What "Extended Me" knows

What 'Extended Me' Knows

Andy Clark

School of Philosophy, Psychology and Language Sciences
Dugald Stewart Building,
3 Charles St,
Edinburgh,
EH8 9AD
Scotland
UK

andy.clark@ed.ac.uk

Tel: +44 131 650 3659
Fax: +44 131 650 3660

Abstract

Arguments for the 'extended mind' seem to suggest the possibility of 'extended knowers' – agents whose specifically epistemic virtues may depend on systems whose boundaries are not those of the brain or the biological organism. Recent discussions of this possibility invoke insights from virtue epistemology, according to which knowledge is the result of the application of some kind of cognitive skill or ability on the part of the agent. In this paper, I argue that there is a fundamental tension in these appeals to cognitive virtue. The tension centers on the presence of a tool or technology as an object of awareness, hence something apt for epistemically virtuous engagement on the part of the agent. I highlight a dilemma: the better something looks as a non-biological element of the machinery of mind, the worse it looks as a potential object of any specifically epistemic skill or ability on the part of the agent. The tension is resolved, I argue, by thinking about sub-personal forms of epistemic hygiene. I examine one such form (rooted in the vision of the 'predictive brain'), and show how it sits neatly with the vision of the extended mind. I end by asking what we can still reasonably expect, given this more complex sub-personal story, by way of agent-level cognitive hygiene.

Keywords: Extended Mind, Extended Knowledge, Virtue Epistemology, Reliabilism

What 'Extended Me' Knows

Andy Clark

O. Abstract/Introduction

Arguments for the 'extended mind' seem to suggest the possibility of 'extended knowers' – agents whose specifically epistemic virtues may depend on systems whose boundaries are not those of the brain or the biological organism. Recent discussions of this possibility invoke insights from virtue epistemology, according to which knowledge is the result of the application of some kind of cognitive skill or ability on the part of the agent. In this paper, I argue that there is a fundamental tension in these appeals to cognitive virtue. The tension centres on the presence of a tool or technology as an object of awareness, hence something apt for epistemically virtuous engagement on the part of the agent. I highlight a dilemma: the better something looks as a non-biological element of the machinery of mind, the worse it looks as a potential object of any specifically epistemic skill or ability on the part of the agent. The tension is resolved, I argue, by thinking about sub-personal forms of epistemic hygiene. I examine one such form (rooted in the vision of the 'predictive brain'), and show how it sits neatly with the vision of the extended mind. I end by asking what we can still reasonably expect, given this more complex sub-personal story, by way of agent-level cognitive hygiene.

1. Extended Knowers: The Story So Far

Proponents of the 'extended mind' (to use the term introduced by Clark and Chalmers (1998)) claim that even quite familiar human mental states, such as states of believing, can be realized, in part, by structures and processes located outside the human organism. It is important to recognize, at the outset, that such claims go far beyond the important but less challenging assertion that human cognizing often leans heavily on various forms of external scaffolding and support. Instead, extended mind theory suggests that the physical machinery that realizes some of an individual agent's cognitive processes and mental states can, under humanly attainable conditions, include elements and devices located beyond the bounds of skin and skull. I shall not attempt to rehearse the arguments for this claim here. Suffice to say that the claim has been widely defended and widely attacked (for a nice sampling, see the essays in Menary (2010), and for a sustained presentation and defense, see Clark (2008)).

For the purposes of the present essay, I shall simply assume the claim to be correct.

There is one aspect of the original argumentation that is, however, crucial to thinking about possible implications for epistemology. That aspect concerns the agent's own attitude, if any, to the physical machinery (neural circuits, iPhones, notebooks) in question. It is here that we locate a partial mismatch between the way the original extended mind story and the way that story has been depicted in some of recent the literature concerning extended knowledge. Thus Pritchard (2010), in his careful and groundbreaking treatment, begins with what he dubs the 'ability intuition' according to which:

> "A true belief, no matter what else of epistemic relevance can be offered in its favour (e.g., that it is safe, sensitive, backed by reasons, epistemically blameless, and so on), will not count as a case of knowledge if it is not the product of cognitive ability." (Pritchard (2010) p.134).

The notion of cognitive ability is designed to overcome some of the apparent limitations of basic 'reliabilist' accounts according to which an agent has knowledge when her beliefs are the products of reliable belief-forming processes (Goldman (1986)). Such accounts fall short, the ability theorist claims, because they do not ensure that the truth of the beliefs is due to the 'cognitive efforts' (Pritchard (op cit) p.140) of the agent herself, and hence can yield counter-intuitive conclusions. Pritchard gives the example of Temp who forms temperature-beliefs by consulting a thermometer in the room. But in fact, the thermometer is faulty. Luckily, though, a hidden agent adjusts the room temperature to match that of the malfunctioning thermometer. Pritchard concludes that:

> "Clearly…Temp does not know the temperature of the room, and the reason for this is that his reliability does not reflect his cognitive ability at all, but merely the helpful assistance of the hidden helper". (Pritchard (2010) p. 136).

Nor, Pritchard notes, does it seem to matter whether the source of such deviant reliability is onboard or (as in the helper case) external. The more general moral is that the reliability of my beliefs (their capacity to track the truth) should not be disconnected from my own cognitive character and agency. For when such disconnects occur (and Prichard gives several many examples, both internally and externally supported, in the paper) it seems as if

my successful believing (my believing what turns out to be true) is not properly creditable to me. This is the crux of the ability intuition.

One reason why we might endorse such an intuition is that it helps make sense of the point of ascribing knowledge (rather than merely true belief) to an individual agent. To credit an agent with knowledge is, plausibly, to credit them with a kind of reliable skill for tracking the truth in certain kinds of situation. Thus Greco (2012) notes that:

> Regarding the value of knowledge, we may note that, in general, we value success from ability over mere lucky success. Our preference for knowledge over mere true belief may now be understood as an instance of this more general valuing. Greco (2012) p. 2)

Thus I may admire someone who hits the bull's-eye repeatedly in dart games, but I would retract my admiration were I to learn that they were using an intelligent dart that could find any target on the dartboard. Under such circumstances, the agent does not possess the skill I thought she did (though she may possess many other skills, including many that involve the use of the intelligent dart). Valuing the agent's abilities in this way is, Greco notes, valuing the ability itself rather than merely valuing its consequences (in our example, consequences such as winning the game). It is valuing an aspect of her 'person-level excellence' (Greco (op cit)). In the case of agents that make claims about the how the world is, we may very reasonably care about finding agents whose abilities include abilities to track the truth – agents who believe the true because they are good at tracking the truth, rather than for some other reason.

Temp, of course, looks as if he is good at tracking temperature truths. But in fact, the temperature truths are tracking him. This should reduce my assessment of his temperature-tracking excellence. In addition, it makes his performance vulnerable in ways that I might not otherwise expect (for example, if the helper falls sick, or decides to stop helping him out). This kind of hidden vulnerability adds an important instrumental dimension, or so it seems to me, even to the 'person-level excellence' model that Greco prefers.

Taking all this onboard, I agree that there is some motivation for linking knowledge to the exercise of abilities that are properly credited to the agent herself (assuming ecologically normal contexts). Thus linked, ascriptions of knowledge are like other ascriptions of skill and excellence – they tell us something important about the agent, and (hence) about the normal range of circumstances in which we can rely on that agent for specific purposes.

But what does it mean to ascribe an ability *to the agent*? It is at this point that the debates surrounding the extended mind seem potentially relevant. For arguments for the extended mind thesis are meant to cast doubt on the idea that simple markers like the boundaries of skin and skull are reliable indicators of the boundaries of the underlying machinery of an individual mind. Whatever cognitive skills and capacities I have due to the underlying machinery of my own mind are surely - by definition - skills and capacities that are properly assigned *to me*. Thus there ought to be a direct route from extended realization bases for individual minds to extended realization bases for whatever skills and abilities are implicated in ascriptions of knowledge. Extended minds breed extended knowers. Or do they?

2. The Dilemma

Pritchard (2010) distinguishes two forms of the appeal to cognitive ability:

> (COGAWEAK) If S knows that p, then S's true belief that p is the product of a reliable belief-forming process which is appropriately integrated within S's cognitive character such that her cognitive success is to a significant degree creditable to her cognitive agency.
>
> (COGASTRONG) S knows that p iff S's true belief that p is the product of a reliable belief-forming process which is appropriately integrated within S's cognitive character such that her cognitive success is primarily creditable to her cognitive agency.

Notice that in the first (weak) form, the requirement is that the agent's cognitive success be credited to her cognitive agency "to a significant degree", whereas in the second, it must be creditable "primarily" to her cognitive agency. In each case, there is in addition an appeal (in the form of a necessary condition) to the integration of the belief-forming process within the agent's 'cognitive character'. This is meant to help cash the intuition that the ability is properly credited to the agent (rather than, in our non-cognitive example, to the intelligent dart). In the dart case, the requirement would thus be that the bull's-eye hitting skill be either significantly or primarily creditable to her athletic (e.g. hand-eye co-ordination) ability. This is meant to explain why Temp lacks knowledge of the ambient temperature. For as noted above, Temp plausibly lacks the skill we thought he possessed. But the strong and weak versions differ over what is required to bring Temp up to scratch.

To improve Temp's situation, imagine (though this is not a thought-experiment that Pritchard himself conducts) that he begins to spot situations in which his temperature-beliefs become unreliable. Imagine that one day a week (every Wednesday, say) the helper takes a break, and that Temp consciously notices that on wednesdays the projects that he and his friends undertake using his temperature beliefs tend to fail. On Wednesdays, thenceforth, Temp does not trust the thermometer at all. From that point on, Temp seems to meet the COGAweak condition.

To meet the more stringent COGAstrong condition, Temp could become aware (here I follow Pritchard's text directly) of the entire original (24/7 helper) underlying scenario. Now the same thermometer- consulting routine looks, even on the COGAstrong reading, apt to yield knowledge. This, Pritchard suggests, is because:

> "Temp is now able to take cognitive responsibility for this cognitive success. In becoming aware of the relevant facts… this hitherto merely reliable belief-forming process becomes integrated within Temp's cognitive character, such that his cognitive success is now primarily creditable to his cognitive agency. In this way, mere reliability in his belief-forming process is converted into the exercise of a genuine cognitive ability, one that can in principle deliver knowledge". Pritchard (2010) p.138 (my emphasis)

In this passage, I have emphasized the phrase 'becoming aware of the relevant facts', as this (as we'll see) is crucial for the issues concerning personal-level involvement that I wish to stress.

Pritchard himself (after examining a 'just slightly weakened' version of COGA strong that need not detain us) opts for COGAweak. COGAweak is sufficiently weak, Pritchard notes, to allow cases where success depends on a good deal of environmental luck to count as cases of knowing. In particular, it is weak enough to accommodate the case of Jenny. Jenny gets off a train in a new town, and asks the first person she meets for directions. The person is (as it happens) reliable and she thus acquires (Pritchard's intuitions, and COGAweak, allow) knowledge of how to reach her destination. To press this intuition, Pritchard fills in a few details, noting that:

> "we are assuming here that Jenny is in an epistemically friendly environment—it is not as if, for example, this town is renowned for its dishonest informants. Moreover, I take it that we are also reading into

> the case that Jenny is suitably responsive to epistemically relevant factors—it is not as if, for example, she would ask someone who would clearly not be a good informant (e.g., someone who was clearly a tourist), and it is not as if she would believe whatever she was told, even when it was obviously false" (Pritchard (op cit) p. 141)

The idea, then, is to allow for an interplay between cognitive ability and the presence of a friendly environment. That seems correct, for all the reasons that Pritchard (and others – see e.g. Palermos (2013)) suggest. To insist on more would be to deprive ourselves of large amounts of prima facie knowledge – all cases, in fact, where that kind of trust plays a significant role. COGAweak is thus meant to deliver a weak requirement upon cognitive agency, but one that nonetheless is claimed to go beyond anything that mere reliabilism could demand (op cit p.138).

It is this weakened version of the ability demand that, Pritchard then argues, turns out to fit nicely with the thesis of extended cognition. To see this, we are asked to consider a staple of the extended mind literature, Otto. Here's how Clark and Chalmers describe the case:

> "Otto suffers from Alzheimer's disease, and like many Alzheimer's patients, he relies on information in the environment to help 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, his notebook plays the role usually played by a biological memory." (Clark and Chalmers (1998) p.12

Here's how Pritchard describes the same case:

> "Otto suffers from Alzheimer's disease, and as a consequence he is gradually becoming aware that his memory is fading. In order to counter this, he starts carrying with him a notebook in which he records the kind of information that he requires on a day-to-day basis." Pritchard (2010) p.144

Notice the difference. Pritchard here makes Otto active and aware. Otto is 'gradually becoming aware that his memory is fading' and adopts the notebook strategy 'in order to counter this'. These glosses are not unreasonable. They probably reflect the general tenor of real-world cases better than the original! But they are adding something that Clark and Chalmers rather deliberately left

out. They are making the notebook strategy an object of Otto's mental focus. The reason we left that out is, of course, because it works subtly but surely against the extended mind claim itself. For this is not the role played by ordinary biological memory. Ordinary biological memory, for the most part, functions in a kind of automatic, subterranean way. It is not an object for us, we do not encounter it perceptually. Rather, it helps constitute us as the cognitive beings we are.

This is not to say that biological memory can never turn up as such an object. Bio-feedback devices sometimes make our inner activity into an object of our own attention. But on the whole, that which forms part of the agent is not encountered and processed like other stuff in the world. Rather, it provides the means and mechanisms for our encounters with the wider world. This, as Heidegger famously noted, is true of the body too. Our hands are not normally objects of our attention, though they may become so when they malfunction of fail to perform. It is thus plausible that real-world Otto would indeed go through the aware and reflective stage that Pritchard describes. But going through such a stage is not essential to the extended mind argument, and remaining locked in such a stage (never automating the procedures of information-deposit and access via the notebook) would be inimical to it for the reasons just described.

By contrast, those 'active' stages (awareness of the problem, and deciding to use the notebook to offset it) play a rather important role in Pritchard's epistemologically-motivated reconstruction. Thus we read that:

> " [A] feature that it is critical to the Otto case is that Otto has self-consciously decided to 'extend' his cognitive process in this way: aware that his (non-extended) memory is failing, this is the means by which he ensures that he can still get access to the information that he requires. " Pritchard (2010) p. 144

This feature, which was not part of the original scenario, plays a pivotal role in Pritchard's attempt to depict the extended mind account as a neat fit with (COGAweak-style) virtue epistemology, as evidenced (for example) in the comment that:

> "The first thing to note about the Otto case is how Otto's acquisition of the notebook, and his systematic use of it, represents a great deal of epistemic virtue on his part. A lesser cognitive agent—i.e., one who was less interested in gaining and retaining true beliefs about his

> environment—would have acquiesced in the loss of his (non-extended) memory and so accepted the epistemic consequences. Moreover, notice that the way in which Otto employs the notebook also reflects his epistemic virtue. An agent less concerned with epistemic goods would not, for example, go to the lengths that Otto goes to in order to ensure that this information resource is readily available to him but really would just use this notebook as a mere incidental aid to his cognition." Pritchard (op cit) p.145

I hope the reader will forgive the long quotations. They serve, I think, to underline the serious extent to which even proponents of the weakest available versions of virtue epistemology may yet retain traces of stronger demands concerning what I will dub the 'active pursuit of epistemic hygiene'. Thus Pritchard continues by suggesting that:

> "In contrast, if Otto had no awareness at all of the source of the reliability of his belief forming processes, nor that it was reliable, then it is hard to see why we would now regard the true beliefs that he forms as a consequence as knowledge" Pritchard (op cit p.144).

This, I suggest, is not the view that best fits with Clark and Chalmers' arguments for the extended mind. 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. The reliabilist – whatever else may be right or wrong about her account – has a much easier time here. For the notebook, even when used without any efforts, at any time, at epistemic hygiene on the part of the agent, may in fact be part of a reliable way of forming and deploying true beliefs (e.g. about the address of the Museum of Modern Art). The kinds of fact that seem essential to the epistemically virtuous use of the notebook are thus precisely the kinds of fact that seem inimical to counting the notebook as part of the realization base of a memory-like capacity in extended-Otto. In sum, the more the notebook figures in active attempts at epistemic hygiene, the less it looks like part of Otto, appearing instead as an external resource in need of careful handling.

The dilemma for the virtue epistemologist may be stated like this:

(Extended Knowledge Dilemma)

> Otto must either consciously encounter the notebook as an object for epistemically hygienic practice, or not. If he does, this makes the notebook look, at that moment, more like external equipment (it may then be a source of knowledge while failing to be part of Otto). If he doesn't, it looks unable (even on these weakened forms of virtue epistemology) to act as a source of knowledge.

So contrary to Pritchard's 'snug fit' hypothesis, it seems that we can either have knowing Otto or extended Otto, but not both. Such a dilemma would also threaten the extended mind hypothesis itself – for it seems odd to think that an agent might be belief-extended without thereby being at least potentially knowledge-extended.

3. First-Pass Responses: Innate Processes, Simple Reliablism, and Minimalist Cognitive Integration.

There are several ways to proceed at this point. Pritchard himself immediately (and rightly) worries that the demands he has just placed upon Otto would be too stringent were they placed on normal biological systems for memory and perception. We may accept the testimony of our senses and our bio-memory without awareness of the underlying processes or the reasons for their reliability (see e.g. Pritchard, op cit p.147).

On the face of it, this is a surprising concession from a virtue epistemologist. For what else, apart from simple reliability, could then be in play in these biologically basic cases? This is where the appeal to 'integration with the cognitive character of an agent' starts to play a major – but I fear somewhat inadequate and unstable – role. The idea is that equipment (neural circuits, sensory organs) that has been present since birth need not be subjected to the kinds of active attempts at cognitive hygiene that Pritchard imagined in the case of Otto. Thus we read (Pritchard, op ct p. 146) that a child may employ her perceptual abilities to gain knowledge of her surroundings even if she lacks all awareness of the source of the reliability of her belief-forming processes. Similarly, Pritchard suggests, for an agent fitted from birth with some kind of non-biological system that delivers reliable information that is then suitably integrated with whatever else the agent comes to know. In such cases the equipment counts towards the agent's cognitive character despite their failing

(at any point) to take any reflective stance on its epistemic standing. Contrast cases in which something alters, and a new device or strategy is suddenly in play. Surely, we are told:

> "such change cries out for the agent to take a reflective stance on the epistemic standing of [the] change….. if we were to imagine [the agent] being fitted with this device at a later stage, then we would require him to form a view as to the reliability of this process, and the source of this reliability, before we would regard the process as knowledge-conducive." Pritchard (2010) p. 147

I do not share this intuition. I believe that a piece of new cognitive technology could be so well-designed as to be immediately assimilated into our daily routines, requiring no reflective window before properly being counted as delivering knowledge. Such a well-fitted device would enable new states of knowing from the very moment it is fitted (much as do the sensory organs of creatures that do not have the luxury, immediately after birth, of spending time 'testing them out' in a nurturing or protected environment). I shall not, however, pursue this line of argument here. For Pritchard goes on, in any case, to make some observations that relax the 'reflective stance' requirement regarding new equipment in important and interesting ways. First, he suggests that even new and unexamined capacities may be weakly integrated just in case their deliverances are subject to a kind of consistency-checking with the deliverances of the rest of the agent's cognitive arsenal, so that the agent "would respond to discrepancies were they to occur" (op cit p. 147). Second, he suggests that over time, even an agent unknowingly augmented by some device would (assuming the previous condition is met) count as having integrated the new device into his cognitive character.

These are reasonable concessions (though the second might be thought to bear less weight than the first). All appeals to the active (agentive) pursuit of epistemic hygiene regarding the new devices have now been dropped, leaving only the requirements of reliability and weak integration. And all that weak integration here requires is something like consistency-checking (so that the deliverances of the new equipment can, in principle, be rejected when other sources of information supervene). This latter kind of integration is, however, routinely achieved using purely sub-personal resources, as we shall later (section 4) see.

A similar diagnosis can be found in Adams (2012) who notes that folk fitted with replacement lenses for cataracts need have no knowledge of the sources of

reliability of the lenses deliverances – all that matters is that, even perhaps from the very moment the are fitted, the agent trusts the deliverances of the lenses and that (as a matter of fact) they allow the lens-augmented agent to track truth in the perceptual domain. These much-weakened versions of ability theory are thus, Adams suggests, little more than terminological variants of tracking theories (though for an interesting caveat, see footnote 8 by Adams).

In fact, the literature offers an even weaker version of the ability thesis, that is still claimed to be distinct from simple reliabilist or tracking accounts. This is the 'minimalist cognitive integration' account due to Palermos (2014). All that cognitive integration requires, Palermos suggests, is that our cognitive capacities *be interconnected* in the right way. As long as this is true, and we are motivated to believe the truth, then (Palermos argues) we may reasonably trust whatever our cognitive abilities deliver simply because we encounter no reason to distrust them. The core idea here is that interconnectedness delivers a kind of background monitoring (one requiring no ongoing reflection or active effort) such that, if there something wrong with one of our integrated external devices or on-board mechanisms, then we will then notice and be motivated to take epistemically protective action. Thus he comments that:

> "This sense of epistemically adequate—yet unreflective—cognitive responsibility can only be achieved by agents like us, whose intellectual capacities are appropriately interconnected such that in cases where there is something wrong with the way we form our beliefs or with the beliefs themselves, we will be able to notice this and respond appropriately. Otherwise—if there is nothing wrong—we can go on about with our daily activities without questioning our epistemic standing with respect to every single of the millions (possibly billions?) of beliefs we enjoy in the course of our days." Palermos (2014) p 1934

This, Palermos argues, is the most minimal requirement that can still do justice to the guiding intuitions behind the appeals to ability and cognitive virtue. Moreover (the argument continues) we must do justice to these intuitions if we are to respect important facts about our own epistemic nature. Process reliabilism (and, we might add, simple versions of tracking theory) fail in this regard since:

> "While it is true that in order to know we do need the way of forming our beliefs to be objectively reliable, this sort of objective justification is not sufficient in its own. What we further need is that we be subjectively justified in the sense that we must be somehow sensitive to the reliability

> of our evidence. Process reliabilism, however, ignores this dimension of our epistemically sentient nature altogether." Palermos (2014) p. 1936

I think there is something importantly right about this worry, and that the simplest versions of reliabilism or tracking theory do therefore leave out important facts about our epistemic nature. But Palermos' way of cashing this out, when we see how it works in a little more detail, turns out (on one reading at least) to share some of the same excess baggage as that of Pritchard. The excess baggage (which we shall attempt to lose in section 4) concerns the role of agent-awareness in the presence of an appropriate sensitivity to the reliability of the deliverances of the integrated (internally or externally supported) capacities.

What Palermos is insisting upon, it seems to me, is that the agent, even though she may be unable, even in principle (op cit p. 1942) to give positive reasons for her belief, should at least be able to spot circumstances in which her belief-forming routine or process is unreliable. Thus we are told that what this kind of minimalist demand ('integration with her cognitive character') delivers is a kind of sensitivity to overall coherence such that "the agent must be able to become aware that the process is unreliable in certain circumstances" (op cit p.1939). I read this to mean that at times an agent who meets Palermos' minimalist condition must be able to become consciously aware that a belief-forming process (such as the notebook-process, in the case of Otto, or the thermometer-process, in the case of Temp) is unsafe.

In the next section, I shall question even this requirement. The most minimal version of virtue epistemology (one that avoids the dilemma laid out earlier and fits snugly with work on the extended mind) is, I shall suggest, one that does not require any kind of conscious or personal-level engagement between the agent and the cognitive process on the part of the agent at all.

4. Predictive Processing

To bring this into focus, it will help to sketch (very briefly) the shape of some recent neurocomputational thinking about the nature of perception and the relation between perception and cognition. The account I have in mind depicts the brain as, in essence, an organ whose job is to predict the incoming sensory streams.

The image of the brain as an engine of prediction can be found in various forms in contemporary neuroscience (for useful surveys, see Kveraga et al

(2007), Bubic et al (2010), and for my own favorite incarnation, see Friston (2009)). The underlying story is large and complex (for introductions, see Clark (2013) and Hohwy (2013)) but two core features of the predictive processing (PP) account stand out as especially relevant to our current concerns. First, perception involves the use of a unified body of acquired knowledge (a 'generative model') to predict the incoming sensory barrage. Second, the use of that knowledge is subject to a constant kind of second-order assessment (known as 'precision estimation') that determines the weighting assigned to specific predictions at all levels of processing, and to different aspects of the incoming sensory signal. These weightings reflect the varying reliability, in context, of differing aspects of the generative model and of the sensory inputs currently available. It is this second feature that (I hope to argue) is suggestive of an important species of sub-personal epistemic virtue. I now expand on each of these points in turn.

Concerning the first feature (the use of acquired knowledge as a 'generative model' able to predict the current sensory input) the idea here is to perform a kind of 'Bayesian flip' upon the standard (passive, feedforward) image of sensory processing. Instead of trying to build a model of what's out there on the basis of a panoply of low-level sensory cues, these models aim, in effect, to predict the current suite of low-level sensory cues from their best models of what's likely to be out there (see Hohwy (2007), (2013), Clark (2013)). The basic effect is neatly illustrated by a simple but striking demonstration (used by the neuroscientist Richard Gregory back in the 70's to make this very point) known as the hollow face illusion. This is a well-known illusion in which an ordinary face-mask viewed from the back (which is concave, to fit your face) appears strikingly convex when viewed from a modest distance. That is, it looks (from the back) to be shaped like a real face, with the nose sticking outwards rather than having a concave nose-cavity. The hollow face illusion illustrates the power of 'top-down' (essentially, knowledge-driven) influences on perception. Our statistically salient experience with endless hordes of convex faces in daily life installs a deep sub-personal 'expectation' of convexness: an expectation that here trumps the many other visual cues that ought to be telling us that what we are seeing is a concave mask. This kind of strategy, if PP is correct, both pervades and makes possible human perception. For the PP claim is that brains like ours are constantly trying to use what they already know so as to predict the current sensory signal. Downward-flowing predictions are attempting to guess (to pre-emptively specify) the states of various neuronal groups along the appropriate visual (and other) pathways. The torrent of prediction concerns all aspects of the unfolding encounter, and is not limited to simple visual features such as shape and colour. It may include a wealth of

multi-modal associations, and a complex mix of motoric and affective predictions. At the higher levels, it will involve knowledge and expectations concerning whole objects (such as cups) and their functions. As you visually inspect your desktop there occurs a rapid exchange (an energetic dance between multiple top-down and bottom-up signals) in which any incorrect downward-flowing guesses yield 'prediction error signals'. These error signals propagate upwards, and are used to leverage better and better guesses. As this process unfolds, top-down processing is trying to generate (at multiple spatial and temporal scales) the incoming sensory signal for itself. When downward-flowing ('top-down') guessing adequately accounts for the incoming signal, the visual scene is perceived.

This is a potentially dangerous process. It would not do to always see what we expect, even if the use of top-down expectations is essential for dealing with noise and uncertainty. This is where the second core feature comes into play. Thus an important feature of the predictive processing account is that the weight that is given to different elements in the processing ensemble can be varied according to their degree of estimated certainty or uncertainty. This is achieved by altering the gain (the 'volume' to use the standard auditory analogy) on appropriate aspects of the prediction error signal. One effect of this is to allow the brain to vary the balance between sensory inputs and prior expectations at different levels (see Friston (2009) p. 299). In this way the weighting of sensory prediction errors (hence the relative influence of sensory inputs and prior expectations) at any level of processing within the whole hierarchical cascade may itself be flexibly modulated. This is sometimes described as optimizing "the relative precision of empirical (top-down) priors and (bottom-up) sensory evidence" (Friston (2009) p. 299). Precision estimations thus allow for context-sensitive control of the weighting of specific prediction errors, reflecting their estimated reliability or inverse variance. Reliable information sources (e.g. vision, when the lighting is good, but not when it is bad) are thus trusted, and prediction errors calculated relative to those sources get to drive processing in the strongest ways. The same technique of variable precision weighting allows us to vary, according to context, which aspects of the generative model (which neural populations and hence which elements of our overall stored knowledge) get to bear the most weight for a given task at a given time.

How is the information (used to drive the predictions and precision estimations) acquired in the first place? Some may be innate, inherent in the basic shape of the neural economy. But a major attraction of the multi-layer predictive processing approach is that it lends itself very naturally to a form of

unsupervised (or, if you prefer, self-supervised) learning in which the attempt to predict powers the learning that makes better predictions possible. In these models, each 'higher' neural population is constantly trying to predict the rolling (ongoing) state of the neural population below it. During learning, that prediction is compared to the state that actually occurs and the neural population generating the prediction gently (automatically) self- tweaked, via standard gradient descent means so as to progressively reduce the error. The prediction task is thus a kind of 'bootstrap heaven'. To predict the next word in a sentence, it helps to know stuff about grammar (and lots more too). But one way to learn a surprising amount about grammar is just to look for the best ways to predict the next words in sentences. So you can use the prediction task to bootstrap your way to the grammar, that you then use in the prediction task in future. The information that will later inform the use of the prediction circuitry is thus acquired by using that very circuitry to try and perform predictions.

Precision estimation is learnt in the same broad fashion. We (our brains) learn that the best way to predict the current sensory signal is to decrease the weighting given to visual information on a foggy day, and to increase the weighting on visual information when lighting conditions are good. But despite being learnt in the same way (via the prediction task), the mechanisms involved in prediction and precision estimation are different (see Friston, Shiner et al (2012)). Functionally, this seems sensible since the role of variable precision weighting is essentially meta-cognitive in nature. Precision estimations reflect what we come (sub-personally) to pre-suppose about the context-varying reliability of different aspects of our own cognitive and sensory processing. Can this meta-cognitive dimension help resolve our dilemma concerning extended knowledge and the extended mind?

5. Towards a Sub-personal Virtue Epistemology

Perhaps it can. The dilemma, recall, took the following form: Otto must either consciously approach his notebook as an object apt for epistemically hygienic practice, or not. If he does, this begins to make the notebook look more like external equipment. If he doesn't, it looks unable (even on weakened forms of virtue epistemology) to act as a source of knowledge. Recall the way we earlier imagined Temp might meet a weakened form of the ability requirement (such as COGAweak). In the special scenario where the hidden helper takes a holiday every wednesday, the requirement would be met if Temp consciously noticed the trouble with Wednesdays, and refrained from consulting the thermometer

on those days. This would clearly also meet Palermos' demand that the agent employ a 'conscientious attitude' (Palermos 2014) p.1941) to the process by *becoming aware* that it is unreliable in certain circumstances.

But there is an even weaker possibility that we should now consider. Imagine Temp2. Temp2 is a predictive processing agent equipped with a rich arsenal of sub-personal mechanisms dedicated to estimated the reliability, in context, of the various sources of information that impinge upon her senses. Over time, sensory cues indicative of the Wednesday context (e.g. seeing the word Wednesday on the computer date-screen that she views every morning before starting work) come to be associated with unreliable information from the thermometer. Sub-personal precision-weighting mechanisms then cause Temp2, even when she looks at the thermometer, to fail to attend to what it says. In effect, she simply ignores the thermometer readings on Wednesdays. Counterfactually, were things to change (were the helper to start taking Tuesdays off instead), the same sub-personal mechanisms would slowly update their estimations.

This is not, in broad outline, an implausible scenario. It is well-known that simple associative learning mechanisms can operate without conscious awareness on the part of the agent, as the large literature on implicit bias (see Bargh (2006)) rather starkly demonstrates. So much so, in fact, that a recent review concludes that:

> "The unconscious is not identifiably less flexible, complex, controlling, deliberative, or action-oriented than is its [conscious] counterpart" Bargh and Morsella (208) p. 73

Non-conscious goals, desires, motivation, and strategies, of surprising scope, sophistication, and flexibility, are ubiquitous in both human and (other) animal cognition. Even the so-called 'central executive' functions, that control our choice of strategies and are called upon to inhibit pre-learnt but (contextually) unwanted behaviors can, it seems, become activated and operate entirely outside the sphere of conscious awareness – see Lau and Passingham (2007), Reuss et al (2011), Soto et al (2011), and for a review and discussion, Dehaene et al (2014).

Given the scope and power of unconscious processing, it should come as no surprise to learn that some forms of epistemic hygiene may likewise be non-consciously acquired and non-consciously deployed. Specifically, the large and

crucial second-order apparatus of precision-weighting of prediction error constitutes a potent means of responding, without need for awareness, to the context-varying reliability of our (inner and outer) information sources. This is because variable precision-weighting provides a tool that can favour some sensory inputs or top-down expectation over others, but also (more generally) one that can enhance the effects of any neural population, changing the moment by moment patterns of 'effective connectivity' (ref) that determine the varying flows of information as we perform our tasks.

Benefitting from the unconscious operation of such potent resources Temp2, I suggest, exhibits a genuine (indeed, fundamental) form of epistemic hygiene. She benefits from unconscious meta-cognitive mechanisms that allow her to rely or not rely upon specific information sources, both internal and external, in ways dictated by their context-varying reliability as truth-tracking (or at any rate, successful behavior-enabling) devices. Importantly, Temp2 can display this kind of epistemic virtue even if she is never consciously aware of (for example) her Wednesday-distrust of the thermometer. That is to say, nothing here requires her to consider her thermometer-consulting process at all. That process need not show up anywhere in her thinking – not even when, on Wednesdays, she fails to trust it.

With this case in mind, consider once more the story about Otto and his notebook. Clearly, neural mechanisms specializing in the estimation of precision are not themselves operating within the notebook. But the notebook (just like the thermometer in the case of Temp2) delivers a stream of information apt for automatic non-conscious meta-cognitive assessment whenever the notebook resource is invoked in active problem-solving. This means that the notebook case can be treated in the same way as Temp2[1]. But with one important difference. The notebook, by being a constant, automatically-invoked, resource, meets the additional requirements that (arguably, following Clark and Chalmers (1998)) warrant treating it as forming part of an extended cognitive circuit. But the information in the notebook, when the notebook is invoked in a bout of active problem-solving, is subject to all the automatic sub-personal checks and balances that apply to information retrieved from bio-memory. These checks and balances (the automatic 'precision-weighting' of information and information sources) in no way require that the notebook be encountered, or even be poised to be encountered, by the agent as an object of active epistemic scrutiny.

To sum up, a core feature of the predictive processing story is that cognitive strategies are selected (moment-by-moment) by neural mechanisms that

automatically estimate the reliability of information. Some of that information is stored in bio-memory. But it can just as well be stored 'in the world' and accessed when required. The very same neural estimation mechanisms may thus 'endorse' the loop out into the world (Otto's notebook, for example) by assigning high reliability to information thus retrieved and using it to guide action. Internally stored information and information stored bio- externally (but accessibly) are thus equally apt to be accorded high 'precision'. This simply means that Otto's brain treats the notebook-loop as a high-grade information source, just as it might treat some (but not all) aspects of its own bio-memory. The same brain-based sub-personal reliability-estimating processes are here able to target both inner and outer sources of information, putting them nicely on a par as contributors to 'what the agent knows'. Because these estimations operate below the level of conscious engagement or agentive effort, the tensions highlighted earlier simply do not arise.

I offer this as a sub-personal take on core insights from virtue epistemology. But another way to think about it is simply as a kind of second-order reliabilist vision[2]. Brains that are able to estimate precision are able to estimate the reliability of their own information sources, crunching together sensory evidence and prior expectations in a broadly Bayesian fashion, and weighting different forms of evidence (e.g. from different sources, senses, and inner circuits) according to task and context. To accept the epistemic merit of these sub-personal modes of meta-cognitive evaluation is thus to endorse a subtle version of reliabilism – one that places non-conscious, automatic epistemic self-evaluation (a species of meta-cognition) centre-stage[3]. What I am calling a weak (because entirely sub-personal) version of virtue epistemology is thus identical with a sophisticated (second-order) reliabilism, in which our brains assess the context-sensitive reliability of their own evidence and resources.

6. Back to The Extended Mind.

Appeals to automatic, continuously operating neural mechanisms like these can seem like a step backwards from the original vision of the extended mind. For such appeals put on-board mechanisms of precision estimation centre-stage. But the point of the extended mind story was never to deny the importance of the brain in enabling and maintaining the environment-exploiting loops that (if you buy the arguments) extend cognitive processing into the world. Rather, it was to show that, considered in the context of an active, cognitively well-endowed organism, certain apparently bodily or worldly goings-on might form parts of the realization base for some cognitive capacities. Nothing in the PP framework materially thus alters, as far as I can tell, the arguments previously

presented (as summarized in e.g. Clark (2008)) both pro and con, regarding the possibility and actuality of genuinely extended cognitive systems. What PP does offer, however, is a specific proposal concerning the shape of the specifically neural contribution to cognitive success.

It is a proposal, moreover, that dovetails neatly with work on embodied cognition. To see this, reflect that actions that engage and exploit specific *external* resources will now be selected in just the same manner as the inner coalitions of neural resources themselves. Thus consider the case where salient high-precision information is available by the use of some bio-external device, such as a laptop or smartphone. The core routine that selects actions to reduce prediction error will now select actions that invoke the bio-external resource. Invoking a bio-external resource, and moving our own effectors and sensors to yield high-quality task-relevant information are here expressions of the same underlying prediction error minimizing strategy, reflecting our brain's best estimates of where and when reliable, task-relevant information is available. Apt precision assignments could, for example, select neural circuits that select actions that manipulate the beads of an abacus, so that the resulting perception-action cycles solve a mathematical problem.

Seen from this perspective, the selection of task-specific inner neural coalitions is entirely on a par with the selection of task-specific neural-bodily-worldly ensembles. In each case, what is selected is a temporary problem-solving ensemble (a 'temporary task-specific device' – see Anderson, Richardson, and Chemero (2012)) recruited as a function of context-varying estimations of uncertainty. As such cycles unfold, no conscious agency or inner homunculus need oversee the repeated soft-assembly of the distributed problem-solving ensembles that result. Instead, such ensembles emerge and dissolve in ways determined by the progressive reduction of precise, high-quality, prediction error. Organismically salient (high precision) prediction error may thus be the glue that, via its expressions in action, binds elements from brain, body, and world into temporary problem-solving wholes.

Moreover, precision estimation itself can be partially outsourced, scaffolded, and amplified using bio-external tools and resources. Talking to oneself or others is, arguably (see Lupyan and Clark (submitted)) a means of artificially manipulating precision estimations, upping the gain on some neural populations at the expense of other. So too is the use of alarms or external devices that alert us when salient new information becomes available. Such an effect is described by Vaesen (2011) in his account of Sissi. Sissi is an airport security scanner operator. But the task is boring and levels of attention can

lapse allowing suspicious objects free passage through the scanner. To remedy this, new scanners occasionally project a false image of a suspicious object, and this helps maintain alertness for such images (and enables supervisory monitoring of alertness too). This is a nice case of an external circuit that is scaffolding neural precision assignments[4].

7. Conscious Epistemic Engagement – Why Bother?

We are left with an important question. Given all that built-in sub-personal epistemic hygiene (automated assessments of the context-variable reliability of internal and external information sources), what does conscious, agentive engagement bring to the table?

This is obviously a huge (though I think under-explored) topic. But I'd like to end with two brief speculations. First, and most obviously, conscious awareness is a necessary (but not sufficient) condition for conscious awareness of our own thought processes. Thus a conscious agent can treat her own thoughts, and those of other agents, as objects for personal-level scrutiny and assessment. Written and spoken language may play a large role in enabling this kind of 'objectification' of thought. It is only thanks to this objectification (see Clark (1998) (2006) for some discussion) that we humans have been able to create and benefit from the kinds of practice that Sterelny (2003) describes as "incremental downstream epistemic engineering." This means we create and benefit from slowly evolved culturally transmitted practices that improve our collective and individual grip on the world. For example, we build designer environments for learning, and for teaching teachers, and even for teaching people how to teach teachers. But second, as noted by Pritchard (2013), what we teach is (or at least can be) precisely how to take an active stance with regard to our own beliefs and knowledge sources. In other words, we don't (or ought not) simply teach facts. Rather, we install methods and practices that help students probe and test their beliefs and knowledge sources, and deepen their understandings. The upshot of such discipline, Prichard concludes, is a greater capacity to cope when environmental conditions alter or new challenges arise.

One way to think about these 'cognitive agency' effects, from the PP perspective, is to see them as providing further ways of contextualizing the complex stack of lower-level capacities upon which they are built. Thus, if the number that appears on my calculator looks way out of line with my rough estimate I may, as a conscious agent, ask myself whether the device is faulty or

I entered the wrong information. It is hard to see how an agent with no access to her own problem-solving procedures could interrogate her own performance in this way. There is a close analogue in sports skills, where only reflective agents can ask what aspect of their golf swing (for example) was at fault when the ball lands in the rough. Perhaps reflective conscious agents are nothing but those agents whose higher-level models include models of their own capacities and of themselves as agents attempting to know their worlds? There is thus a certain sense (if you buy the dilemma described earlier) in which consciousness is a necessary condition of a form of self-alienation – a form of self-alienation that opens up the space for all these more deliberate forms of epistemic engineering.

8. **Conclusions: Minds as Means**

The kinds of informed awareness and vigilant use that would allow (at least on strong virtue-based models) the outputs of some external cognitive aid to count as my own knowledge are, prima facie, a poor fit with the standard profile of (putatively) extended minds. This is because typical extended mind scenarios rely upon fluid unreflective use as one of the markers of incorporation into my own extended cognitive architecture, thus distinguishing true incorporation from mere - even if careful - tool use. By contrast, strong appeals to cognitive virtue suggest the kinds of personal-level awareness, vigilance, and monitoring that best characterize our early encounters with external knowledge sources – ones that, for all their practical value, do not extend the realization base of human minds. I have argued that we ought not demand, of a putative mind-extending device, that it be subject to *any* form of agent-aware scrutiny whatsoever – not even the weaker kinds of agentive engagement imagined by Pritchard and (I think) Palermos.

The good news, however, is that when minds like ours encounter a world, they do so using sub-personal mechanisms that already build-in huge and game-changing amounts of sub-personal epistemic care. In particular, work on the predictive brain highlights the importance of a class of second-order processes that estimate the reliability (the 'precision' or inverse variance) of both internal and external sources of information, altering these estimates in ways determined by task, conditions, and context. Variations in these contextually-determined weightings allow optimal or near-optimal combinations of information from multiple sensory sources, and disturbances to these second-order mechanisms are currently thought to result in various cognitive pathologies such as the emergence of delusions and hallucinations. Importantly,

the outputs of non-biological devices (or other biological agents) can be fluently incorporated into these sub-personal precision assignment schemes. In this way the deliverances of non-biological information sources may be rendered (still without need for personal-level overseeing or involvement) as epistemically secure, or insecure, as those of our biologically-based onboard devices.

In real-world cases, however, our first encounters with new tools or technologies will often invite additional forms of conscious, agentive epistemic care - although Adams' case of the replacement lenses provides a nice exception here, perhaps because it is so easy to 'check' that they are reliably doing the job from the moment they are first fitted. Such exceptions aside, the bulk of real-world cases will conform to Pritchard's intuition that there is some epistemically important difference – in the early stages - between innate and subsequently-added capacities. What should still be resisted, however, is any lingering thought that agentive epistemic care is itself part of, or in some way supportive of, the extended mind scenario. For if anything, active agentive care works temporarily against that scenario. What human and animal minds (and all their constituent parts) provide first and foremost is a means, usually invisible to the agent, to encounter a world. The active interrogation of those means, though of great epistemic significance, is – as least as far as mindedness itself goes – simply icing on the cake. An interesting question for future research is thus in what ways designers and users of new tools and technologies might exercise due epistemic caution while *simultaneously* aiming for the fluid incorporation of those tools and technologies deep into our cognitive repertoires. Drawing attention to this delicate balancing act is important both for our epistemic health and for our evolving understanding of the nature and limits of true cognitive extensions.

* Thanks to Adam Carter, Orestis Palermos, Duncan Pritchard, Mikkel Gerken, and an anonymous referee, for helpful comments on earlier drafts of this material. This work was supported in part by the AHRC funded 'Extended Knowledge' project, based at the Eidyn research centre, University of Edinburgh.

References

Adams, F. (2012). Extended cognition meets epistemology. *Philosophical Explorations*, *15*(2), 107–119. doi:10.1080/13869795.2012.670717

Anderson, M.L., Richardson, M.J., and Chemero, A (2012) Eroding the Boundaries of Cognition: Implications of Embodiment *Topics in Cognitive Science* 4: 4: p.717–730

Bargh, J. A. (2006). Agenda 2006 What have we been priming all these years ? On the development , mechanisms , and ecology of nonconscious social behavior, *168*(January), 147–168.

Bargh, J. a., & Morsella, E. (2008). The Unconscious Mind. *Perspectives on Psychological Science*, *3*(1), 73–79. doi:10.1111/j.1745-6916.2008.00064.x

Bubic, A., von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, *4*(March), 25. doi:10.3389/fnhum.2010.00025

Clark, A (1998) Magic Words: How Language Augments Human Computation, in P. Carruthers and J. Boucher (Eds) *Language And Thought: Interdisciplinary Themes* (Cambridge University Press: Cambridge,) P.162-183

Clark, A (2006) Language, Embodiment and the Cognitive Niche *Trends in Cognitive Sciences* 10:8:370-374

Clark, A (2008) *Supersizing the Mind: Action, Embodiment, and Cognitive Extension* (Oxford University Press, NY)

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, *36*(3), 181–204. doi:10.1017/S0140525X12000477

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

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews. Neuroscience*, *10*(1), 48–58. doi:10.1038/nrn2536

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. doi:10.1016/j.tics.2009.04.005

Friston, K., & Adams, R. (2012). Perceptions as hypotheses: saccades as experiments. *Frontiers in Psychology 3*(May), 1–20. doi:10.3389/fpsyg.2012.00151

Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., … Bestmann, S. (2012). Dopamine, affordance and active inference. *PLoS Computational Biology*, *8*(1), e1002327. doi:10.1371/journal.pcbi.1002327

Goldman, Alvin I. (1986). *Epistemology and Cognition*, Cambridge, MA: Harvard University Press.

Goldman, A (2012) *Reliabilism and Contemporary Epistemology: Essays* (New York:Oxford University Press).

Greco, J. (2012). A (different) virtue epistemology. *Philosophy and Phenomenological Research*, 85(1), 1-26.

Hohwy, J. (2007). Functional integration and the mind. *Synthese*. Retrieved from http://link.springer.com/article/10.1007/s11229-007-9240-3

Hohwy, J (2013) *The Predictive Mind* (Oxford University Press, Oxford, UK)

Kornblith, H (2012) *On Reflection* (Oxford University Press, Oxford)

Kveraga, K., Ghuman, A. S., & Bar, M. (2007). Top-down predictions in the cognitive brain. *Brain and Cognition*, *65*(2), 145–68. doi:10.1016/j.bandc.2007.06.007

Lau HC, Passingham RE (2007) Unconscious activation of the cognitive control system in the human prefrontal cortex. *J Neurosci* , 27:5805-5811

Lupyan, G. (forthcoming). Cognitive penetrability of perception in the age of prediction: Predictive systems are penetrable systems In *Review of Philosophy and Psychology*.

Menary, R. (ed) (2010) *The Extended Mind* (MIT Press, Camb. MA).

Nagel, S. K., Carl, C., Kringe, T., Märtin, R., & König, P. (2005). Beyond sensory substitution--learning the sixth sense. *Journal of Neural Engineering*, *2*(4), R13–26. doi:10.1088/1741-2560/2/4/R02

Palermos, S. O. (2014). Knowledge and cognitive integration. *Synthese*, *191*(8), 1931–1951. doi:10.1007/s11229-013-0383-0

Pritchard, D. (2010). Cognitive ability and the extended cognition thesis. *Synthese*, *175*(S1), 133–151. doi:10.1007/s11229-010-9738-y

Pritchard, D. (2013). *Intellectual Virtue , Extended Cognition , And The Epistemology Of Education*. *Journal of Philosophy of Education*, *47*(2), 1–20.

Reuss H, Kiesel A, Kunde W, Hommel B (2011) Unconscious activation of task sets. *Conscious Cogn* , 20:556-567. 41. _

Soto D, Mantyla T, Silvanto J (2011) Working memory without consciousness. *Curr Biol* 21:R912-R913

Sterelny K (2003) *Thought in a Hostile World: The Evolution of Human Cognition* (Blackwell, Oxford)

Vaesen, K. (2011). Knowledge Without Credit: Exhibit 4 -Extended Cognition. *Synthese* 181:3: 515-529,

[1] Thanks to an anonymous referee for encouraging me to clarify the shape of the argument at this point.

[2] Recent treatments that highlight varying forms of second-order reliablist intuition include Goldman (2012) and Kornblith (2012).

[3] The proper functioning of these non-conscious epistemic resources is, however, plausibly a pre-condition for the proper functioning of their conscious counterparts. Thus agents whose basic precision-assignment mechanisms malfunction will, it has been suggested, be prone to both sensory hallucinations and delusions (see Fletcher and Frith (2009)). Sub-personal epistemic virtues may thus provide the necessary bedrock for conscious, personal-level epistemic achievements.

[4] Sissi may or may not constitute a case of extended cognition (it will depend on how you feel about temporary, highly environment-specific brain-body-world coalitions – contrast Otto who always carries the notebook). But it is easy to imagine Sissi-style functionality being implemented using a constantly-carried or worn resource – e.g. a 'phone that gives a gentle (potentially below conscious awareness) buzz whenever some particular kind of shop is nearby, so that over time the agent starts to feel they 'just know' when to linger in that area or explore a new back-alley. Such technologies have already been explored using 'FeelSpace' belts that constantly channel information about the position of magnetic north (Nagel et al (2005)).