This is a penultimate draft. Please cite the final version forthcoming in Philosophy and Technology.

Should we discourage AI extension?

Abstract

We might worry that our seamless reliance on AI systems makes us prone to adopting the strange errors that these systems commit. One proposed solution is to design AI systems so that they are not phenomenally transparent to their users. This stops cognitive extension and the automatic uptake of errors. Although we acknowledge that some aspects of AI extension are concerning, we can address these concerns without discouraging transparent employment altogether. First, we believe that the potential danger should be put into perspective – many unreliable technologies are unlikely to be used transparently precisely because they are unreliable. Second, an agent who transparently employs a resource may also reflect (opaquely) on its reliability. Finally, agents can rely on a process transparently and be yanked out of their transparent use when it turns unreliable. When an agent is responsive to the reliability of their process in this way, they have epistemically integrated it, and the beliefs they form with it are formed responsibly. This prevents the agent from automatically incorporating problematic beliefs. Responsible (and transparent) use of AI resources – and consequently responsible AI extension – is hence possible. We end the paper with several design and policy recommendations that encourage epistemic integration of AI-involving belief-forming processes.

Keywords: phenomenal transparency, artificial intelligence, cognitive extension, adversarial attack, cognitive integration

1 Introduction

Some have argued that our fluent and seamless reliance on AI systems may cause us to incorporate into our cognitive systems the strange errors these systems commit (Carter et al. 2018; Wheeler 2019, 2021; Hernández-Orallo and Vold 2019). One such error is committed by deep neural networks (DNN), which may succumb to adversarial exemplars and misidentify what to us is clearly an instance of one object (e.g. a car) as something else (e.g. a non-car) (Szegedy et al. 2014).

According to the thesis of extended cognition (Clark and Chalmers 1998; Menary 2010), our cognitive states may sometimes be realised at least partially outside our bodies,

for instance in notebooks, phones, and other devices. When our cognitive processes are AI-extended, the AI's faults threaten to become our own.

One proposed solution to this problem, owed to Michael Wheeler, is to design Deep Neural Network-based AI systems such that they are not employed transparently (Wheeler 2019, 2021). The concept of transparency used here is borrowed from the phenomenological literature and describes the experience of skilful fluent tool use. When we use a tool (say, a hammer) in such a way, it may disappear from our focus of attention so that we are focusing directly on the task at hand (say, hammering a nail).[1] Several proponents of the extended cognition thesis have argued that phenomenal transparency is a necessary condition for cognitive extension (Clark 2003, 2008; Thompson and Stapleton 2009; Wheeler 2019).

Transparently employing a process causes an AI-extended process to disappear from the agent's experience. This seems to imply that the agent will not be able to think about the properties of the resource and that the resource's faults are therefore likely to escape notice. When that happens, the agent is left without defences against problematic (extended) beliefs. By advocating the *in*transparent (or opaque) use of AI technologies, Wheeler aims to prevent cognitive extension and, consequently, the incorporation of the AI's faults.[2]

This paper argues that we can learn to responsibly extend into AI systems. More specifically, we first show that the problem at hand is less devastating than Wheeler thought. Many technologies are so unreliable that they are unlikely to be employed transparently or at all. Secondly, agents can reflect on the processes that they transparently employ, and focusing on designing *in*transparent AI is therefore unnecessary. Using the virtue reliabilist concept of epistemic integration, we show how agents can rely on external resources even when they employ them transparently. Epistemic integration, we argue, allows us to understand how we can responsibly extend into AI systems.

In addition, we apply our understanding of epistemic integration to propose design and policy recommendations for responsible AI extension. These recommendations aim to make it more likely that agents become aware when their extended processes fail. We think many of these strategies apply also to Wheeler's examples concerning adversarial

[1] A second concept of transparency is used in discussions around AI to refer to the degree to which agents are able to understand how a technology works. The main subject of this paper is phenomenal transparency as we will describe and illustrate in the next section. See Andrada, Clowes, and Smart (2022) for an introduction to the varieties of transparency.

[2] Before we present our arguments, one qualification about Wheeler's recommendation is important. In Wheeler (2019), he focuses on discouraging extension to DNN networks because they can succumb to adversarial attacks. In Wheeler (2021), he recommends the same first, but then seems to pivot to a more nuanced stance according to which we should look for a sweet spot between transparency and intrusion. While the final position isn't entirely clear, it is clear that Wheeler doesn't give an account of how we could find such a sweet spot. Our paper address this question head-on.

exemplars, lessening the force of his argument.

Our paper is structured in the following way. After this first section, section 2 outlines the concept of transparency and shows that many unreliable technologies are not employed transparently. In section 3, we argue that transparency and opacity are not mutually exclusive. Section 4 introduces two routes to epistemic integration and describes how we can extend into AI systems responsibly. Section 5 suggests some design and policy recommendations to encourage the epistemic integration of AI systems. In section 6, we apply our approach to adversarial attacks.

2 Transparency and reliability

We must first take a closer look at the concept of transparency. This will reveal that many unreliable processes are not used transparently (or at all) and that there are, therefore, far fewer cases of problematic extension than we might think. Moreover, it's possible to use unreliable technologies in reliable ways.

Imagine EarSpeak, a wearable computer vision technology. It comes with an earpiece with which an AI calls out the names of objects located in front of you. It can even describe entire scenes. Saira, who is blind, has been using EarSpeak for a year. She is no longer aware of EarSpeak describing the world to her; if you asked her, she'd say that she's simply aware that there is, for instance, a door in front of her.

When a resource is employed transparently, it disappears from the focus of attention. Saira doesn't need to focus on EarSpeak to form beliefs with it. She doesn't believe that there's a door in front of her because she hears a certain utterance which she then interprets using her knowledge of EarSpeak's design. Rather, she focuses directly on the task of navigating her environment – for instance, of opening and passing through doors. EarSpeak is transparent to Saira: she 'sees' the world with the technology rather than focusing on it.

Transparency involves the skilled use of a resource (Heidegger 1976; Merleau-Ponty and Landes 2012; Dreyfus and Dreyfus 1988; Clowes 2019; Andrada 2020). When Saira had just received her EarSpeak, the sudden streams of language issuing forth from her device confused and overstimulated her. She had to pause to catch up with what was said and then form beliefs about the structure of the visual scene. This is similar to the well-known example of the carpenter and her hammer (Heidegger 1976). A novice carpenter will need to focus on the hammer, its weight and its shape, and carefully handle it to strike the nail in the right way. The master carpenter, in contrast, doesn't think about the hammer – she looks at the nail and strikes it.

When an agent skilfully and transparently employs a resource, she is attuned to how

it helps achieve her goals. The master carpenter has learnt that moving the hammer so-and-so reliably sinks the nail into the wood. She knows that moving the hand in a certain way will cause the nail to sit nicely flush with the wood. She doesn't need to rely on any thoughts about the hammer's properties (its shape or weight) to figure out how to move her arm to achieve her goal (Grush and Springle 2019).

Wheeler thinks employing AI technologies transparently is problematic and therefore should be discouraged (Wheeler 2019, 2021). When a resource is transparent to the agent, its properties aren't at the focus of attention and, therefore, the agent cannot easily detect when things go wrong. In contrast, when we use AI technologies opaquely, we can bring to bear the entirety of our conscious cognitive processing on our interaction with them, and can therefore more easily and reliably detect when things go awry. Technologies become objects we attentively interact with, a problem to be solved rather than a part of the machinery with which we solve problems (Clark 2003, 2008).

Suppose when EarSpeak first came out, it failed at detecting doors. It would remain silent instead of warning users of the closed doors in their paths, and so they bumped into them constantly. As a result, many users continued using their canes to check for objects in front of them.

Amna was one of the users of this early iteration of EarSpeak. She was very annoyed about never knowing whether to expect a door when EarSpeak remained silent. She had to constantly pay attention to her surroundings to gauge if EarSpeak's silence could reasonably mean that no objects were ahead. This required her to pay constant attention to her use of EarSpeak, making transparent use impossible. Amna decided to return EarSpeak.

A large class of unreliable AI technology is unlikely to be used transparently as their use requires constant attention to avoid costly mistakes. Agents either use such technologies opaquely or not at all. Since transparency is necessary for cognitive extension, there cannot be extension in such cases, and without extension, we don't have to worry about odd AI errors becoming a part of our cognition.

However, this is not to say that it's impossible for a user to transparently employ a somewhat unreliable technology. A reliable technology is different from a reliable cognitive process and the latter doesn't necessitate the former. Imagine Amna is told she can't return EarSpeak. As she doesn't want to write off the expensive purchase, she decides to try using the device a little longer. After all, she figures, it shouldn't be so hard to only listen to EarSpeak when it says something, but abstain from drawing inferences when it stays silent. Sure, that would limit EarSpeak's usefulness somewhat, but at least it's an easy rule to follow. Unexpectedly, as time passes, Amna expends less and less effort to detect the situations in which EarSpeak is reliable until, eventually, she starts using the technology transparently.

An agent who can distinguish problematic from unproblematic inputs – and can do so without conscious processing – may learn to reliably and transparently employ a (partially) unreliable process. In this way, a (partially) unreliable technology extends her reliable cognitive process. Amna achieves this when she learns to seamlessly ignore EarSpeak's silence.

Thus, there are at least two ways in which unreliable technologies can be unproblematic. First, unreliable technologies are often unlikely to be employed transparently, making extension impossible. Second, agents may come to only rely on those aspects of a (partially unreliable) technology that are reliable. In both cases, erroneous information from AI technologies is not automatically accepted.

3 Transparent and yet opaque

It's not always possible to constrain one's transparent employment of a resource to only those aspects that are reliable – however, as we'll see, this is also not required. We may employ a resource transparently and still have defences against problematic inputs. First, the transparent employment of a resource doesn't make opacity impossible, and agents can inspect resources they transparently employ. Additionally, agents that transparently employ resources can become aware when their processes malfunction. Thus, we can responsibly employ AI technologies even while transparently relying on them.

Agents may not always be able to constrain their transparent use of a technology to only those aspects which are reliable. First, AI technologies may fail rarely or subtly. Suppose EarSpeak develops a new fault, this time involving the identification of tables. But rather than always failing to identify them, it only fails once every 100th time. Or, alternatively, it merely misidentifies tables as desks. When an agent isn't exposed to sufficiently many clear instances of failure, making out dependable patterns becomes challenging.

Second, AI technologies may fail randomly or in ways that appear random to the user. Maybe EarSpeak fails due to the random fluctuation in some internal component or only when some complex set of factors play into one another in a specific way. Here, too, finding a pattern in the failures may be challenging (or impossible) for the agent. And without such a pattern, the agent will be hard-pressed to recognise when not to use the technology. In such cases, there's a risk for agents to transparently employ unreliable technologies.

Saira has been using her reliable EarSpeak device for many months and has come to transparently rely on it. While things have been going well so far, this is now changing as her device has developed one of the hard-to-detect problems mentioned above and doesn't reliably identify tables anymore.

Is Saira left without defences to this change in reliability? Not necessarily. Let's sup-

pose Saira is a computer engineer and has access to the device's code and data. She studies the device and realises why she's been bumping into tables: EarSpeak is malfunctioning (while her biological faculties are working fine).

The case shows that we can use a resource transparently and still focus our attention on it – either *while* employing it transparently or *at another time*. In the latter case, the agent stops or pauses her transparent use of the technology to take a look at how the technology works. In the former case, transparent use of the resource never stops – in fact, we may even use some technology to examine the very same technology. Saira may, for instance, use EarSpeak to look at the visual output of some analysis software to gain information *about* EarSpeak.

Saira may also use EarSpeak transparently until a malfunction makes the device stand out and catch her attention. Imagine Saira is standing in her kitchen, where she knows a table is located. She is transparently using EarSpeak, and EarSpeak is silent about the table in front of her. Since Saira knows that there's a table in front of her, she may become aware that EarSpeak is malfunctioning. In such situations, the device can become opaque as a tool that has failed (Heidegger 1976; Wheeler 2021). This is analogous to the example of the master carpenter whose hammer fails. When that happens, the carpenter's attention is drawn to the hammer, which now becomes a problem to be solved.

We can become aware of failures even when we transparently employ some resource – no matter whether that resource is internal or external to the body. Imagine you're walking down a dark city street and catch a glimpse of a lion lurking in the shadows. You know the object you're seeing cannot be a lion – after all, you're in the middle of the city – and you therefore dismiss the belief outright. When that happens your visual processes, which you normally employ transparently, become opaque. They make themselves known as something that may fail (in particular, that may fail when it's too dark).[3]

Making sure that a transparently employed resource is used so that it becomes opaque when it goes wrong is reminiscent of a specific kind of cognitive integration: *epistemic integration*.

4 Epistemic integration

The virtue reliabilist concept of cognitive integration (hereafter, epistemic or e-integration) posits that when we responsibly (and transparently) rely on a process, we are in the position

[3] See Andrada (2020), Smart, Andrada, and Clowes (2022), and Facchin (2022), for a discussion of transparency, cognitive extension and how we employ our internal cognitive faculties. See also Clark (2022) for a predictive processing-based subpersonal account that explains how we transparently employ processes that we can reflect on.

to become aware when it stops functioning reliably. On this view, we can counter Wheeler's worry by demanding that agents e-integrate AI technologies. Then, instead of internalising AI's errors, agents become aware when their extended processes stop being reliable.

In short, it's by e-integration that processes (habits of inquiry, skilled uses of technologies, and so forth) become a part of our cognitive systems in such a way that we can responsibly employ them to form beliefs (and thus potentially acquire knowledge) (Greco 1999, 2010). And only in the case of e-integration can we speak of *responsible* extension (Pritchard 2010; Palermos 2011, 2014). So, if Saira e-integrates her EarSpeak process, she can responsibly form beliefs with it and therefore extend responsibly.

Responsible cognitive extension is different from mere cognitive extension. For cognitive extension, metaphysical integration (hereafter, m-integration) suffices. M-integration only demands reciprocal and continuous interaction between the new process and other processes in the agent's cognitive system(Carter and Kallestrup 2020).[4] Responsible extension, in contrast, requires e-integration, which demands that the agent be sensitive to the reliability of her integrated process.

E-integration depends on new beliefs cohering with existing beliefs, and therefore only *belief-forming processes* e-integrate. This means, first, that what e-integrates is a *process*. Thus, as we've already said previously, Saira doesn't e-integrate EarSpeak (the technology) but her process of using it. Second, the process needs to be *belief-forming*. Some cognitive processes are belief-forming, but not all, and only belief-forming ones can e-integrate.

Further, only *reliable* belief-forming processes can e-integrate. A reliable belief-forming process is one that forms far more true beliefs than false ones (Palermos 2021). In contrast, m-integration (and therefore mere cognitive extension) doesn't require that one's process be reliable.

There are two routes to e-integration, the strong and the weak (from here on, we will simply call these strong and weak e-integration) (Pritchard 2010).[5] Simply put, the differ-

[4] Following Carter and Kallestrup (2020), we use the term epistemic integration (shortened to e-integration) and contrast it with metaphysical integration or mere cognitive extension. The early literature, for instance Greco (2010) and Pritchard (2010), doesn't distinguish between these two varieties of integration and simply uses the term cognitive integration for epistemic integration.

[5] Even if inspired by Pritchard's (2010) account of strong integration, our strong route to e-integration must not be confused with it. His footnotes 8 and 9 suggest that he is undecided whether an agent must employ her new process for some period of time to build a web of interrelations between it and her pre-existing processes or whether having access to the process's reliability suffices. We think that using the process and forming beliefs with it is necessary (on top of being aware that it's reliable) for strong e-integration. To e-integrate a new process is to cultivate a new cognitive ability (or know-how), and it's this ability that allows the agent to responsibly form beliefs with her process. Developing such an ability requires using the process until we are familiar with it. Simply knowing-that our process is reliable will not help us develop the necessary ability.

ence between the two consists in the level of the agent's active involvement. For strong e-integration, an agent ought to have a perspective on the reliability of her process. We understand this to involve reflective access to the reliability of the process and an understanding of why the process is reliable.[6]

In strong e-integration, after acquiring a perspective on the reliability of her belief-forming process, the agent ought to employ her process for a certain period of time to form a variety of beliefs. These beliefs will cohere with her existing beliefs and become inputs for other belief-forming processes (Palermos 2014). This results in the production of yet more beliefs, which, in turn, become inputs in further processes. E-integration is achieved if the dense and reciprocal cooperation of the agent's processes issues a metacognitive sensitivity to the new process's proper functioning. When that happens, the integratedness of the processes makes her counterfactually sensitive to the reliability of the new (extended) process, so that, if the process were to go astray, she would become aware of it. The agent has metacognitive cues that interrupt the transparent and fluent use of the resource (Proust 2014). Importantly, this sensitivity is effective even when the agent is employing her resource transparently.

We have, on this account, a promising solution to our initial worry, that is, there seems to be a way to rely on AI systems seamlessly and transparently and still become aware when there is something wrong with them. As long as our AI-involving processes e-integrate, we are able to responsibly employ them.

Here is strong e-integration illustrated with an example: Recall that Saira is a computer engineer and has access to EarSpeak's algorithm and the data used to train it. Suppose she also follows several blogs that describe in detail how machine learning engineers rectify functioning quirks. She also reads about updates to EarSpeak and what makes it reliable in different conditions. Confident in the technology's promise, she uses it regularly over a period, forming many beliefs with it. These beliefs cohere with her existing beliefs and become input for her existing processes. As time passes, she stops consciously apprehending what EarSpeak says (learns to employ it transparently), and is instead simply aware that, say, a table is located in front of her. If tomorrow EarSpeak were to fail at identifying tables again, she would become aware that something is amiss.

The strong route to e-integration is ideal, but it won't always do the trick in the case of AI integration. Firstly, AI algorithms, especially DNNs, are opaque black boxes even to the machine learning engineers who develop and train them (Petrick 2020). Neither the people operating these AIs nor the ones creating them have a perspective on what makes them reliable. This stands in the way of strong e-integration.

[6]The idea of a perspective on the reliability of one's process is from Greco (2010), but in our employment its requirements are somewhat stronger. Instead of only involving reflective access to the reliability of one's process, we demand that the agent also understand what makes her process reliable.

Moreover, some AI technologies alter their own algorithms as they collect data, and so, even if they remain reliable, the way in which they are reliable changes. This entails that even if an agent has a perspective on the reliability of their process, it may no longer be appropriate after the AI has changed. Thus, such changes in reliability hinder e-integration.

Luckily for us, a perspective on the reliability is not necessary for e-integration. Weak e-integration presents another – less demanding – way that agents can learn to responsibly use a belief-forming process. The agent may lack a perspective on the reliability of her process, and yet, by employing it for a sufficiently long period, she can obtain counterfactual sensitivity to its reliability (Pritchard 2010).[7]

It's worth emphasising that given the lack of an initial perspective on the process's reliability, it is likely to take an agent longer to develop the requisite sensitivity. Nonetheless, once e-integrated, the agent is – just as in cases of strong e-integration – in the position to become aware when her integrated process fails.

It's no surprise that weak e-integration tends to take considerably longer than strong e-integration. Having a perspective on how and why a new belief-forming process is reliable comes with a number of initial beliefs about the functioning of the resource. Absent these beliefs, we have to acquire the relevant information in another way – namely, by using the resource frequently and over a lengthy period of time.

Ideally, agents would pass quickly from mere m-integration to e-integration. This minimises the time agents spend without defences against problematic beliefs. By achieving e-integration quickly, agents ensure that sooner they are in a place where they can become aware if their process were to stop working reliably. In other words, the agent can start responsibly employing her process sooner.

We have seen that strong e-integration is hindered by AI's black-box-ness, and that it's therefore hard to have a perspective on the reliability of AI-involving belief-forming processes. However, this doesn't mean that we cannot have any knowledge about their functioning. Similarly, just because weak e-integration tends to be slow, it doesn't follow that we cannot design AI technology in a way that makes e-integration a little faster. More on these two points in the next section.

5 Defeaters, design, and policy

Taking inspiration from the weak and strong paths to e-integration, we now turn to the question of how to encourage e-integration rather than mere m-integration. We elaborate on a number of design and policy recommendations that enable agents to more quickly

[7] Pritchard (2010) illustrates how our internal biological processes weakly e-integrate into our cognitive system.

develop a sensitivity to the reliability of their processes – and therefore extend responsibly into AI systems.

It goes (almost) without saying that we should strive to develop *reliable* AI technologies. We think this is important to highlight nonetheless. Today, many companies focus on releasing their products fast and early, with reliability often only playing second fiddle.

Here, we must remember an earlier point: what matters isn't the reliability of the technology but the reliability of the belief-forming process. Other things being equal, it should be easier to form a reliable belief-forming process with reliable technology. However, even if the technology is somewhat unreliable, the agent may learn to use only the technology's reliable aspects. Recall how Amna bumped into doors when EarSpeak was first released. She later learned to ignore EarSpeak's silence about doors and to use it only when reliable.

One way to minimise the risk of agents automatically incorporating AI errors is by helping them more easily detect when some resource is unreliable. For instance, EarSpeak could tell the agent not only about the objects it identifies but also the confidence with which these are identified. Suppose EarSpeak doesn't just say the names of objects but also how certain it is about identifying them correctly. By providing information about its own reliability, a technology can help agents constrain their belief-forming process to only use the reliable parts of the technology. While there is much more to say about this, we leave this subject for another paper.

Once the agent has constrained an AI technology to a reliable belief-forming process, to responsibly employ the process, she ought to become (counterfactually) sensitive to its reliability. As discussed previously, this is achieved when the agent is in the position to become aware of her process turning unreliable. One way to cultivate such sensitivity – and to do so quickly – is to have pre-existing beliefs about the domain in which the AI operates.

Consider, for instance, a cardiologist who has been using a surgical AI for many years. The AI suggests cutting an important vein in the heart. Since the cardiologist's cognitive processes are extended to the AI, she forms the belief that she should cut the said vein. However, this new belief is contradicted by her prior belief – instilled by years of education and practical experience – that the said vein must be handled with great care. This makes her suspect a fault in her (AI-involving) belief-forming process, and she identifies the AI as the culprit (rendering it thus opaque).

The case above exemplifies a rebutting defeater (Bergmann 2005; Palermos 2021). A rebutting defeater is a proposition that undermines the truth of an agent's belief. The cardiologist's pre-existing belief that the said vein must be handled with care is a rebutting defeater for the new problematic belief she forms using her AI-extended process. When an agent has prior beliefs that can function as potential rebutting defeaters, she is in a

position to detect when her extended belief-forming process goes awry. In other words, these defeaters can allow the agent to be sensitive to the reliability of her process and, consequently, extend responsibly.

When an expert uses an AI in her domain of expertise, she has a large store of potential defeaters and is therefore likely to be sensitive to the reliability of her extended belief-forming process. Experts are likely to (quickly) e-integrate rather than merely m-integrate. Therefore, one way to ensure that AI errors aren't incorporated into agents' cognitive systems is to mandate that expert AIs be used by experts. We mustn't replace human expertise with AI expertise, but should rather focus on using AIs to complement and improve our cognitive abilities. So, training experts remains as important as it is now.

However, as it has become obvious, AIs aren't always expert systems, and they aren't only used by experts. Think, for instance, of the newest wave of LLMs such as ChatGPT, which provide information on a vast range of topics and are used by the general public. Since the general public isn't knowledgeable in all the topics covered by these AIs, such technologies are difficult to e-integrate (rather than merely m-integrate).

Faiza asks a GPT system about deep-sea creatures' sources of energy. Unless Faiza is a deep-sea expert, she will generally fail to determine whether the AI is producing credible information. If this is so, she is not sensitive to the reliability of the resource and therefore fails to e-integrate with it.

However, just because Faiza isn't an expert in deep-sea creatures, she needn't be completely defenceless against problematic beliefs. She may possess 'ballpark' knowledge about the domain, which can function as potential defeaters. For instance, Faiza might know that the deep-sea is completely dark and thus if the AI informed her that deep-sea creatures gain energy directly from the sun, she would know that this cannot be right. This is akin to how a child who has been taught how to calculate rough estimates is able to detect when her calculator's results are completely off the mark. Note, however, that this won't work if the AI fails in sufficiently subtle ways – say, if it (wrongly) proclaimed that deep-sea creatures gain energy by eating certain rocks. Thus, it's important to cultivate ballpark knowledge across a wide range of domains, this will only rarely get us all the way to e-integration.

Because ballpark knowledge is by definition constrained, another kind of defeater – undercutting defeaters – is especially important in the case of general AI. Undercutting defeaters provide evidence against the reliability of the source of a belief (Bergmann 2005; Palermos 2021). For instance, when calculators flicker or fail to show any answer at all, they indicate that something is amiss. And the humanoid Star Wars protocol droid *C3PO* often tells people that it's not working optimally. When such defeaters are easily recognisable, they can indicate even to non-experts when a resource cannot be trusted.

The lesson we want to draw is that we should design AI technologies to fail in highly salient ways. An EarSpeak that tries to provide the best estimates even if some of its functions fail might *seem* superior to one that simply shuts down on the earliest sign of a problem – but in the present case, it might not be. Since there is a risk of automatically incorporating EarSpeak's errors in transparent use, it's important for the technology to fail so that the agent is yanked out of transparent use. Only then can she apply the full force of her conscious processing to her employment of the resource.

Similar to how an agent may be an expert in the domain for which the AI is used, she may also be an expert in the AI's functioning. Saira, being a computer scientist, understands how AIs are designed and trained and is able to detect a variety of subtle signs that indicate that EarSpeak is failing. She is starting integration with a bigger store of pre-existing beliefs about the process (and potential undercutting defeaters) and is, therefore, able to responsibly extend to EarSpeak quicker than someone who doesn't understand information technology.

Acquiring potential undercutting defeaters can set an agent on the path to strong e-integration discussed in the previous section. Recall that strong e-integration requires an agent to form a perspective on her process's reliability. To build such a perspective, she ought to learn how her process works, what makes it reliable, and – consequently – how it can lose its reliability. This also means that acquiring a perspective gives the agent potential undercutting defeaters – it allows her to recognise indicators of the unreliability of her process.

Much can be done to provision agents with a bigger stock of undercutting defeaters. On the one hand, we can foster computer literacy with the aim of giving most people at least some knowledge of how AIs function (and fail).[8] On the other hand, we can demand that AI designers and corporations disclose how their systems work. While some of these systems are black boxes, there is still a lot that the corporations can disclose about their networks, like the trained models, the algorithm used, the data employed in training, the kind of training, and so forth.

Disclosing information isn't just important because it allows AI experts, such as computer engineers, to understand specific models. AI experts can also play an important role by disseminating their knowledge among the general public, allowing even non-experts to acquire potential rebutting defeaters. To this aim, we think it's important to build structures which encourage such dissemination and training.

Finally, there will be cases when AIs fail in ways that are so subtle that they cannot be detected during transparent employment. Therefore, as a principle of caution, we believe

[8] Here, Heersmink (2018) and Schwengerer (2021)'s discussion on cultivating intellectual virtues to responsibly extend into smart technologies resonates with us.

that AIs that are prone to a sufficiently large number of subtle errors should be designed to be used opaquely.

6. Adversarial attacks

We want to conclude this paper by responding to Wheeler's worry about adversarial attacks.

First, note that adversarial exemplars are carefully crafted to trick an AI system [szegedy2014; Freiesleben (2021)]. This means that AI systems that are vulnerable to these attacks may typically function reliably across a wide range of inputs. Therefore, we do not think that susceptibility to adversarial attacks alone warrants a demand for opaque design – such systems can be reliable enough to be a part of reliable belief-forming processes.

Second, an agent who employs an AI-extended belief-forming process may be able to detect when an adversarial attack turns the process unreliable and, hence, e-integrate it. The agent may, for instance, have ballpark knowledge that can function as a defeater to the odd errors AIs may commit. If my computer vision technology identifies some non-Panda as a Panda while I'm out for a stroll in the city, I will likely doubt the resulting belief.[9] If AI technologies fail this oddly, then spotting errors is easy, and it's also easy for agents to develop sensitivity to the reliability of their process.

However, there is a range of cases that is more problematic. Minute changes to road signs may, in the eye of an AI, turn them from, say, a stop sign into a right-of-way sign (for such cases, see Pavlitska, Lambing, and Zöllner 2023). Here, the main difficulty is that the agent might not detect that there is something wrong with their process. When we're next to an intersection, we're just as likely to encounter a stop sign as a right-of-way sign, and so that makes it unlikely for the agent to have a rebutting defeater.

This is a serious problem, but not an insurmountable one. First, note that such a change may not even constitute a relevant change in reliability (and thus not something the agent needs to be responsive to in order to responsibly use the process). Reliability need not be absolute – and mostly isn't – for a process to be a candidate belief-forming process. If we assume that the agent's belief-forming process has always been susceptible to certain rare errors due to adversarial attacks, then any false beliefs resulting from such attacks are regrettable but do not jeopardize the agent's capacity to maintain a generally reliable extended belief-forming process.

Suppose, however, numerous road signs are altered in the sneaky ways described above so that the computer vision technology becomes unreliable. Here we want to emphasise that we can strive to make an agent's environment safer by, for instance, enforcing laws

[9] The Panda example is inspired by a case in Goodfellow, Shlens, and Szegedy (2015).

that prohibit altering crucial information like road signs. It's not reasonable to demand that the individual become aware of all the minute and subtle ways in which their belief-forming processes may go wrong. The agent may be manifesting sufficient cognitive agency, or be sufficiently sensitive to the reliability of her integrated belief-forming process, even if she fails to become aware of problems in a sufficiently pernicious environment.[10] Laws and customs therefore play an important role by ensuring that our belief-forming processes generally encounter environments where they function reliably.

Note that this is no different from the case of internal processes. Messing with road signs is prohibited because it leads agents to form false beliefs that can endanger them and others. By prohibiting such meddling, we enforce an environment in which we can responsibly employ our processes (by being reasonably sensitive to the reliability of our processes). The absence of a stop sign warrants our belief that we may safely cross the intersection without stopping the car. Such a belief is responsibly formed – even if it is counterfactually possible that someone removed the road sign.

We must – and do – take great care to construct our environment so that it scaffolds our cognition. When our cognitive processes extend to AI technologies, these scaffolds must be suitable for AI-extended belief-forming processes.

7. Conclusion

In this paper, we have argued that it's possible to employ AI technologies without automatically incorporating their strange errors. First, we showed that many unreliable processes are unlikely to be employed transparently, making problematic extensions in such cases impossible. Additionally, even when an agent transparently employs a resource, they do not necessarily lose the ability to (opaquely) reflect on it.

Moreover, agents may responsibly extend their belief-forming processes into AIs. For this, agents need to epistemically integrate (e-integrate) their AI-extended processes, that is, they need to be in the position to become aware of their process becoming unreliable. Such responsible extension means an agent may transparently use an AI without automatically incorporating its strange errors.

We detailed a strong and a weak route to e-integration. The strong route – unlike the weak route – starts with a perspective on the reliability of our AI-involving belief-forming

[10] One way to understand this is by following Pritchard's (2007, 2010) anti-luck intuition on knowledge. The idea is – roughly speaking – that even if our environment isn't safe from knowledge-underminding luck, we may manifest cognitive ability and be sufficiently sensitive to our process's reliability. In such cases, while we responsibly employ our belief-forming process, we fail to achieve knowledge because of extraordinary circumstances.

process. Both routes require the agent to use the resource over some period of time to develop a familiarity with it (though this may happen faster on the strong route). This familiarity allows the agent to become sensitive to her process's reliability.

Our framework allows us to formulate a number of design and policy recommendations geared towards speeding up the process of e-integration. Among them, we mentioned the importance of having AI technologies fail saliently, training domain experts for expert systems, fostering the general public's AI literacy, and making information about the technologies publicly available.

References

Andrada, Gloria. 2020. “Transparency and the Phenomenology of Extended Cognition.” *Límite: Revista de Filosofía y Psicología* 15 (0). https://philarchive.org/rec/ANDTAT-11.

Andrada, Gloria, Robert W. Clowes, and Paul R. Smart. 2022. “Varieties of Transparency: Exploring Agency Within AI Systems.” *AI & SOCIETY*, January. https://doi.org/10.1007/s00146-021-01326-6.

Bergmann, Michael. 2005. “Defeaters and Higher-Level Requirements.” *The Philosophical Quarterly* 55 (220): 419–36. https://doi.org/10.1111/j.0031-8094.2005.00408.x.

Carter, J. Adam, Andy Clark, Jesper Kallestrup, S. Orestis Palermos, and Duncan Pritchard, eds. 2018. *Extended Epistemology*. Vol. 1. Oxford University Press. https://doi.org/10.1093/oso/9780198769811.001.0001.

Carter, J. Adam, and Jesper Kallestrup. 2020. “Varieties of Cognitive Integration.” *Noûs* 54 (4): 867–90. https://doi.org/10.1111/nous.12288.

Clark, Andy. 2003. *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford ; New York: Oxford University Press.

———. 2008. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Philosophy of Mind. Oxford ; New York: Oxford University Press.

———. 2022. “Extending the Predictive Mind.” *Australasian Journal of Philosophy* 0 (0): 1–12. https://doi.org/10.1080/00048402.2022.2122523.

Clark, Andy, and David Chalmers. 1998. “The Extended Mind.” *Analysis* 58 (1): 7–19. http://www.jstor.org/stable/3328150.

Clowes, Robert W. 2019. “Immaterial Engagement: Human Agency and the Cognitive Ecology of the Internet.” *Phenomenology and the Cognitive Sciences* 18 (1): 259–79. https://doi.org/10.1007/s11097-018-9560-4.

Dreyfus, Hubert L., and Stuart E. Dreyfus. 1988. *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. 1. paperback ed. New York: The Free Pr.

Facchin, Marco. 2022. “Phenomenal Transparency, Cognitive Extension, and Predictive Processing.” *Phenomenology and the Cognitive Sciences*, July, 1–23. https://doi.org/10.1007/s11097-022-09831-9.

Freiesleben, Timo. 2021. “The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples.” *Minds and Machines*, 77–109. https://doi.org/10.1007/s11023-021-09580-9.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2015. “Explaining and Harnessing Adversarial Examples.” March 20, 2015. https://doi.org/10.48550/arXiv.1412.6572.

Greco, John. 1999. “Agent Reliabilism.” *Nous* 33 (s13): 273–96. https://doi.org/10.1111/0029-4624.33.s13.13.

———. 2010. *Achieving Knowledge: A Virtue-Theoretic Account of Epistemic Normativity*. Cambridge ; New York: Cambridge University Press.

Grush, Rick, and Alison Springle. 2019. “Agency, Perception, Space and Subjectivity.” *Phenomenology and the Cognitive Sciences* 18 (5): 799–818. https://doi.org/10.1007/s11097-018-9582-y.

Heersmink, Richard. 2018. “A Virtue Epistemology of the Internet: Search Engines, Intellectual Virtues and Education.” *Social Epistemology* 32 (1): 1–12. https://doi.org/10.1080/02691728.2017.1383530.

Heidegger, Martin. 1976. *Sein und Zeit*. 13. unveränd. Aufl. Tübingen: Niemeyer.

Hernández-Orallo, José, and Karina Vold. 2019. “AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI.” In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 507–13. Honolulu HI USA: ACM. https://doi.org/10.1145/3306618.3314238.

Menary, Richard, ed. 2010. *The Extended Mind*. Life and Mind. Cambridge, Mass: MIT Press.

Merleau-Ponty, Maurice, and Donald A. Landes. 2012. *Phenomenology of Perception*. Abingdon, Oxon ; New York: Routledge.

Palermos, Spyridon Orestis. 2011. “Belief-Forming Processes, Extended.” *Review of Philosophy and Psychology* 2 (4): 741–65. https://doi.org/10.1007/s13164-011-0075-y.

———. 2014. “Knowledge and Cognitive Integration.” *Synthese* 191 (8): 1931–51. https://doi.org/10.1007/s11229-013-0383-0.

———. 2021. “System Reliabilism and Basic Beliefs: Defeasible, Undefeated and Likely to Be True.” *Synthese* 199 (3-4): 6733–59. https://doi.org/10.1007/s11229-021-03090-y.

Pavlitska, Svetlana, Nico Lambing, and J. Marius Zöllner. 2023. “Adversarial Attacks on Traffic Sign Recognition: A Survey.” July 17, 2023. https://doi.org/10.48550/arXiv.2307.08278.

Petrick, Elizabeth R. 2020. “Building the Black Box: Cyberneticians and Complex Systems.” *Science, Technology, & Human Values* 45 (4): 575–95. https://doi.org/10.1177/0162243919881212.

Pritchard, Duncan. 2007. “Anti-Luck Epistemology,” 22.

———. 2010. “Cognitive Ability and the Extended Cognition Thesis.” *Synthese* 175 (S1): 133–51. https://doi.org/10.1007/s11229-010-9738-y.

Proust, Joëlle. 2014. “Epistemic Action, Extended Knowledge, and Metacognition.” *Philosophical Issues* 24 (1): 364–92. https://doi.org/10.1111/phis.12038.

Schwengerer, Lukas. 2021. “Online Intellectual Virtues and the Extended Mind.” *Social*

Epistemology 35 (3): 312–22. https://doi.org/10.1080/02691728.2020.1815095.

Smart, Paul R., Gloria Andrada, and Robert W. Clowes. 2022. “Phenomenal Transparency and the Extended Mind.” *Synthese* 200 (4): 335. https://doi.org/10.1007/s11229-022-03824-6.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. “Intriguing Properties of Neural Networks.” February 19, 2014. https://doi.org/10.48550/arXiv.1312.6199.

Thompson, Evan, and Mog Stapleton. 2009. “Making Sense of Sense-Making: Reflections on Enactive and Extended Mind Theories.” *Topoi* 28 (1): 23–30. https://doi.org/10.1007/s11245-008-9043-2.

Wheeler, Michael. 2019. “The Reappearing Tool: Transparency, Smart Technology, and the Extended Mind.” *AI and Society* 34 (4): 857–66. https://doi.org/10.1007/s00146-018-0824-x.

———. 2021. “Between Transparency and Intrusion in Smart Machines.” *Perspectives Interdisciplinaires Sur Le Travail Et La Santé (PISTES)*.