AI & SOCIETY (2019) 34:857–866
https://doi.org/10.1007/s00146-018-0824-x

ORIGINAL ARTICLE

The reappearing tool: transparency, smart technology, and the extended mind

Michael Wheeler[1]

Received: 14 April 2017 / Accepted: 24 January 2018 / Published online: 7 February 2018

Abstract
Some thinkers have claimed that expert performance with technology is characterized by a kind of disappearance of that technology from conscious experience, that is, by the transparency of the tools and equipment through which we sense and manipulate the world. This is a claim that may be traced to phenomenological philosophers such as Heidegger and Merleau-Ponty, but it has been influential in user interface design where the transparency of technology has often been adopted as a mark of good design. Moreover, in the philosophy of cognitive science, such transparency has been advanced as necessary for extended cognition (the situation in which the technology with which we couple genuinely counts as a constitutive part of our cognitive machinery, along with our brains). By reflecting on concrete examples of our contemporary engagement with technology, I shall argue that the epistemic challenges posed by smart artefacts (those that come equipped with artificial-intelligence-based applications) should prompt a reassessment of the drive for transparency in the design of some cases of technology-involving cognition. This has consequences for the place of extended minds in the contemporary technological context.

Keywords Artificial intelligence · Extended cognition · User interface design · Skilled tool use · Phenomenological transparency

1 Extended senses and extended minds

Designed by *Cyborg Nest*, a group of digital pioneers and transhumanists, *North Sense* is a recent addition to the personal technology market (https://cyborgnest.net/products/the-north-sense, last accessed 24 March 2017; for an introduction and discussion, see Emslie 2017). At first sight, it might seem that this product is not something to get overly excited about. Its only mechanical function is to vibrate gently when it faces magnetic north, and there is nothing particularly new about technology detecting the Earth's electromagnetic field for us. That's what compasses do. However, any such indifference would be misplaced, because *North Sense* is pushing at the door of our species' much-heralded cyborg future. Tiny, and encased in body-compatible silicone, it is designed to be fixed onto the upper part of the wearer's chest using piercings. This distinctive and unusual corporeal anchoring means that the magnetic-north-tracking vibrations produced by the device are felt in an intimate way by the wearer. In addition, *North Sense* is designed to be left permanently turned on, meaning that its effects are standing features of the wearer's ongoing experience.

First-person reports from early adopters of *North Sense* provide evidence that the technology quickly becomes deeply integrated into the wearer's cognitive life. Most strikingly, orientation and position start to play a bigger-than-usual role in the structuring of memory. Scott Cohen, one of the co-founders of *Cyborg Nest*, and also one of the first to have *North Sense* fitted, describes this phenomenon as follows: 'It is hard to put into words only a few hours after attaching the North Sense, but the feeling I am left with is profound. The impact of immediately sensing my position created a permanent memory. I vaguely recall the colours and sounds in the room, but I remember my position vividly.' (https://www.cyborgnest.net/, last accessed 21 March 2017) What *North Sense* offers, then, is a powerful combination of corporeal and experiential close coupling, permanent presence, and altered cognitive processing—a transformative cocktail of entanglement and enhancement.

✉ Michael Wheeler
m.w.wheeler@stir.ac.uk

[1] Division of Law and Philosophy, University of Stirling, Stirling FK9 4LA, UK

Indeed, the *Cyborg Nest* team claims that, with *North Sense* installed, one does not merely deploy technology that allows one to detect the Earth's magnetic field; one senses that field. By these lights, *North Sense* is an artificial sense organ that expands our human perceptual capacities. And that, it might surely be argued, really is new.

As I interpret these claims, given that *North Sense* reliably detects magnetic north—something that human beings cannot ordinarily do with their purely organic sensory machinery—it seems largely straightforward that if *North Sense* is genuinely part of its wearer's sensory machinery, on a par with her eyes and ears, then it has succeeded in endowing that person with a new sense. The hard part, I think, is securing the antecedent in this conditional, that is, the claim that *North Sense* genuinely counts, in an entirely non-metaphorical way, as a constituent part of its wearer's sensory machinery. So what precisely is it about a corporeally installed *North Sense* device that might conceivably justify this conclusion?

Let us widen our view. For although the specific features of *North Sense* are intriguing, the question just posed exposes an issue that reaches well beyond that particular piece of technology and, indeed, well beyond perception as a psychological capacity. Here, then, is the more general question of interest: when, if ever, does an item of technology-in-use genuinely count as part of one's psychological machinery (sensory or otherwise), as opposed to part of the external world? This is a question that has been exercising a number of philosophers of cognitive science in recent years, in the debate over the so-called *extended mind hypothesis* (henceforth *ExM*; see Clark and Chalmers 1998; Clark 2008 for canonical treatments, and for a more recent collection that contains criticisms, defences, and developments of the view, see; Menary 2010). Advocates of ExM (or of the hypothesis of extended cognition—I shall use the terms 'mind' and 'cognition' interchangeably) hold that the physical machinery of mind sometimes extends beyond the skull and skin. More precisely, according to ExM, there are actual (in this world) cases of intelligent thought and action, in which the material vehicles that realize the thinking and thoughts concerned are spatially distributed over brain, body, and world, in such a way that certain external elements are rightly accorded fundamentally the same cognitive status as would ordinarily be accorded to a subset of your neurons.

Two clarificatory remarks regarding ExM: First, when one wonders what sorts of external elements might figure in instantiations of an extended mind, it is immediately tempting to point to the array of information-sensitive, information-gathering, information-storing, and information-manipulating devices readily found in our high-tech modern world, devices such as smartphones, tablets, and instances of wearable computing. However, it is worth noting that, as far as the letter of ExM goes, less sophisticated items of 'information technology' such as notebooks (the old-fashioned kind), tally sticks, and abacuses would, under the right circumstances, do just as well. To borrow Andy Clark's rich and suggestive phrase, we human beings are *natural born cyborgs* (Clark 2003) who have been coupling with technology pretty much for as long as we've been on the planet as a species. Nevertheless, there seems little doubt that much of the contemporary interest in ExM stems from the way in which it chimes with, and helps us to come to terms with, our modern experience of technology.

The second clarification is that, in the characterization of ExM given above (which, as a formulation of the view, is not in any way idiosyncratic), the term 'extended' has the sense of spatial (environment-encompassing) extension, not of performance enhancement. Of course, in some cases of (spatially) extended cognition, psychological performance will, indeed, be enhanced, and thus, cognition will be extended in that other sense too. For example, if *North Sense* is genuinely part of its wearer's sensory machinery, then it extends her psychological machinery in both senses of the term. Thus, although it is physically connected to the skin in an intimate way, it remains an external element (i.e., it is not a biological part of the human organism), so it provides a case of ExM; and it allows the wearer to track magnetic north—something that she could not do prior to having it fitted—so it enhances psychological performance. Of course, even if *North Sense* is not genuinely part of its wearer's sensory machinery, it still might enable her to respond cognitively to new environmental stimuli, in which case it extends her mind in the enhancement sense of the term, but not the spatial sense.

The foregoing discussion should remind us once again that it is the 'if' in phrases such as 'if *North Sense* is genuinely part of its wearer's sensory machinery' that remains to be substantiated. In other words, we still need an account of when some external element, such as an item of technology, qualifies as a constituent part of one's psychological machinery, that is, in the relevant sense, as part of one's extended mind. As one might expect, there are several proposals out there for delivering this result, and here is certainly not the place to rehearse them all or to explore the sometimes bad-tempered debate that has surrounded them (Menary 2010, is a good place to start; for my own favoured way of arguing for ExM, see e.g.; Wheeler 2010, 2011a, 2013). Therefore, I shall explore just one thought in the vicinity, which is that cognitive extension depends on the phenomenological property of transparency. The concept of transparency at work here has its roots in certain prominent philosophical—more precisely, phenomenological—analyses of tool use. The next section of this paper will explain the notion more detail. Just to get us going, however, the rough and ready idea is this: where external technology is used in a skilled and hitch-free manner, that technology disappears

from the conscious apprehension of the user. It is invisible to her. The proposal that will concern us here, then, is that this sort of disappearance or invisibility on the part of the technology—henceforth, its transparency—is necessary for that technology to meet the target constituency condition and thus for it to furnish us with a case of ExM. In saying that this proposal will concern us here, I should stress that my aim is not to defend it, but rather to lay it out and then to discuss certain issues that come to light once one takes it as a point of departure. Ultimately, I will be interested in what the consequences are when transparent technology—technology which might, partly in virtue of its transparency, count as, or at least be on the way to counting as, part of our psychological machinery—is itself smart, that is, equipped with artificial intelligence (AI).

2 The disappearing tool

To bring the pivotal notion of transparency into proper view, let us begin by taking a peak at some of its philosophical past. At one point in his magnum opus, *Being and Time*, Heidegger (1927) presents an analysis of our everyday experiences with entities. Famously, he argues that we ordinarily encounter entities as (what he calls) equipment, that is, as being for certain sorts of tasks (cooking, hair-care, text-editing, navigation, and so on). Designed tools clearly provide the paradigm case of equipment, although not the only instance. According to Heidegger, when we skillfully manipulate equipment in a hitch-free manner, we have no conscious apprehension of the items of equipment in use as independent objects, that is, as something like identifiable bearers of determinate states and properties. Thus, to use Heidegger's most-quoted example, while engaged in trouble-free hammering, the skilled carpenter has no conscious recognition of the hammer, the nails, or the work-bench, in the way that one would if one stood back and thought about them. In other words, tools-in-use become phenomenologically transparent. All we experience (often in an indeterminate and non-thematic way) is the ongoing task (e.g., the hammering). Therefore, to be more specific, it is not only the tool itself but one's interface with it that disappears. Of course, transparency is not the only possible way of relating to equipmental entities. When skilled practical activity is disturbed, say by a broken or malfunctioning tool that is in need of repair, that item of equipment is certainly no longer phenomenologically transparent to the user. It is itself apprehended in that user's conscious experience, either as presenting a barrier to skilled activity or, in extreme cases, as a context-free object with, for example, a certain weight or size. Nevertheless, the idea is that, under the right circumstances, equipment becomes transparent.

It is worth pausing here to make a clarificatory remark about a rather different notion of transparency that is in the same technology-oriented ballpark as our target concept. My aim here is merely to register the fact that this alternative notion of transparency exists and is genuinely different, and then to put it to one side, so that it does not confuse matters. Sometimes, technology is described as being transparent when a specified class of users is able to understand precisely how it functions. This is a perfectly reasonable notion of transparency, but note that a device which is transparent in this sense may be broken or malfunctioning, and so will not be transparent in the phenomenological sense, and that a device which is phenomenologically transparent in use may be impenetrable in its inner workings, and so will not be transparent in the 'open to understanding' sense. Therefore, there is a double dissociation between the two concepts. I will be concerned only with the phenomenological notion.

To return to the main plot, in our ordinary way of thinking about things, tool use is a matter of action, of changing or manipulating the world: think of using a hammer, to bang in nails, to build something, that is, to bring something new into existence. However, as a famous example from Merleau-Ponty (1945) indicates, there is also a perceptual dimension to some cases of tool use. Merleau-Ponty observes that a blind person using her cane in a skilled and hitch-free manner does not consciously apprehend the cane itself. On a first impression, it might seem that this is simply another case of transparency in action: the blind person uses the cane for finding her way around, and when she does so in an expert, smooth, and undisturbed fashion, the cane disappears from conscious apprehension. However, now notice that the cane is also a device that enables the blind person to access the world—to locate things in space. From this perspective, when one says that the blind person no longer consciously apprehends the cane in use, one might well conclude that, in that respect, the cane is just like the biological machinery that constitutes one of her (properly functioning) organic sense organs. When hearing is going well, she does not experience her ears as information-gathering objects. In fact, in ordinary usage, she does not experience her ears as such at all. They are transparent to her. What happens is that she experiences the world through her ears. The situation is similar with regard to her cane. She doesn't experience the cane as such. It is transparent to her. She experiences the world through her cane. Put another way, the blind person's experiential interface is with the world beyond the cane, not with the cane itself.

In some cases, then, one accesses the world through technology, and in doing so, one accesses the world in a new way. Moving beyond the analyses offered by phenomenological philosophy, one might add detail to this picture by focussing on the experiences of so-called sensory substitution subjects. The phenomenon of sensory substitution

occurs when technological augmentation enables one sensory modality to support the kind of environmental access and interaction ordinarily supported by a different sensory modality. The seminal work in this area began with Bach-y-Rita's (1972; Bach-y-Rita and Kercel 2002) research on tactile-vision sensory substitution (henceforth TVSS). In one version of this work, congenitally blind subjects are equipped with a head- or shoulder-mounted camera that conveys information, from video images, via the activation of an array of vibrators located on the subject's back, abdomen, or thigh. After a short period of adaptation, those TVSS subjects who actively control the information received, either by manipulating their bodies or by manipulating the camera, are able to make reliable judgments about things such as the number, relative size, and position of distal objects in three-dimensional space, and to perform actions such as reaching out and picking up objects. Moreover, sensory substitution subjects routinely report a shift in perceptual experience, with some organically blind users of sensory substitution systems reporting experiences that might be categorized as visual qualia, such as experiences of phosphenes (the seeing of light without light actually entering the eye) (Ortiz et al. 2011). Therefore, blind subjects now access and experience the world in ways that are characteristic of the distal sense of vision, a sense that, organically speaking, they do not have. The most straightforward (but also the most controversial) claim in the vicinity here would be that post-adaptation TVSS subjects genuinely see (enjoy authentic visual phenomenal consciousness), in spite of the fact that the relevant channel of proximal stimulation is tactile (vibrations on the skin caused by the TVSS technology). It may be, however, that what one ought to say is that TVSS engenders a transformation in perceptual consciousness, such that, even though the proximal stimuli remain tactile in character, the post-adaptation conscious experience is not correctly categorized as one of touch, even if it is not vision.

For present purposes, the most important feature of TVSS is that post-adaptation subjects do not normally have the conscious tactile experience of the vibrations taking place on the surface of their skin. Rather, they simply experience the distal world that the technological augmentation makes available to them. And although it is possible for some subjects to switch sensory mode so as to experience the vibrations, that requires a deliberate cognitive effort on their part, at which point the experiences of the distal world that the technology supports are lost to them. Therefore, even though the transformative effect of sensory substitution technology is arguably greater than that of the blind person's cane, the fundamental phenomenological signature exhibited remains in force: in hitch-free usage—including usage that is not disrupted by purposeful willings by the subjects concerned—TVSS devices are transparent. Their users do not experience the technology itself; they experience the world through the technology.

TVSS is a case of sensory substitution, since environmental access and conscious experiences usually achieved via the eyes are achieved via a different route. This remains a plausible way of describing things, even if one takes the outcome to be vision-like, rather than strictly vision. Now, we have just seen that TVSS technology-in-use is routinely transparent. However, what about cases of sensory enhancement (of cognitive extension in the secondary sense identified earlier), cases in which a device allows its user to be sensitive to environmental stimuli that are not accessible by the technologically unaugmented organic human senses? One might reasonably wonder whether the transition from substitution to enhancement is a phenomenological game changer here. More specifically, is the hitch-free use of sensory enhancement technology ever transparent? To bring the answer to this question into view, we need to register the fact that we have been making the distinction between sensory substitution and sensory enhancement from a species-level perspective, with the dividing line set by the evolved organic sensory repertoire that human beings standardly possess. From that vantage point, TVSS and, say, *North Sense* fall on opposite sides of the substitution-enhancement line. However, from the perspective of the individual congenitally blind subject, TVSS is already a case of sensory enhancement. The congenitally blind person's sensory repertoire does not include vision, so the TVSS technology is a sensory enhancement rather than a sensory replacement. In other words, once we adopt the first-person perspective, TVSS and *North Sense* are equivalent and there is no reason to think that their experiential profile need be any different. In fact, *North Sense* itself is a somewhat complicated example, because its principal epistemic effect is in terms of an enhanced appreciation of spatio-bodily position and orientation, rather than in terms of, say, how many objects there are on the other side of the room. Nevertheless, and intriguingly, verbal reports from early adopters of the *North Sense* package suggest that the transparency condition is sometimes met. For example, Scott Cohen, mentioned earlier, talks of '*immediately* sensing [his] position' when using the device (https://www.cyborgnest.net/north-sense, last accessed 15 March 17, my emphasis), and Liviu Babitz (also from *Cyborg Nest*) reports that, after several weeks of wearing the device, he sometimes fails to feel the vibrations, even though the device is functioning successfully (interview for the BBC's *Tomorrow's World*, http://www.bbc.co.uk/guides/zs4btv4?intc_type=singletheme&intc_location=tomorrowsworld&intc_campaign=tomorrowsworld&intc_linkname=guide_brainhacking_contentcard15, last accessed 15 August 2017). It seems, then, that enhancement does not undermine the possibility of transparency.

The final stop on our tour of transparency takes us beyond both the idea of tools designed for specific actions and the idea of tools as dedicated sensory channels. It is already the case that we routinely offload data storage (one kind of 'memory') onto our personal technology. To give just one everyday example, in the age of mobile phones with contacts lists, most people commit far fewer phone numbers (if any) to their organic memory stores. However, the bulk of the contextual reasoning by which we decide, for example, precisely why and when we need to make a call, and whom we need to call, remains a neural achievement. But now, imagine a not-too-distant future in which advances in AI enable us to offload the contextual reasoning too onto technology. This is only just science fiction. For example, James Kozloski at IBM has patented the technology for what he calls a cognitive assistant, a device that uses a combination of surveillance (monitoring its owner's activity, perhaps, as a wearable), machine learning (to become knowledgeable about patterns, regularities, events, and situations in its owner's life), and predictive processing (to determine what information its owner needs in her present situation). The device aims either to provide its owner with the information she needs or to give her appropriate prompts, so that she can retrieve that information from her own brain (see LaFrance 2016). What on-the-horizon technology such as the cognitive assistant demonstrates is that the notion of technological enhancement is not limited to new sensory discriminations or to the effects that such discriminations may have on the character of memory, but stretches into the very heart of cognition central—to capacities such as categorization (what sort of situation am I in?) and decision-making (what should I do?). In addition, given that, as I have argued, the possibility of transparency is not disrupted by the fact of enhancement, there seems to be no particular reason, on grounds of the psychologically central nature of the enhancement concerned, to think that such smart tools will not become transparent in use. One might think that there is a snag with this generalization, namely that the interface between the user and, say, the cognitive assistant must be visible to the user. However, this conclusion depends on how one imagines that interface. If one thinks of Kozloski's device on the model of an app such as Siri or Cortana, or a device such as Alexa—that is, on the model of technology with which the user converses in language—then, the visibility of the interface seems much more likely than if we think in terms of information projected into one's visual field by an optical head-mounted display, in the way that Google Glass once promised. Surely, one could incorporate the information transmitted in this way into one's action and reasoning in such a way that one would simply fail to notice its external technological source or its mode of presentation.

Therefore, when skilled tool use is trouble-free, the tools in question may disappear from conscious apprehension. This plausibly applies as much to smart technology that has been smoothly integrated into one's perception-action cycles as it does to an expertly wielded hammer. Unsurprisingly, perhaps, this phenomenological account of how things are for tool users has been adopted in a normative key by some technology designers, and, perhaps, most obviously by some computer interface designers. As Wendt (2013) observes, interface designers have often been attracted by the thought that, when it comes to interfaces, 'invisible' goes hand in hand with 'seamless, efficient and functionally optimized', while 'visible' goes hand in hand with 'cumbersome, inefficient and functionally suboptimal'. Thus transparency is often advanced as a sign of high levels of efficiency and functionality, and, therefore, as something for which designers should aim. As the philosopher Alva Noë remarks: 'You never ask, when confronted with a doorknob, What is this? For the question even to come up is for the doorknob's utility already to have been undermined. If you even notice the knob, it's potentially bad design' (Noë 2015, p. 101). In what follows, I'll argue that, in the case of smart technology that occupies cognition central, we should be suspicious of the urge to design for transparency. To bring this worry into proper view, however, we need first to explore the connection between the transparent tool and the extended mind.

3 Tools are us

As we have seen, in cases of extended cognition, the machinery of mind stretches beyond the skull and skin, in the sense that certain external elements are, like an individual's neurons, genuine constituents of the material realizers of that individual's cognitive states and processes. Put in a somewhat attention-grabbing way, if using your mobile phone counts as an instance of extended cognition (for example, if it is part of your memory, such that, in a sense that would need to be carefully unpacked, you know the phone numbers stored in it), then losing that phone would be equivalent to losing some of your neurons. By contrast, in cases of what is now often called embedded cognition, the machinery of mind remains internal, but the performance of that inner mental machinery is causally scaffolded in significant ways by certain external factors. On this view, the external elements of interest are not genuine constituents of the material realizers of our cognitive states and processes, even though thought and action depend on the causal contributions they make. If using your mobile phone counts as an instance of embedded cognition (for example, if it enables you to fluidly and reliably access phone numbers that are not stored in your memory), then losing that phone might well be disruptive—distressing even—but at least your mental machinery would still be intact. The difference between extended cognition and embedded cognition is the difference between a *North*

Sense device genuinely being part of its wearer's sensory machinery, on a par with her eyes and ears, and a *North Sense* device being 'no more than' a necessary causal factor in producing different inner experiences and differently shaped inner memories.

What is the relationship between the transparency of tools and the truth or otherwise of the extended mind hypothesis (ExM)? Here are two possibilities: (1) given a situation in which an organic human being is using a tool, the transparency of that tool when in use is sufficient for it to be a genuine part of that individual's mental machinery—if the tool is transparent, then what we confront is a case of extended cognition; (2) given a situation in which an organic human being is using a tool, the transparency of that tool when in use is necessary for it to be a genuine part of that individual's mental machinery—if the tool is not transparent, then what we confront is not a case of extended cognition. I do not know of anyone who openly defends (1) in the bald form that I have stated it here. Kiverstein and Farina (2012) arguably treat transparency as sufficient for agent and tool to count as a single system, but additional considerations figure in their justification for treating that system as a whole as a cognitive system. In addition, Wheeler (2005) presents a neo-Heideggerian analysis in which transparency defeasibly indicates that some of the distinctive features of the skilled intelligent behaviour in question must be traced to the causal contributions of beyond-the-skin factors, but, as the distinction between embedded and extended cognition indicates, that is not equivalent to ExM. However, the lack of straightforward advocates for claim (1) need not detain us here, since claim (2) will be enough for our purposes. So what about the claim that transparency is necessary for extended cognition? Where might we find that view making itself heard?

Consider the following passage from Carter et al. (2018, 4): '[I]n normal operation, mind-extending tools should seek to by-pass the epistemic gatekeepers of deliberate, conscious, slow, careful, agentive attention. The best new-you-bits need to join the 'cognitive party' without being constantly stopped at the sensory gates and asked to show their invitations and IDs!' So, Carter et al. claim that for a tool to be genuinely mind-extending (to be a 'best new-you-bit'), it is necessary (as indicated by the use of 'need' in the second sentence), for that tool to be transparent to (to 'by-pass') the deliberative, attentive, reflective conscious mind, the kind of mind that would apprehend it precisely as an independent object. That is how the tool avoids having to show its invitation or ID at the sensory gates! In other words, for Carter et al., transparency is necessary for cognitive extension.

Taking transparency to be necessary for cognitive extension underlies a dynamic version of ExM. Minds grow beyond the skin and shrink back to the boundary of the skin, depending, in part, on the phenomenological dynamics of our couplings with technology. When a tool is transparent, that is a necessary condition met for its constitutive incorporation into the user's mental machinery. When a tool becomes visible, due to, for example, damage or malfunction, or when, as in the case of some sensory substitution subjects, a deliberate, conscious effort on the part of the user resets the mind-world boundary at the skin, that means that cognitive extension is no longer operative.

This dynamic, transparency-driven version of ExM is at odds with those accounts of when an element, biological or otherwise, counts as a genuine part of one's mental machinery that stress persistence rather than transience. For example, Rupert (2009) argues that our theorizing about the mind ought to track a distinction, prevalent in the empirical models produced by cognitive psychologists, between the persisting cognitive architecture, characterized by a relatively fixed set of elements with relatively stable relations among them, and a shifting set of non-cognitive causal factors that sometimes combine with that persistent architecture to produce intelligent behaviour. According to Rupert, if one follows his recommendation, it turns out that ExM is empirically false, since all the genuinely cognitive components turn out to be body-side phenomena (for critical discussion, see, e.g., Clark 2011).

Despite Rupert-style intuitions to the contrary (which here I note merely for completeness), I think that we should be untroubled by the idea that each of us possesses a dynamically growing and shrinking extended mind. That said, the advocate of the transparency-driven version of this idea will need to say something about the following issue. As Clark notes (although in a rather different context):

> 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. (Clark 2015, 8)

Therefore, although, as noted earlier in relation to our sense organs, our own biological machinery is standardly transparent to us in experience, it can be made visible to us. It is certainly possible, then, for our own brains to be consciously apprehended during thought and action. But why should this trouble the transparency-citing fan of ExM? The reason is that proponents of ExM are wont to play an even-handedness card against their neuro-centric opponents (see, most famously, the parity principle, as developed and deployed by Clark and Chalmers 1998). Very roughly, the debate proceeds like this: in the face of the neuro-centrist's claim that the brain is the seat of cognition, the extended mind theorist complains that some external element of

interest is playing exactly the same role or possesses exactly the same property (whatever that role or property may be), in intelligent thought and behaviour, as some internal element, and if the internal element counts as cognitive, then so should the external element. There are, of course, some important questions of detail to be settled here, such as how to decide which roles and properties are relevant and how to determine sameness of contribution (for my responses to these questions, see e.g. Wheeler 2010, 2011a). The core point, however, should be clear enough: the neuro-centrist needs to play fair. But now, if we apply this reasoning in the present context, a problem arises for transparency-based ExM, because even-handedness cuts both ways. If meeting the transparency condition is necessary for an external element to count as cognitive, then, to preserve fairness, it is also necessary for some internal element to count as cognitive. However, as we have just seen, it is possible for one's own brain to fail to meet the transparency condition, so it is possible for one's own brain to become temporarily divested from one's cognitive architecture. And that is a highly counter-intuitive result.

What this seems to tell us is that the fan of ExM cannot combine the even-handedness principle with the claim that the transparency of some external element is necessary for that element to have cognitive status. It looks as if something will have to give, although the mere fact that there is allegedly a tension here does not tell us which of the two thoughts should be rejected, and it is certainly consistent with ExM to hold that transparency is a necessary condition for external resources, but not for internal resources, to count as cognitive. Nevertheless, there might be a different, less disruptive response available. In the case of using bio-feedback to make one's neural states and processes available as objects to consciousness, there is a sense in which those states and processes are simultaneously both transparent and visible. After all, one is still experiencing the world through those states and processes, so the transparency condition is satisfied. It is just that the world thereby revealed contains those very states and processes as objects. From this perspective, there is no tension between the even-handedness principle and the transparency condition.

If the reflections of this section are correct, then, in endeavouring to design transparent technology (or transparent human-technology interfaces), we are endeavouring to meet at least a necessary condition for cognitive extension, that is, for ExM to be true. Therefore, we are at least on the way to designing extended minds. Of course, if we try to design transparent technology, but fail in the attempt, then we have thereby failed to meet a necessary condition for cognitive extension. Moreover, if we were to have a good reason to shy away from designing transparent technology, then we would thereby be refusing to design extended minds. In the next section, I shall argue that, in cases where the technology of interest is equipped with a certain sort of increasingly popular AI, just such a reason exists.

4 Invisible adversaries

Recently, so-called deep neural networks have been all the rage in AI. Such networks typically deploy multi-layered cascades of nonlinear processing units that deploy (supervised or unsupervised) machine learning algorithms to perform pattern analysis and classification tasks, by deriving higher level features from lower level features to build hierarchical representations spanning different levels of abstraction. Such systems have famously learnt to play games to high levels of proficiency, culminating in Google's AlphaGo, a deep learning system for playing the game Go that, in March 2016, recorded a 4-1 victory over Lee Sedol, one of the highest ranked human players in the world.

In influential research, Szegedy et al. (2013) demonstrate that deep neural networks are systematically prone to so-called adversarial exemplars. Let us consider one of Szegedy et al.'s own examples, a network that had successfully learnt to categorize images into two groups—'cars' and 'not cars'. Szegedy et al. proceeded to systematically generate a range of minutely altered images of cars. The deformations were very small changes made at the pixel-level, meaning that, to the unaided human eye, the new images looked identical to other images to which the network had been exposed, and which it had learnt to categorize correctly as cars. The in-advance prediction would surely have been that the network would correctly classify these altered images as cars. However, it did not. With spectacular incorrectness, it classified them as non-cars; hence, the status of those images as adversarial exemplars.

What lessons should we draw? The first thing to stress is that, once in possession of the knowledge that such networks are prone to adversarial exemplars, their designers can, of course, systematically include such exemplars in the networks' training sets. This is clearly a practical response to adversarial exemplars. However, given finite time constraints, there is surely a danger that its effect will sometimes be akin to flattening out a lump under a carpet. There is a tendency for the lump simply to reappear somewhere else. The overarching worry, then, is that deep learning networks are learning to categorize the world in ways that do not coincide with the way that their human users will categorize the world (cf. Carter et al. 2018, 7, who, in a discussion of the consequences of such networks for our epistemic hygiene in light of extended cognition, make the closely related point that deep networks 'learn ways of solving problems that are opaque to their human developers').

Three further points establish the relevance of the foregoing observation to our present concerns. First, as Metz (2016) reports in the technology magazine, *Wired*, deep learning systems are 'already pushing their way into real-world applications. Some help drive services inside Google and other Internet giants, helping to identify faces in photos, recognize commands spoken into smartphones, and so much more'. If deep learning networks systematically classify the world's patterns in ways that are at variance with our ordinary human classifications, and if those networks are lodged in the workings of the technology that organizes and shapes our cognitive lives, then those lives will be organized and shaped by those variant classifications. However, surely we will notice this divergence, I hear you say. It is here that a second point becomes relevant. What if the networks in question are fluidly and expertly integrated into our everyday activities, such that they are transparent in use? Imagine such networks operating as part of a cognitive-assistant-style wearable that classifies situations and transmits the results via an optical head-mounted display. We have already concluded that such behaviour-guiding technology, even though it enhances cognitive performance, and even though it is operating in cognition central, could be transparent. On some occasions, no doubt, its variant classifications of the world would lead to mismatches to which the human user will be sensitive before anything detrimental occurs. However, it seems just as likely that subtle changes in one's engagement with the world—changes that, for example, have potentially damaging social consequences for how one classifies others—might continue to by-pass conscious apprehension, at least until it is too late for the user to take successful remedial action. The final aspect of this worrying scenario comes to light once one realizes that deep neural network applications that meet the transparency condition are at least on the way to being correctly treated as genuine parts of the user's own cognitive architecture. In other words, if ExM is true, there is a natural path to counting such network-driven variant classifications as *our* classifications. This outcome would surely have epistemic implications and perhaps moral ones too. If a deep neural network application to which I am transparently coupled qualifies as part of my cognitive architecture and thus as part of me, then the classifications in question—classifications that unconsciously guide my behaviour—will be part of what I unconsciously believe to be the case, and thus presumably will have the same status as my more familiar, internally realized unconscious beliefs when it comes to any moral judgments that are made about my resulting thoughts and actions.

How should we respond to these somewhat dystopian conclusions? To bring my own proposal into view, I shall all-too-briefly connect the foregoing analysis with certain related themes that arise in the recent work of Mark Hansen (2015). According to Hansen, 21st century media possess the operational signature of by-passing consciousness, that is, they 'go directly to behavioural, biometric and environmental data that are increasingly able to capture our "attention" without any awareness on our part' (ibid. 58). The scare quotes around the word 'attention' here signal the fact that what concerns Hansen is the capacity of modern media to steer decision-making and behaviour independently of any explicit conscious apprehension on the part of the user, thereby delivering a kind of 'digital insight' (ibid. 60). This should sound familiar since, as we have seen, deep learning networks may display the same operational signature. Moreover, Hansen suggests (as have I) that such invisibility to consciousness is impeded neither by the shift from outsourcing to enhancement (what he calls the transition from a 'prosthetic operation of surrogacy [to] the visible inauguration of new, properly technical domains of sensation', ibid. 54), nor by the shift from specialized couplings to more centrally cognitive couplings (which Hansen discusses in the context of a transition from existing medical devices that mediate particular bodily activities to imagined applications that intervene in cognitive decision-making, ibid. 59–60). However, Hansen traces the phenomenological invisibility with which he is concerned not to the skilled, hitch-free use of 21st century media as tools (a claim that connects new media with old media), but rather to the fact that 21st century media distinctively function primarily in networks of machine-to-machine (as opposed to machine-to-human) communication that involve micro-temporal scales and thresholds that are beyond our conscious apprehension. In other words, Hansen's key point is that consciousness necessarily lags behind the operational effects of such media.

It is Hansen's focus on what is necessarily beyond our discriminatory awareness that, I suspect, ultimately leads him to offer a radical response to 21st century media that builds on the work of Whitehead to deliver a reconceptualization of consciousness itself. Here is not the place to investigate Hansen's proposed conceptual shift, but rather to note that although his analysis and mine exhibit the clear parallels just identified, my focus on transparency through undisturbed expert use mandates a response to the threat of epistemically divergent smart technology that is, philosophically speaking at least, more conservative in nature, turning on the idea (more on which in a moment) of making the critical operationality of that technology show up to consciousness as ordinarily conceived. Here, then, is my suggestion: given that our cognitive lives will increasingly be saturated by smart devices that may routinely classify the world in ways at potentially damaging epistemic variance with our technologically unaugmented practices, perhaps, we need to resist the urge to design for transparency. In addition, if transparency is indeed necessary for extended

cognition, then this proposal to resist designing for transparency is simultaneously a proposal to resist designing for extended cognition. What form might such resistance take?

5 The reappearing tool

An alternative to the goal of transparency is identified by the architect Usman Haque in his manifesto for a genuinely interactive kind of intelligent building. Increasingly, architects will be designing buildings that, via permanently installed computational systems, will be able to autonomously modify the spatial and cognitive environments of the people dwelling within them, given what those buildings 'believe' about the needs, goals and desires of the people concerned. As an example, consider an exploratory architectural project called Evolving Sonic Environment, due to Haque and Davis (reported in Haque 2006; see also http://www.haque.co.uk/evolvingsonicenvironment.php, last accessed 14 April 2017). In this structure, people walk around inside an acoustically coupled 'spatialized' neural' network (a spatial web of interconnected simple processing units). The movements of the occupants (detected via sound) affect the organization of the network (the architectural environment) through the operation of local learning algorithms active at each of its nodes. This results in the network adapting over time to different patterns of occupancy, often (in a feature that is prescient in relation to our foregoing discussion of deep neural networks) developing perceptual categories for reflecting those patterns that do not necessarily correspond to categories that the human observer would employ.

In his treatment of such smart architectural environments, Haque argues for a model 'in which people build up their spaces through "conversations" with the environment, where the history of interactions builds new possibilities for sharing goals and sharing outcomes' (Haque 2006, 3). Such conversational interfaces 'would provide us with a method for comparing our conception of spatial conditions with the designed machine's conception of the space' (Haque 2006, 3). Haque's model of an optimally functioning human-AI interface is one of constructive interactive dialogue rather than smooth invisibility. In the interactive mode of encounter, the technology in question is not transparent (it is present precisely as another intelligence), and thus, given the position that we have been investigating, the individual's mental machinery does not extend so as to incorporate that technology.

However, at the end of the day, is an interactive dialogue with our technology what we really want? Perhaps, the pendulum has now swung too far back in the direction of intrusive non-transparency. Indeed, it might seem that whatever it is that we desire from our smart technology, it is not normally the dialogical possibility of misunderstanding and argument. Therefore, maybe, we are looking for a sweet spot in the space of possible interfaces, one located somewhere between the two extremes of transparency and intrusion. This is a speculative thought that requires more unpacking than I can give it here, but maybe an analogy will help. As David Byrne explains in his book *How Music Works*, the goal of the Muzak corporation was to smooth out the curves of workforce inefficiency using specially recorded versions of popular music tracks (Byrne 2012, 326–7). It had been noted in workplace studies of the 1930s that American workers were alert at some times of the day and lacking in energy at others, while the bosses wanted a steady and energetic workflow throughout the day. This is where the Muzak corporation came in. Its strategy was to manage productivity in the workplace by piping in calm music during the high-energy periods of the day and mildly livelier music during the low-energy periods. This was a technique known as 'stimulus progression'. Specially arranged versions of the chosen pieces were recorded, with the dynamics of the music (higher and lower pitches, jumps in volume) flattened. It was soul-less music, hence the disparaging cultural sense of the term 'muzak'. But then, as Bing Muscio of the Muzak Corporation pointed out at the time, it was music to be heard, but not listened to. And that, for us, is the point of the analogy with smart technology. We do not always want our smart technology to be muzak—a transparent factor that manipulates our activities in ways that we do not realize; sometimes, we want our smart technology to be music—not for the most part challenging music that interrogates us and makes us feel uncomfortable (although there is certainly a place for such music and for such technology), but music that solicits listening from us, and so in that way shows up in our conscious experience. In other words, just as music is, for the most part, designed precisely so as to attract our conscious attention, but without its invasion of consciousness being evidence of any breakdown in our engagement with it, so, the proposal goes, the kind of smart technology in which we are interested should be designed with a similar aim in mind, that is, to disrupt transparency without disrupting the skilled use of that technology. That way we can check whether the discriminatory verdicts reached by such technology coincide with our current take on the world. Illuminating models for this phenomenon might be road-signs or directions spoken aloud by mobile map applications, structures that are designed to enter our consciousness to guide navigation without disrupting that skilled activity, but which are thereby available for critical assessment.

If the foregoing reasoning has any force, and I think that it does, then, when it comes to designing the kind of smart technology that we have been considering, we should be

striving for music not muzak. That is what will enable us to restrict the dangers posed by transparent AI-equipped devices that categorize the patterns of the world in ways that depart from the classificatory norms we recognize. However, if we also maintain the position that the phenomenological transparency of a device in use is necessary for extended cognition, then this amounts to the recommendation that, when it comes to such smart technology, we should be designing for a particular kind of embedded mind, rather than an extended mind. To be clear, the conclusion here is not that we should not ever design for transparency or for extended cognition. After all, there will be many devices and applications (especially 'dumb' ones) that do not pose the sort of risk that we have been working to contain. In these cases, arguments which conclude that transparency is a mark of good design may, for all I have said, prevail. Nevertheless, perhaps recent advances in smart technology should encourage us to design for minds that do not always leak into the world, but which sometimes remain in close, productive but visible coupling with it, and thus for devices that sometimes solicit our conscious attention rather than fade from view. The tool, it seems, has reappeared.

Acknowledgements For useful discussion of the ideas presented here, many thanks to seminar audiences in Amsterdam and Dusseldorf, and at Microsoft Research in Cambridge. Helpful comments from two anonymous referees also helped me to improve the paper. Some short passages of text have been adapted with revision from (Wheeler 2011b).

References

Bach-y-Rita P (1972) Brain mechanisms in sensory substitution. Academic Press, New York

Bach-y-Rita P, Kercel S (2002) Sensory substitution and augmentation: incorporating humans-in-the-loop. Intellectica 2(35):287–297

Byrne D (2012) How music works. Canongate, Edinburgh

Carter AJ, Clark A, Palermos SO (2018) New humans? Ethics, trust and the extended mind. In: Carter AJ, Clark A, Kallestrup J, Palermos SO, Pritchard D (eds) Extended epistemology. Oxford University Press, Oxford **(forthcoming)**

Clark A (2003) Natural-born cyborgs: minds, technologies, and the future of human intelligence. Oxford University Press, New York

Clark A (2008) Supersizing the mind: embodiment, action, and cognitive extension. Oxford University Press, New York

Clark A (2011) Finding the mind. Philos Stud 152(3):447–461

Clark A (2015) What "extended me" knows. Synthese. https://doi.org/10.1007/s11229-015-0719-z

Clark A, Chalmers D (1998) The extended mind. Analysis 58(1):7–19

Emslie K (2017) This artificial sixth sense helps humans orient themselves in the world. Smithsonian Magazine. http://www.smithsonianmag.com/innovation/artificial-sixth-sense-helps-humans-orient-themselves-world-180961822/. Accessed 15 March 2017 **(published online)**

Hansen MBN (2015) Feed-forward: on the future of twenty-first-century media. University of Chicago Press, Chicago

Haque U (2006) Architecture, interaction, systems. Extended version of a paper written for Arquitetura and Urbanismo, AU149, Brazil. http://www.haque.co.uk/papers/ArchInterSys.pdf. Accessed 14 April 2017

Heidegger M (1927) Being and time, translated by J. Macquarrie and E. Robinson in 1962. Basil Blackwell, Oxford

Kiverstein J, Farina M (2012) Do sensory substitution devices extend the conscious mind? In: Paglieri F (ed) Consciousness in interaction: the role of the natural and social context in shaping consciousness. John Benjamins, Amsterdam

LaFrance A (2016) A search engine for your memories, The Atlantic. https://www.theatlantic.com/technology/archive/2016/01/sorry-dave-afraid-i-cant-do-that/431559/. Accessed 13 April 2017 **(published online 28/01/2106)**

Menary R (ed) (2010) The extended mind. MIT Press, Cambridge

Merleau-Ponty M (1945) Phenomenology of perception, translated by C. Smith in 1962. Routledge, New York

Metz C (2016) Google's AI wins fifth and final game against go genius Lee Sedol, Wired. https://www.wired.com/2016/03/googles-ai-wins-fifth-final-game-go-genius-lee-sedol/. Accessed 14 April 2017 **(published online 15/03/2016)**

Noë A (2015) Strange tools: art and human nature. Hill and Wang, New York

Ortiz T, Poch J, Santos JM, Requena C, Martínez AM, Ortiz-Terán L, Turrero A, Barcia J, Nogales R, Calvo A, Martínez JM, Córdoba JL, Pascual-Leone A (2011) Recruitment of occipital cortex during sensory substitution training linked to subjective experience of seeing in people with blindness. PLoS One 6:8. https://doi.org/10.1371/journal.pone.0023264

Rupert R (2009) Cognitive systems and the extended mind. Oxford University Press, Oxford

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. arXiv:1312.6199 (**preprint**)

Wendt T (2013) Designing for transparency and the myth of the modern interface, UX Magazine, article No 1077. http://uxmag.com/articles/designing-for-transparency-and-the-myth-of-the-modern-interface. Accessed 14 April 2017 **(published online 26/08/2013)**

Wheeler M (2005) Reconstructing the cognitive world: the next step. MIT Press, Cambridge

Wheeler M (2010) In defense of extended functionalism. In: Menary R (ed) The extended mind. MIT Press, Cambridge, pp 245–270

Wheeler M (2011a) Embodied cognition and the extended mind. In: Garvey J (ed) The Continuum companion to philosophy of mind. Continuum, London, pp 220–238 **(reprinted in paperback version published as The Bloomsbury companion to philosophy of mind (2015), pp 220–238)**

Wheeler M (2011b) Thinking beyond the brain: educating and building from the standpoint of extended cognition, computational culture 1 (on-line journal) **[reprinted in Pasquinelli M (ed) Alleys of your mind: augmented intelligence and its traumas. Meson Press, Lüneburg (2015), pp 85–104]**

Wheeler M (2013) Is cognition embedded or extended? The case of gestures. In: Radman Z (ed) The hand, an organ of the mind: what the manual tells the mental. MIT Press, Cambridge