Draft of paper to be published in *Pistes*, issue #1.

To be published in French as *Entre la transparence et l'intrusion des machines intelligentes* (translation from English to French by Tyler Reigeluth).

Between Transparency and Intrusion in Smart Machines

Michael Wheeler
University of Stirling

Introduction

Smart machines, as I shall use the term here, are devices endowed with artificial intelligence (AI) applications. Some smart machines are examples of autonomous AI, in the sense marked out by attention-grabbing innovations such as self-driving cars and autonomous weapons systems, that is, intelligent technology that, in normal circumstances, operates independently of a human user. By contrast, other smart machines – such as many mobile phone applications and AI-augmented financial decision-making programs – operate in tight coupling with human users within human-technology-world loops.

In what follows, I shall explore what we might call the phenomenological landscape that accompanies smart machines of the second sort. I have navigated parts of this landscape before (Wheeler 2019, 2020) and some of the present treatment revisits many of the same themes and issues, although, in important respects, with revised analyses and arguments. My aims are (i) to map out two phenomena, namely transparency and intrusion (characterized below), that together give shape to the territory in question, (ii) to confront a design-problem – one with significant social implications – that, as I will argue, comes into view, once that shape is established, and (iii) to develop the beginnings of a response to that problem that requires us to find a way of encountering smart machines that is located between transparency and intrusion. Let's begin with the phenomenon of transparency.

Transparency

What does it mean to say that an item of technology is transparent? There are, in fact, at least two legitimate senses of the term 'transparent' that one might deploy in the vicinity of technology. In a first sense of the term, an item of technology is transparent to some highlighted individual or group just when that individual or group is able to understand precisely the inner workings of the device in question. Thus my computer may be transparent to some software or hardware engineer in a way that it is certainly not transparent to me. At a crucial point in our discussion, this notion of transparency (more precisely, its failure to apply to a particular class of cases) will become relevant, but it will not be our primary focus in what follows. Instead, we will mostly be interested in a second sense of transparency, one that is perhaps most readily traced to the work of phenomenological philosophers such as Heidegger (1927), Merleau-Ponty (1945) and more recently Dreyfus (1990). It will take us a little longer to bring this second notion into proper view.

Consider everyday tool-use. The thinkers just mentioned all give essentially the same account of such activity, an account according to which, when a tool is manipulated with skilled expertise and in a hitch-free manner, it *disappears from*, or is *invisible to*, the conscious experience of its user. Put another way, during the unhindered skilled use of a tool, the expert user has no conscious apprehension of that tool as an independent object, that is, as something like the identifiable bearer of determinate states and properties. Rather, what characterizes the experience of the user – what she is aware of during her skilled activity, in an indeterminate and non-thematic way – is the smooth ongoing performance of the task. In effect, the user *sees through* the tool-in-use (through the tool as object) to the task at hand, and that's why it is rightly identified as *transparent*. Heidegger's over-quoted example is that of the skilled carpenter engaged in trouble-free hammering. Although non-thematically aware of her hammering activity, she 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 (Heidegger 1927).

It should be clear enough by now that, although there is undoubtedly much to be said about the relationship between our two senses of transparency, they are not equivalent. On the one hand, a tool or device whose states and mechanisms are understandable to, say, a qualified technician may be fully the target of the user's conscious attention. On the other hand, a tool or device which is invisible in smooth expert use may be wholly impenetrable to its user in its inner workings. From now on, where I mean to refer to the phenomenological concept of transparency, I shall write simply of 'transparency'. Where I mean the fact that some specified individual or group knows precisely what makes a device work as it does, I shall write of 'transparency in the open-to-understanding sense'.

We need to delve deeper into the phenomenon of transparency. Tools are usually thought of as items of technology that allow us to manipulate the world more effectively than would otherwise be possible for us, if we were restricted only to our naked organic machinery. This way of thinking about things is not wrong, but it threatens to obscure an important dimension of the way in which technology intervenes in our lives, one that illuminates further the phenomenon of transparency itself. For technology allows us not only to manipulate the world, but to *access* the world. This is by no means a novel observation. Consider Merleau-Ponty's (1945) example of a blind person using her cane, in a skilled and hitch-free manner, to get around the world. In using her cane in this way, this person will plausibly experience the characteristic transparency of technology when in expert use. Moreover, under these conditions, she may well come to manipulate the world more effectively than would otherwise have been possible for her. The salient point, however, is that the cane enables her to locate things in space, that is, to access her environment. Looked at this way, the cane is just like the biological machinery that, in conditions of proper functioning, constitute a person's organic sense organs. Under such conditions, a sighted individual does not experience her eyes as information-gathering objects. Rather, she experiences the world *through* her eyes. One might put it like this: under normal conditions, a sighted individual's experiential interface is positioned not between her and her eyes, but between her (including her eyes) and the spatial world beyond her eyes. Similarly, under the right conditions – that is, as a result of the cane being smoothly integrated into her perception-action cycles – the blind person's experiential interface is

positioned not between her and her cane, but between her (including her cane) and the spatial world beyond her cane.

Whether this additional focus on accessing the world mandates (i) a broader understanding of what counts as a tool, or perhaps of what counts as manipulation, or alternatively (ii) a switch to thinking more widely in terms of technology in general, rather than merely in terms of tools narrowly conceived, is probably a matter of taste. But however one settles that issue, one might observe that, although hammers and vision-substituting canes are, of course, world-changing in their effects on human lives, the fact remains that, these days, they might be thought of as rather pedestrian examples of the ways in which artefacts may contribute to the ways in which we manipulate and/or access the world. A number of examples will figure in the discussion that follows, but, even just in relation to sensing the world, one might mention technologies that either substitute for non-working organic senses by enabling one sensory modality to support the kind of environmental access and interaction ordinarily supported by a different sensory modality (see e.g. Bach-y-Rita and Kercel 2002) or that endow their users with sensory capabilities beyond our ordinary biological endowment, such as the *North Sense* device that, through intimate physical coupling, allows its 'wearer' to sense magnetic north (product developed by the pioneering technology company *Cyborg Nest*; for an introduction and discussion, see Emslie 2017). In broader terms, one might highlight the myriad ways in which AI is increasingly becoming embedded in our everyday use of technology, performing tasks ranging from Internet search to identifying faces in photos, from recognizing spoken commands to contributing to driving performance through in-vehicle applications, alongside categorization tasks that guide financial decisions and steer medical diagnoses, and many others besides.

At this juncture, one might be tempted to complain that, at least in many cases of such smart machines, the notion of transparency will fail to apply. The thought would go something like this: Although there is some evidence that sensory enhancement technologies might become transparent in use, the processing contributions that these sorts of technologies make within the human-world loop are not of a kind that, at least intuitively, we would count as cognitively sophisticated. Thus, although some user reports of the aforementioned *North Sense* provide evidence of the transparency of that technology in use (see discussion in Wheeler 2019), 'all' the device itself actually does is vibrate gently when its 'wearer' is facing magnetic north, and there's nothing especially surprising about the fact that vibrations on the skin can disappear from conscious apprehension, once the device in question has been in use for a while. By contrast, so the argument continues, once we are focused on the embedding of smart machines in the human-world loop – and thus on applications that perform quite sophisticated feats of categorization, inference and reasoning within that loop – what those machines are doing, judged against our sense of our own workings, qualifies as more intelligent, or (straining the language) more cognitive. And the consequence of this that these devices and applications will not disappear in use. In short, smart machines, unlike hammers, are not transparency-friendly.

In truth, it is hard to know whether the final step in the foregoing reasoning has any force at all. For starters, it is certainly not obvious that just because a technological contribution is, as we might loosely put it, cognitively sophisticated, that simply guarantees that the technology in question cannot become transparent. Consider, once again, the analogy with

the organic brain. If there's one thing cognitive psychology has established, it's that we remain blissfully unaware of many of the operations of categorization, inference and reasoning which are performed by that venerable organ in the ordinary course of our advanced cognitive capacities. So it doesn't stretch things very far to say that, in many examples of hitch-free sophisticated psychological performance, our brains are wholly transparent to us. Of course, our critic might reply by pointing out that such unconscious psychological processes take place on the 'inner' side of the perception-action boundary, whereas the applications on our smart devices are on the 'outer' side of that boundary, meaning that we access them differently, namely with our eyes, ears, and so on. But, as we know from the disappearing hammer, simply being on the outer side of that boundary does not, in and of itself, establish a barrier to transparency.

Obviously these are only the first moves in what could turn out to be a long and drawn-out debate, but, for now anyway, it seems at least to be an open question whether any smart technology is somehow necessarily resistant to the phenomenon of transparency. And that leaves in place the possibility that at least some examples of smart technology will positively invite such transparency, when used smoothly and with appropriate skill. And that's enough for us, since it makes transparency in smart machines a phenomenon worth investigating, especially given certain concerns associated with that possibility that I am going to identify later in this paper. Before we turn to those concerns, however, we need to complete our tour of the relevant phenomenological landscape.

From Transparency to Intrusion

It probably goes without saying that, when it comes to the spectrum of ways in which we can experience technology, then whether we are talking about dumb tools or smart machines, transparency certainly isn't all there is. Here it is worth touching base again with Heidegger.

Famously, Heidegger had a term for (roughly) the transparency of tools-in-use. He called it *readiness-to-hand*. And juxtaposed with readiness-to-hand are two further ways in which we encounter entities in our experience, namely *un-readiness-to-hand* and *presence-at-hand*. Taking the latter first, entities emerge as present-at-hand when, for example, they are the targets of our natural sciences, or when our sensing of them takes place purely in the service of reflective or philosophical contemplation. In such cases, the entities in question are no longer folded into our practical activity and they are no longer transparent. They figure in our detached conscious apprehension as fully independent objects, that is, as the bearers of certain context-general determinate or measurable properties (size in metres, weight in kilos etc.). *In between* readiness-to-hand and presence-at-hand comes un-readiness-to-hand, a phenomenon which occurs when skilled practical activity is disturbed by, for example, the breaking or malfunctioning of a tool. Like present-at-hand entities, un-ready-to-hand entities are no longer transparent. However, they are not yet fully present-at-hand. A driver does not encounter a punctured tyre as a lump of rubber of measurable mass, but rather as a damaged piece of equipment. In fact, the category of un-readiness-to-hand seems to involve a range of cases, stretching from examples in which we are barely aware of the disturbance and almost effortlessly adjust our skilled behavior accordingly, which leaves us near the border with the readiness-to-hand, to examples in which we are

forced to deploy explicit theoretical knowledge of how tools and devices work in order to restore their functioning, which leaves us near the border with presence-at-hand. Thus compare a case in which an expert driver adjusts her driving activity to compensate for a mildly under-performing clutch to one in which a qualified and experienced mechanic is required to diagnose a recalcitrant reoccurring engine problem. One might think of the spectrum of cases traversed by un-readiness-to-hand as one characterized by increasing levels of *intrusion* into our experience. Roughly speaking, the more intrusive the technology is, the more we have to think about it.

So what? In areas such as user interface design, transparency has often been adopted as a mark of good design, that is, as something for which designers should aim. By contrast, intrusion, as exemplified by the more disruptive end of un-readiness-to-hand, is often thought of as something to be avoided where possible, that is, as a mark of poor design. As just two examples, this normative framework is plausibly at work behind the title of Krug's influential book on web design called *Don't Make Me Think* (Krug 2005) and, with a more general application, behind the following remark from the philosopher Alva Noë: '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, 101). One thing that this indicates is that, as a phenomenon, intrusion is not straightforwardly equivalent to being broken or to malfunctioning. A device that is designed in such a way that it intrudes on my consciousness during use may not be broken or be not functioning as it was designed to. Nevertheless, Heidegger's appeal to the broken and the malfunctioning to help scope out the domain of the un-ready-to-hand is suggestive of the idea that where intrusion reigns, things could be going better. And, just as the improvement that could be delivered in fixing a broken item of technology would, in Heidegger's framework, be marked by the (re-)establishing of transparency, so the same phenomenological transition would mark an improvement in the design of the properly functioning but intrusive device. To re-use Noë's example, we have designed a better doorknob when we have eliminated intrusiveness.

There's no doubt that the sort of reasoning just rehearsed has a certain momentum to it. However, at least when we turn our attention to smart machines – to technology in which AI has been embedded – there is reason to think that it might need to be questioned, and perhaps even stoutly resisted. To see this, consider a class of AI systems known as deep learning networks. Deep learning networks deploy multi-layered cascades of nonlinear processing units alongside (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 networks are increasingly prevalent in all sorts of contexts, from face and speech recognition systems to more openly life-critical applications for detecting earthquakes, predicting heart disease and controlling driverless vehicles.

As Szegedy et al. (2013) have demonstrated, deep learning neural networks are systematically prone to so-called *adversarial exemplars*. Consider one of Szegedy et al.'s own examples, a network that had successfully learnt to sort a set of images into images of cars and images of not-cars. What the researchers did next was to generate some minutely altered images of cars from the training set. The deformations were so small that, to the

unaided human eye, the new images looked identical to images on which the network had been trained. One might have expected the network to classify these altered images 'correctly', as images of cars. Surprisingly, however, it classified them as non-cars, hence their adversarial status. Of course, armed with the knowledge that adversarial exemplars exist, designers can systematically generate such items and include them in their networks' training sets. But, especially given finite time constraints, there is surely a strong likelihood that other adversarial exemplars will persist in the operational space, even after the retraining process.

For our purposes, the general message to be extracted from Szegedy et al.'s work is this: deep learning networks will sometimes divide up the world in ways that do not coincide with our ways of dividing up the world. This is already a prompt for reflection, in that a capacity for reliable categorization – more specifically, the consistent partitioning of the world into morally-relevant and action-relevant categories (e.g. pedestrians and non-pedestrians, disease-indicating phenomena and non-disease-indicating phenomena) – is at the heart of what we are increasingly asking our smart machines to do for us. In the case of autonomous AI (e.g. self-driving cars with no ongoing routine human monitoring), this poses directly the question of whether we should in fact be comfortable in ceding control to such systems in the strong sense of allowing them to operate in a genuinely independent fashion (for discussion, see e.g. Wheeler 2020). But the issue that concerns the present treatment comes into view if we take seriously the possibility (and I have argued earlier that we *should* take it seriously) that smart machines may often be transparent to us, as a result of (i) AI applications, such as deep learning networks, being embedded in our personal technology (wearables, smartphones and so on) and then (ii) the operations of those applications in the device in question being folded seamlessly into our ongoing activity in a skillfully deployed, hitch-free manner.

To make things more concrete, consider the iCart, an augmented shopping trolley with a display built into its handle that successfully changed supermarket shopper behaviour by providing the user with simple nutritional information, namely which of three health categories a selected product fell into (Katsikopoulos and Fasolo 2006). In fact, let's consider a hypothetical next-generation iCart, a deep learning version of the original technology, one that had learnt its classifications in a training regime. On the basis of the aforementioned research into adversarial exemplars in such networks, it seems very likely that our smart machine, working in the wild, will sometimes make classifications that diverge from user judgments. We can assume that the nutritional information is provided visually by a simple colour-coding system. This would facilitate smooth integration into the user's perception-action cycle and thereby establish conditions of transparency. Shopper behaviour would thus end up being affected by the network's divergent classifications and, crucially, the user wouldn't notice.

The general lesson, then, is that our activity may be nudged, influenced and maybe, by extension, wholly determined by the divergent classifications of a deep learning AI, in ways that won't be registered in our deliberative consciousness, or at least not until it is too late for remedial action. Of course, thinking about this in relation to the next-generation iCart example won't chill anyone to the bone. The shopper might go home with a few unhealthy snacks that she believes have nutritional benefits, but that's all. However, given the kinds of

morally charged scenarios in which we are increasingly putting AI to work alongside human users, it is but a short step to more worrying cases.

It might seem that there is an obvious corrective to the direction of travel here, which is to give up the identification of transparency with good design. But that is easier said than done, at least for some contexts in which we expect smart machines to support cognitive activity. To see why, we can draw on recent work by Reicherts and Rogers (2020). The *Tableau* application is a data analytics tool that enables lay users (those who are not experts in the domain from which the data is drawn) to conduct analyses of complex data sets by generating visualizations of that data, and to do so in just a few clicks. Given the easy-to-navigate interface, it is plausible that *Tableau* is transparency friendly. So far so good for the standard view. However, a problem surfaces when one takes account of the fact that Tableau actually enables users to generate any type of visualization that is consistent with the kind of data under analysis. This poses the question of how the lay user is supposed to know which specific type of visualization would be the most illuminating in some particular case. Using the 'Wizard of Oz' paradigm in human-computer interaction (where subjects believe they are interacting with an autonomous device but in fact are interacting with a hidden human being), Reicherts and Rogers introduced a natural language conversational user interface (CUI) that explicitly prompted and guided lay users regarding their choice of visualization. The general conclusion drawn by Reicherts and Rogers is that graphical user interfaces (like the standard *Tableau* interface) are appropriate when users know what they are looking for or are familiar with the task, whereas CUIs are appropriate when analytical or problem-solving domains are, as they put it, 'new to the user, ill-structured, open-ended or exploratory' (Reicherts and Rogers 2020, 1).

In the context of the present investigation, what does this tell us? It is plausible that the CUI developed by Reicherts and Rogers ought to be categorized as phenomenologically intrusive. Indeed, one might think that it works precisely by disrupting the smooth clicking of the user. Understood this way, their study suggests that there are problem-solving domains in which good design may be achieved through intrusiveness. Equally, however, it suggests that there are many problem-solving domains in which, ignoring the worries about divergent classifications, transparency remains the experiential gold standard for which to aim. This presents a problem once one deploys deep learning networks in the latter class of domains, because it invites a perfect storm of divergent classifications and transparency. Of course, anyone moved by such worries may be tempted to stamp their foot about now and demand that we make the relevant interfaces intrusive, precisely because of the threat of divergent classifications. However, the indisputable and impressive success of deep learning AI, which routinely outperforms human experts in many domains (just ask Lee Sedol; see Metz 2016), threatens to render such protests largely ineffectual. And, as a final sting in the tale, this is the point at which the alternative sense of transparency highlighted earlier in this paper – transparency in the open-to-understanding sense – becomes relevant. It is a recognized fact that although we understand the learning algorithms that enable deep learning networks to be trained to classify data in powerful ways, there are many cases In which the reasons why the trained-up networks make the decisions that they do (that is, the principles by which they make their in-use classifications of data) remains opaque to us. In other words, they are definitely not transparent in the open-to-understanding sense. Hence all the recent fuss about so-called explainable AI. Thus, as things stand anyway, it is far from

obvious that, for the relevant class of systems, we could in practice deliver an interface that was intrusive in the specific sense of making the system's own reasoning processes explicit and thus available to our deliberative consciousness for processes of reflective evaluation and critique.

Putting that final concern to one side, here is a suggestion. We have implicitly been assuming that we are working with a dichotomy between transparency and intrusiveness, such that our intelligent devices need to be in one category or the other. But that is not the situation in which, as we saw earlier, Heidegger leaves us, following his analysis of the phenomenological space in question. Between transparency (readiness-to-hand) and intrusiveness (the disruptive end of un-readiness-to-hand) come cases that solicit our conscious attention without this being accompanied by any breakdown in our worldly engagement (the non-disruptive end of un-readiness-to-hand). What is distinctive of these in-between cases is that transparency is removed without that removal resulting in any disruption – or at least any significant disruption – to the skilled use of the technology in question. And maybe that's what we should be aiming for, at least sometimes, in our design of smart devices. Call this the *sweet spot* between transparency and intrusiveness. (The term 'sweet spot' is adapted from Kalnikaite et al 2013; more on their use of the term in a moment.) If we can find this sweet spot, we would have devices whose presence and processing explicitly enter our consciousness, and which thereby become poised for critical assessment, in the sense of enabling us to monitor whether the discriminatory verdicts they reached coincide with our current take on the world. Nevertheless, they would continue to scaffold ongoing skilled activity. Admittedly, this idea remains speculative and under-specified, but in the next and final section of this paper I shall present some considerations that, I suggest, constitute some provisional hooks on which a more developed analysis could perhaps be hung.

Searching for the Sweet Spot

Yet again I shall take my cues from Heidegger (1927), and more specifically from his discussion of how hitch-free skillful activity may be mediated by *signs*. According to Heidegger, signs have the function of indicating, and in doing so they modulate the practical interaction of an agent with its environment. One of Heidegger's examples is the south wind (see Heidegger 1927, 111-2). When an agent approaches the south wind with a particular practical attitude, it emerges as a very specific kind of equipment, one with the function of indicating the possibility of rain. But, as Heidegger puts it, as a sign it is also 'an item of equipment which explicitly raises a totality of equipment into our circumspection so that together with it the worldly character of the ready-to-hand announces itself' (Heidegger 1927, 110). Roughly speaking, and without the Heideggerian language, the south-wind-as-sign does not simply tell the meteorologically savvy agent that it will rain soon; it also makes her aware that she must actively use her other competences in order to cope with that change in the weather. The south wind is of course a naturally occurring sign, but the same analysis applies to signs that we design, such as road-signs. Given a road-sign-savvy driver trying to get to Edinburgh Castle, a sign which tells her that the castle is down a road to the right not only makes her aware of that fact; it also makes her aware that she should actively use her other driving competences in order to turn into that road.

For our purposes, what is crucial about all this is that, for Heidegger, signs are phenomenologically special structures that enter our consciousness in order to scaffold and guide skilled activity *without* disrupting that activity, and indeed are standardly designed precisely so as to do that. Moreover, in relation to how such artefacts figure in our experience, it seems that, although they are not intrusive, they are nevertheless poised for critical assessment. *If* what the road sign indicates clashes with the driver's existing way of dividing up Edinburgh into castle and non-castle regions, there is at least the potential for that conflict to be available to her consciousness, for assessment, reasoning and appropriate action. In summary, signs are neither transparent nor intrusive, but rather are located between these two categories in a way that renders them poised for critical assessment, and that's precisely the kind of phenomenon we are endeavouring to expose. Thus, designing our smart machines to have the phenomenological profile of signs suggests a way to solve the problem of divergent classifications.

Of course, the foregoing, rather abstract reflections, raise the question of just how to design smart machines so that they possess the desirable profile just sketched. Here is an example that, I think, points us in the right direction. We are going back into the supermarket. Kalnikaite et al (2013) observe that, for familiar, low-cost purchases, supermarket shoppers typically apply 'fast and frugal' heuristics, short-cut strategies that focus on a limited range of information. By contrast, mobile phone shopping applications that are designed to augment the shopping experience tend to overload shoppers with vast amounts of product information. This overload tends to be disruptive. What designers need to find, according to Kalnikaite et al, is 'the information sweet spot that technological aids should aim for so as to enhance the shopping experiences'. To make headway into this space, that should look familiar, they designed a device that clips onto the shopping trolley and which supplies 'at-a-glance', barcode-driven LED and emoticon-based real-time feedback on different brands of the same product. This feedback supplies information about: (a) food mileage, indicated by the number of lit LEDs; (b) whether the brand is organic, indicated by the colour of the LEDs; and (c) how the shopper's choices compare to some relevant social norm (other people's behaviour) regarding e.g. food miles, indicated by an emoticon. The device includes a barcode scanner so that the relevant operations are 'hands free'. Given our concerns, the key results of the study are that shoppers found the device intuitive to use, commenting that it did not disturb their shopping activity, although it did slow them down. It seems plausible that this slowing down effect is evidence that the device was appropriately poised for critical assessment, and indeed the results show that when the emoticon display was neutral or negative, shoppers scanned and checked the mileage on more product alternatives than when the emotion was positive. What this data tells us, I think, is that the device in question is neither transparent nor intrusive, but rather is positioned between these two categories, with its operations poised for critical assessment. In other words, it has the phenomenological profile of a sign. We can of course easily imagine that deep learning is used in the design – say, in the fine-tuning to individual user preferences – of a system such as this. In such a scenario, and given what we have just learned, the sign-like phenomenological status of the device may allow designers to address the problem of divergent classifications.

Smart machines are with us and they are here to stay. If the foregoing analysis is on the right track, that means that the problem of divergent classifications is waiting in the wings,

but it also means that maybe, through a canny combination of philosophical reflection and ingenious design, that problem can be prevented from making a disquieting entrance onto the stage.

Acknowledgments

For useful discussion of the ideas presented here, many thanks to audiences in Grenoble and Stirling, and at Microsoft Research in Cambridge. Some short passages of text have been adapted with revision from (Cappuccio and Wheeler 2010, Wheeler 2019, 2020).

References

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

Cappuccio, M., and Wheeler, M., "When the twain meet: could the study of mind be a meeting of minds?", in J. Reynolds, J. Chase, J. Williams and E. Mares (eds.), *Postanalytic and metacontinental: crossing philosophical divides*, Continuum Studies in Philosophy, London, Continuum, 2010, p.125-144.

Dreyfus, H. L., *Being-in-the-world: a commentary on Heidegger's Being and Time, Division I*, Cambridge, Mass., MIT Press, 1990.

Emslie, K., "This artificial sixth sense helps humans orient themselves in the world", *Smithsonian Magazine*, published online, 2017, http://www.smithsonianmag.com/innovation/artificial-sixth-sense-helps-humans-orient-themselves-world-180961822/, last accessed 04/02/2020.

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

Katsikopoulos K.V. and Fasolo. B., "New tools for decision analysts", *IEEE Transactions on Systems, Man and Cybernetics A*, 36, 2006, p.960–967.

Kalnikaitė, V., Bird, J. and Rogers, Y., "Decision-making in the aisles: informing, overwhelming or nudging supermarket shoppers?", *Personal and Ubiquitous Computing*, 17, 2013, p.1247–1259.

Krug, S., *Don't make me think: a common sense approach to the web (2nd edition)*, USA, New Riders Publishing, 2005.

Merleau-Ponty, M., *Phenomenology of perception*, translated by C. Smith in 1962, New York and London, Routledge, 1945.

Metz, C., “Google’s AI wins fifth and final game against go genius Lee Sedol”, *Wired*, published online 15/03/2016. https://www.wired.com/2016/03/googles-ai-wins-fifth-final-game-go-genius-lee-sedol/, last accessed 04/02/2020.

Noë, A., *Strange tools: Art and human nature*, New York, Hill and Wang, 2015.

Reicherts, L. and Rogers Y. “Do make me think! How CUIs can support cognitive processes”, *Proceedings of CUI '20: 2nd Conference on Conversational User Interfaces*. 2020. DOI: 10.1145/3405755.3406157

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R., “Intriguing Properties of Neural Networks”, *arXiv preprint arXiv:1312.6199*, 2013.

Wheeler, M., “The reappearing tool: transparency, smart technology, and the extended mind”, *AI and Society*, 34, 2019, p.857–866.

Wheeler. M., “Autonomy”, in M. D. Dubber, F. Pasquale and S. Das (eds.), *Oxford handbook of ethics of AI*, Oxford, OUP, 2020, p.343-358.