognitive character depends in part on their etiology. Dispositions that are acquired through natural devel-

opment (or that are otherwise innate) do not need to be consciously integrated, while dispositions that are acquired later in life, might have to be integrated through conscious endorsement of their truth conduciveness (Pritchard 2010). If someone implanted a chip into your brain without your knowledge, which caused you to form true beliefs about the results of the latest baseball games, then those beliefs would hardly qualify as knowledge. If, however, you were to come to know that someone implanted this annoying, but perfectly reliable chip into your brain, those beliefs would qualify as knowledge. In some cases, sub-conscious integration might be enough. For example, if your other senses constantly confirmed the outputs of your newly and unconsciously acquired belief-forming disposition, the belief-forming disposition would, at some point, be integrated into your cognitive character, and you would be rational to trust the deliverances of that disposition. In fact, our senses are interconnected in the kind of way that they constantly confirm the outputs of each other. You hear a sound of a car driving by and see it a split second later. You smell the exhaust fumes of the car and feel the water splash on your neck as the car drives through the puddle. What you taste is the bitterness of life, which is not directly related to the car, but is still, in part, caused by it. All of these sensations help to confirm that a car just passed by. This kind of minimal interconnectedness can suffice for integration if it occurs over a prolonged period of time.[10]

From these remarks we can derive when a belief is virtuously formed:

A subject S's belief that *p*, which belongs to a field of propositions F, is virtuously formed via method V, in circumstances C and environment E if and only if:

(i) S has an inner disposition D, which is integrated into S's cognitive character, to attain correct doxastic attitudes with respect to propositions that belong to F while in C/E,
(ii) S is in C/E,
(iii) the fact that S believes that *p*, via V, is due to exercising D.

It is vital to note that the idea that the subject's belief has to be a formed via exercising an inner disposition to attain correct doxastic attitudes in order to be virtuously formed, is not in tension with the extended mind thesis. After all, if the extended mind thesis is true, cognition can extend beyond the boundaries of skin and skull, and therefore one's 'inner' dispositions could have a physical basis that extends beyond one's biological body. Therefore, Otto's notebook in OTTO could be part of a stable inner disposition to attain and maintain true beliefs.

With these virtue-theoretic considerations in mind, it is quite easy to see that Otto's belief in JOKESTER is not virtuously formed. After all, someone has tampered with his notebook. It is as if he had been brainwashed. The constitution of his external memory has been undermined. This does not seem to be the case in FORGETFULNESS, however. After all, even those with excellent memory forget things from time to time and this does not undermine the fact that their memory constitutes an epistemic virtue. Therefore, while Otto*'s belief is plausibly thought to be virtuously formed, Otto's is not. If the safety condition is relativized to virtuous methods of belief formation, as I

[10] Palermos (2014, p. 1934) argues that this kind of unreflective integration allows us to trust the deliverances of our cognitive abilities provided that we lack any reasons for negating our beliefs and that we are motivated to believe what is true.

have argued, then Otto*'s belief can be safe, while Otto's belief is not even a candidate for a safe belief.

Once the details of the cases are brought to light, it is reasonable to think that JOKESTER and FORGETFULNESS are not analogous after all. An intracranial analogue of JOKESTER would be a case in which a prankster has deliberately altered the memories of Otto*. Crucially, if Otto* has been brainwashed in such a way, the intuition that he has knowledge vanishes, since his beliefs would not be the products of an epistemic competence because the constitution of that competence would have been undermined by the prankster's actions.

Once the safety condition is properly understood and relativized to virtuous methods of belief formation it is not in tension with the parity principles. Carter anticipates that this problem could be dissolved along the above lines, since he writes that:

> Though anti-luck epistemology seems to get the right result across a spectrum of cases, we need a more precise account of what to hold fixed under the description of the relevant way the belief was formed in the actual world, when moving out to nearby worlds. Pritchard (2007) himself has described the account on offer as vague on this point. Anti-luck epistemologists need to do better, and when they do, perhaps this will help deal with cases of extended cognition—cases for which the matter of what precisely to hold fixed is of special importance. (Carter, 2013, p. 4212)

Now that we have a satisfactory solution to the problems raised by Carter it is time to consider Kelp's argument, which aims to establish the conclusion that the safety condition can be a necessary condition for knowledge only if the extended mind thesis is true.

6 A hostage situation?

Kelp has argued that the safety condition can be a necessary condition for knowledge only if the extended mind thesis is true. Kelp's argument rests on the following thought experiment:

> TIMEKEEPER: The timeseeker looks at a public clock, sees that it reads 2.30 and on that basis comes to believe that it is 2.30. The clock has an outstanding track-record of functioning properly and the timeseeker has no reason to think that it is currently not accurate. Her belief is true. It is in fact 2.30. Unbeknowst to the timeseeker, however, the clock has stopped exactly twelve hours ago. As it happens, this episode is observed by the timekeeper, who has been called in to fix the stopped clock. Using his two radio clocks, the timekeeper confirms that the reading of the stopped clock is accurate. Had the stopped clock reading been inaccurate, the timekeeper would have alerted the timeseeker to this fact. (Kelp, 2014, p. 236)

Kelp intuits that the timeseeker knows that it is 2.30. He takes this thought experiment to be a counterexample against virtue epistemological theories of knowledge, since the subject's cognitive abilities are not manifested in the truth of her belief. He also considers it as a counterexample against sensitivity and safety-based theories of knowledge.[11] There are plenty of nearby possible worlds where the timeseeker believes that it is 2.30, while her belief is false, since she looks at the stopped clock a bit earlier. There is no reason to suppose that the involvement of the timekeeper is modally robust, in that she would have told the timeseeker that the clock has stopped in all nearby possible worlds. In fact, Kelp (2014, p. 238) thinks that we can stipulate that the involvement of the timekeeper is modally fragile, in that she is not guarding the timeseeker from error in most nearby possible worlds. But if that is so, then the sentence "Had the stopped clock reading been inaccurate, the timekeeper would have alerted the timeseeker to this fact" is clearly false. After all, if the involvement of the timekeeper is modally fragile, she would not have alerted the timeseeker to the fact that the clock's reading is inaccurate, since in the vast majority of possible worlds, where the reading is inaccurate, the timekeeper is not alerting the timeseeker to this fact. If the involvement of the timekeeper is indeed a modally fragile feature of the case, then I have to acknowledge that I lack the intuition that the subject lacks knowledge. But let us put my intuitions aside for now.

Kelp (2014, p. 246) claims that virtue epistemological-, safety- and sensitivity-based theories of knowledge are able to deliver the verdict that the timeseeker knows only if the case is construed as a case of extended cognition. Kelp (2014, p. 244) maintains that the timekeeper should be understood as a monitoring process, which extends beyond the timeseeker's body. It is easy to see why the safety condition would be satisfied by the timeseeker if the timekeeper was part of the cognitive process that is to be kept fixed in relevant possible worlds. After all, in all of the possible worlds where the timekeeper is there to alert the subject of the fact that the clock is displaying the wrong time the timeseeker will not form the relevant belief, and hence her belief that it is 2.30 is true in all relevant possible worlds where she holds the belief. But if this is the only way for the safety theorist to deal with this case, then the safety condition is hostage to the possible truth of the extended mind thesis. Given that the extended mind thesis is a controversial thesis, and far from an obvious truth, this commitment is hardly welcome. Worse, it seems that the cognitive extension in TIMEKEEPER occurs all too easily. After all, the timeseeker is not even aware of the timekeeper's presence in the actual world! He just happened to walk by. It does not seem to be the case that the timekeeper is integrated (in anyway) to the timeseeker's cognitive character.

Most proponents of the extended mind thesis would agree that TIMEKEEPER is not a case of cognitive extension, since they hold that cognitive extension can occur only if the so-called *trust and glue* conditions are satisfied. These conditions state that cognitive extension can occur only if:

(i) the resource is reliably available and typically invoked,

[11] According to the sensitivity condition a subject S's belief that p is sensitive just in case if it were the case that not-p S would not believe that p.

(ii) any information retrieved or gained via it should be more-or-less automatically endorsed, and
(iii) the information contained should be easily accessible when required. (Clark, 2010, p. 46)[12]

Kelp thinks that cognitive extension can occur without there being a reliable coupling, as long as the extended system achieves functional integrity for the short period that it lasts. He relies on Wilson and Clark (2009, p. 65), according to whom cognitive extensions might be short lived and fleeting. If cognitive extension can indeed occur as easily as suggested by Wilson and Clark, then it seems that one's cognition can extend without it being the case that the physical basis of one's cognitive character extends, since such short-lived couplings are plausibly not integrated into one's cognitive character. Indeed, Clark (2015) has argued, contra Pritchard (2010) and Palermos (2014), that the extended mind thesis does not fit snuggly with virtue epistemological accounts:

> As far as that argument goes, it should make no difference at all whether or not Otto is now, or ever was, aware of the source of the reliability of the notebook involving process. Indeed—and here comes the promised dilemma—there is a very real sense in which the more he is aware of such matters, the less the notebook will seem to be playing the same kind of functional role as biological memory. For as we noted, our biological memory is not typically subject to agentive scrutiny as a process at all, much less as one that may or may not be reasonably judged to be reliable by the agent. (Clark, 2015, p. 3763)

However, as we noted earlier, cognitive integration need not always involve conscious awareness of the reliability of the process, unreflective integration is also possible, and Clark (2015, p. 3773) acknowledges this. However, Clark (2015, p. 3754) thinks that no kind of cognitive integration is necessary, since he holds that an implant which caused beliefs about the ambient temperature, which is installed without the agent's knowledge, would generate beliefs that amounted to knowledge from the very moment it delivered its first output. Clark is, however, alone with his intuitions on this score, since the vast majority of epistemologists think that brain lesion-type cases are cases of ignorance.

Moreover, if we abandon the trust and glue conditions, we risk incurring *cognitive bloat*. To weaken them would result in counting processes that are genuinely non-cognitive as cognitive and to an unwelcome explosion of dispositional beliefs (Clark 2008, p. 80). With the trust and glue conditions in place, a downloaded book in your dropbox would not count as an extension of your cognition but Otto's notebook would. If the extended cognition thesis leads to overextending our cognition we have a good reason to reject the thesis. Therefore, we should hold onto the trust and glue conditions.[13] The fact that we need to abandon the trust and glue conditions in order to

[12] Clark and Chalmers (1998) also offer a fourth condition, according to which the information in the notebook would have to have been consciously endorsed by Otto in the past, but suggest that this condition might be too stringent.

[13] In fact, many have argued that the trust and glue conditions are too weak, and fail to specify sufficient conditions for when cognition extends. See Farkas (2012, pp. 444–4445) and Wikforss (2014, p. 475). For

conceive TIMEKEEPER as a case of extended knowledge is a good reason to think that it is not a case of cognitive extension.

Finally, Kelp seems to recognize that the intuition of knowledge that he hopes to elicit is rather weak, and maintains that one can alter the case in such a manner that the involvement of the timekeeper is modally robust in order to strengthen the intuition that the timeseeker knows (2014, p. 249). He claims that the case will still be a counterexample to virtue epistemological theories of knowledge that accept the idea that in cases of knowledge one's cognitive abilities have to be manifested in the truth of one's belief. In my view this does in fact strengthen the intuition that the timeseeker acquires knowledge. After all, if the involvement of the timekeeper is modally robust, then the clock is either displaying the correct time, or the timekeeper is alerting the timeseeker to the fact that the clock is displaying the incorrect time. Crucially, however, the reformulated version of the case no longer serves as a counterexample against the sensitivity or safety conditions, which Kelp fails to mention.

In essence, Kelp is playing with two sets of cards. The case can be a counterexample against sensitivity and safety conditions only if the involvement of the timekeeper is a modally fragile feature of the case, in which case the intuition that the subject knows diminishes, while it cannot be a counter example against the necessity of these conditions if the involvement of the timekeeper is modally robust, in which case the intuition that the subject knows is more robust. But if this really is so, then the evidence actually supports the necessity of the modal conditions, since the intuition of knowledge varies with the modal features of the case.

7 Concluding remarks

To conclude, the case that Kelp uses in order to argue that the safety and sensitivity conditions can be necessary conditions for knowledge only if the extended mind thesis is true does not succeed. Plausibly, the case does not feature cognitive extension, since the timekeeper really is external to the timeseeker's cognitive character, and hence does not even partially constitute the timeseeker's cognitive abilities. Therefore, safety theorists have no reason to think that the timekeeper should be held fixed in all relevant possible worlds, and can deliver (in my mind) the intuitive verdict that the timeseeker does not acquire knowledge, given that the timekeeper is a modally fragile feature of the case. If, however, the involvement of the timekeeper is a modally robust feature of the case, then safety and sensitivity conditions have no problem with delivering the verdict of knowledge in TIMEKEEPER. The fact that the intuition of knowledge varies with the modal robustness of the timekeeper only speaks in favour of the safety and sensitivity conditions. Given that the safety condition delivers the correct verdict regarding the case irrespective of whether the extended mind thesis is true, it is not hostage to the possible truth of the extended mind thesis. Moreover, I argued earlier that Carter's arguments fail to create tension between the safety condition and the parity principles, once the safety condition is properly understood. Therefore, we have found

Footnote 13 continued
a critical assessment and discussion of attempts to confine cognition that do not resort to the trust and glue conditions, see Allen-Hermanson (2013).

no reasons for thinking that there is any tension between the extended mind thesis and the safety condition. This is a welcome conclusion both to proponents of the extended mind thesis as well as to the safety theorists.

Acknowledgements I would like to that Duncan Pritchard, Adam Sanders, Pii Telakivi and an anonymous referee at *Synthese* for insightful comments that helped to improve this paper.

References

Allen-Hermanson, S. (2013). Superdupersizing the mind: Extended cognition and the persistence of cognitive bloat. *Philosophical Studies, 164,* 791–806.
Bonjour, L. (1980). Externalist theories of empirical knowledge. *Midwest Studies in Philosophy, 5,* 53–73.
Carter, J. A. (2013). Extended cognition and epistemic luck. *Synthese, 190,* 4201–4214. https://doi.org/10.1007/s11229-013-0267-3.
Carter, J. A. (2016). Robust virtue epistemology as anti-luck epistemology: A new solution. *Pacific Philosophical Quarterly, 97*(1), 140–155. https://doi.org/10.1111/papq.12040.
Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford: Oxford University Press.
Clark, A. (2010). Memento's revenge: The extended mind, extended. In R. Menary (Ed.), *The extended mind* (pp. 43–66). Cambridge: MIT Press.
Clark, A. (2015). What the 'Extended Me' knows. *Synthese, 192,* 3757–3775.
Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis, 58*(1), 7–19.
Coffman, E. J. (2007). Thinking about luck. *Synthese, 158*(3), 385–398. https://doi.org/10.1007/s11229-006-9046-8.
Farkas, K. (2012). Two versions of the extended mind thesis. *Philosophia, 40,* 435–447.
Gendler, T., & Hawthorne, J. (2005). The real guide to fake barns: A catalogue of gifts for your epistemic enemies. *Philosophical Studies, 123*(3), 331–352.
Goldman, A. (1976). Discrimination and perceptual knowledge. *Journal of Philosophy, 73,* 771–791.
Greco, J. (2010). *Achieving knowledge: A virtue theoretic account of epistemic normativity*. Cambridge: Cambridge University Press.
Greco, J. (2012). A (different) virtue epistemology. *Philosophy and Phenomenological Research, 85*(1), 1–26. https://doi.org/10.1111/j.1933-1592.2011.00567.x.
Hawthorne, J. (2004). *Knowledge and lotteries*. New York: Oxford University Press.
Hirvelä, J. (2017). Global safety: How to deal with necessary truths. *Synthese*. https://doi.org/10.1007/s11229-017-1511-z.
Hirvelä, J. (2018a). On virtue, credit and safety. *Grazer Philosophische Studien, 95*(1), 98–120.
Hirvelä, J. (2018b). No safe haven for the virtuous. *Episteme*. https://doi.org/10.1017/epi.2018.15.
Kelp, C. (2014). Epistemology extended. *Philosophical Issues, 24*(1), 230–252. https://doi.org/10.1111/phis.12032.
Kim, J. (1976). Events as property exemplifications. In M. Brand & D. Walton (Eds.), *Action theory: Proceedings of the Winnipeg conference on human action* (pp. 159–177). Dordecht: Springer.
Lehrer, K. (1990). *Theory of knowledge*. Boulder, CO: Westview.
Miracchi, L. (2015). Competence to know. *Philosophical Studies, 172,* 29–56.
Palermos, S. O. (2014). Knowledge and cognitive integration. *Synthese, 191*(8), 1931–1951. https://doi.org/10.1007/s11229-013-0383-0.
Plantinga, A. (1993). *Warrant: The current debate*. Oxford: Oxford University Press.
Pritchard, D. (2005). *Epistemic luck*. Oxford: Oxford University Press.
Pritchard, D. (2007). Anti-luck epistemology. *Synthese, 158*(3), 277–297. https://doi.org/10.1007/s11229-006-9039-7.
Pritchard, D. (2010). Cognitive ability and the extended cognition thesis. *Synthese, 175,* 133–151. https://doi.org/10.1007/s11229-010-9738-y.
Pritchard, D. (2012). Anti-luck virtue epistemology. *Journal of Philosophy, 109*(3), 247–279.
Pritchard, D. (2015). The modal account of luck. In D. Pritchard & L. Whittington (Eds.), *The philosophy of luck*. Hoboken, NJ: Wiley-Blackwell.

Rabinowitz, D. (2018, May 22). "The Safety Condition for Knowledge", *The Internet Encyclopedia of Philosophy*. Retrieved from http://www.iep.utm.edu/.
Sosa, E. (1991). *Knowledge in perspective—Selected essays in epistemology*. Cambridge: Cambridge University Press.
Sosa, E. (2007). *A virtue epistemology: Apt belief and reflective knowledge* (Vol. I). Oxford: Oxford University Press.
Sosa, E. (2009). *Reflective knowledge: Apt belief and reflective knowledge* (Vol. II). Oxford: Oxford University Press.
Sosa, E. (2010). How competence matters in epistemology. *Philosophical Perspectives, 24*(1), 465–475. https://doi.org/10.1111/j.1520-8583.2010.00200.x.
Sosa, E. (2015). *Judgment and agency*. Oxford: Oxford University Press.
Vaesen, K. (2011). Knowledge without credit, exhibit 4: Extended cognition. *Synthese, 181*(3), 515–529. https://doi.org/10.1007/s11229-010-9744-0.
Wikforss, Å. (2014). Extended belief and extended knowledge. *Philosophical Issues, 24*(1), 460–481. https://doi.org/10.1111/phis.12043.
Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.
Williamson, T. (2009). Reply to John Hawthorne and Maria Lasonen-Aarnio. In P. Greenough & D. Pritchard (Eds.), *Williamson on knowledge* (pp. 313–329). Oxford: Oxford University Press.
Wilson, R., & Clark, A. (2009). How to situate cognition: Letting nature take its course. In M. Aydede & P. Robbis (Eds.), *The Cambridge handbook of situated cognition*. Cambridge: Cambridge University Press.
Zagzebski, L. (1996). *Virtues of the mind: An inquiry into the nature of virtue and the ethical foundations of knowledge*. Cambridge: Cambridge University Press.