Philosophy & Technology (2024) 37:91
https://doi.org/10.1007/s13347-024-00774-4

RESEARCH ARTICLE

Should We Discourage AI Extension? Epistemic Responsibility and AI

Hadeel Naeem[1] · **Julian Hauser**[2]

Received: 22 November 2023 / Accepted: 21 June 2024 / Published online: 15 July 2024

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, even an agent who transparently employs a resource may reflect on its reliability. Finally, agents can rely on a process transparently and be yanked out of their transparent use when it turns problematic. 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.

Keywords Phenomenal transparency · Artificial intelligence · Cognitive extension · Adversarial attack · Cognitive integration

✉ Hadeel Naeem
hadeel.naeem@khk.rwth-aachen.de

Julian Hauser
julian@julianhauser.com

[1] Käte Hamburger Kolleg, Cultures of Research RWTH University Aachen, Theaterstrasse 75, 52062 Aachen, Germany

[2] LOGOS - Universitat de Barcelona, Facultat de FilosofiaC. Montalegre, 6-8, desp. 4009, Barcelona 08001, Spain

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 & 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 & 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 only using AI technologies while they are present in experience, Wheeler aims to prevent cognitive extension and, consequently, the incorporation of the AI's faults.[2]

This paper argues that we can learn to (epistemically) 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 AI to never be transparent is therefore unnecessary. Using the virtue reliabilist concept of cognitive integration (Greco, 2010; Pritchard, 2010), we show how agents can rely on external resources even when they employ them transparently. Cognitive integration (which we prefer

[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 addresses this question head-on.

to call epistemic integration), we argue, allows us to understand how to 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 exemplars, lessening the force of his argument.

Our paper is structured in the following way. After this first section, Sect. 2 outlines the concept of transparency and shows that many unreliable technologies are not employed transparently. In Sect. 3, we argue that transparency doesn't entail that a resource cannot be present in experience. 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 Sect. 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[3]. 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.[4]

Transparency involves the skilled use of a resource (Heidegger, 1976; Merleau-Ponty & Landes 2012; Dreyfus and Dreyfus 1988; Clowes, 2019; Andrada, 2020; (Farina and Lavazza 2022). When Saira had just received her EarSpeak, the sudden

[3] We thank a reviewer for pointing out that a similar technology that already exists in Microsoft's *Seeing AI* (Novet, 2017).

[4] As noted by a reviewer, it is currently unknown to what extend speech perception depends on attention. If an agent needs to pay attention to the source of an utterance to parse it, using a device like EarSpeak transparently will not be possible. This doesn't affect the core of our argument, as other AI-enabled devices could repose on modalities where the possibility of transparent employment is well established. We might, just to give an example, think of a device that — in a fashion akin to tactile-visual sensory substitution devices (TVSS) (Clark, 2003) — communicates with the user through touch. In the following we will proceed on the assumption that the transparent use of EarSpeak is possible.

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 & 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 the focus of attention and, therefore, the agent cannot easily detect when things go wrong. In contrast, when we use AI technologies while they are present as objects of experience, 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 while attending to them or not at all. If, like Wheeler (2019) and certain others (Clark, 2003, 2008; Thompson & Stapleton 2009), you believe that transparency is necessary for cognitive extension, then you will agree that 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.[5]

[5] If you agree with those authors (Andrada, 2020; Facchin, 2022; Smart et al. 2022) who have recently argued against the claim that transparency is necessary for cognitive extension, then the argument here won't be quite as decisive. However, note that in this paper, we argue that we can epistemically responsibly employ resources even when these are transparent to the user. To the extent that this claim is true, it's *a fortiori* possible to responsibly employ a resource in the sort of seamless fashion that is important even to authors critical of the role of transparency.

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 responsibly and transparently employ a reliable process that involves employing a (partially) unreliable technology.[6] 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 Present

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 it impossible for the object to be present in experience, and agents can inspect resources they transparently employ. Additionally, agents that transparently employ resources can become aware when their processes malfunction. Thus, we can be epistemically responsible when relying on AI technologies transparently.

Agents may not always be able to constrain their transparent use of 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 compo-

[6] It is important to note here that the agent can learn to responsibly employ a partially faulty technology but their belief-forming process of employing the said technology ought to be reliable. We do not mean to say that an epistemic agent can learn to responsibly employ an unreliable belief-forming process. Instead, we want to simply highlight that there is a difference between the technology or resource and our process of employing the resource.

nent 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 suppose 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 present 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, are rendered present to you. They make themselves known as something that may fail (in particular, that may fail when it's too dark).[7]

Making sure that a transparently employed resource is used so that it is rendered present to experience when it goes wrong is reminiscent of what some virtue epistemologists (Greco, 2010; Pritchard, 2010; Palermos, 2014) call 'cognitive integration'.

[7] 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.

4 Epistemic Integration

The virtue reliabilist concept of cognitive integration posits that when we responsibly rely on a process, we are in the position to become aware when it stops functioning reliably. On this view, we can counter Wheeler's worry by demanding that agents cognitively integrate AI technologies. Then, instead of internalising AI's errors, agents become aware when their extended processes malfunction.

First, a bit of background on the concept of cognitive integration: Greco describes cognitive integration as a "function of cooperation and interaction" (2010, 152) with which new belief-forming 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 employ them to responsibly form beliefs (and thus potentially acquire knowledge). The main idea is that a new belief-forming process integrates into an agent's cognitive system when the agent has formed a link to the reliability of the said belief-forming process and is in the position to become aware when and if her process malfunctions.

Pritchard (2010) and Palermos (2011, 2014) employ the concept of cognitive integration to show how virtue reliabilism fits neatly with the thesis of cognitive extension. While Pritchard highlights how cognitive integration is similar to cognitive extension, Palermos (2014) explicitly argues that the conditions required for cognitive integration are the same as those required for cognitive extension.

Meanwhile, in the philosophy of mind, a distinct concept of cognitive integration arose as a challenge to what's now called the first wave of thinking on extended cognition. Authors such as Menary (2007) argued that the study of the extended mind needn't be based on instances where body-external resources have functional parity with body-internal structures but should instead focus on investigating how two objects (body and external resource) integrate to give rise to a unified system. He mentions three complementary ways of understanding integration: reciprocal causation as studied in dynamical systems theory, incorporation of the external resource into the agent's body schema (that is, the neural representation of our body's posture, shape, and movement (Gallagher 2005), and the manipulation of external objects to achieve an agent's cognitive tasks. Note how these three proposals are not, as such, constrained to studying how agents responsibly employ extended processes to form beliefs.

The difference between the epistemologists' notion of cognitive integration and that of philosophers of mind was noted by Carter and Kallestrup (2020),[8] who distinguish between epistemic integration (hereafter, e-integration) and metaphysical integration (m-integration). We adopt this distinction as we agree that m-integration need not fulfil the same conditions as e-integration. We think this is evident, for instance,

[8] Carter and Kallestrup (2020) also argue that there are no necessary and sufficient conditions for these two integrations and propose a kind of cluster approach to understand these integrations. Note that we don't subscribe to the cluster approach – and the paper doesn't require that we endorse or reject the idea. We only take their distinction of the two integrations and provide our own reasons for why it's reasonable to distinguish between the two integrations.

in the fact that m-integration can concern all sorts of cognitive processes, whereas e-integration is about belief-forming processes.[9]

More importantly, perhaps, the distinction is at the root of the problem we discuss, at least as it has been formulated by Wheeler (2019). His worry is that we may integrate extended belief-forming processes in such a way that we are not in a position to ascertain whether the resulting beliefs are problematic. In that case, we have m-integrated but not e-integrated the process: we form extended beliefs, but these are not *responsibly* formed.

As mentioned, e-integration depends on new beliefs cohering, or at least not being inconsistent, with existing beliefs, and it is therefore *belief-forming processes* that 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 e-integration is concerned with belief-forming processes.

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 (Goldman, 1979; Sosa, 1992; Alston, 1995; Goldberg, 2010).

Here's one way an agent could e-integrate a belief-forming process: 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.

This is an example of the *reflective route to e-integration*. Saira has reflective access to the reliability of her belief-forming process so that she gains — to use a term employed by Greco (2010) —a perspective on its reliability. She knows when and why her process is reliable, and she can use this knowledge to identify when it malfunctions.

The reflective route to e-integration is ideal, but it won't always do the trick in the case of AI integration. Inasmuch as building a perspective on the reliability of a process requires understanding what makes a process reliable, it is possible that we may not be able to reflectively integrate AI algorithms. This is mainly because AI algorithms, especially DNNs, are opaque black boxes even to the machine learning engineers who develop and train them (Petrick, 2020).[10] It may be that the people

[9] A reviewer has pointed out to us that knowledge-how may arguably also e-integrate. This is an intriguing and plausible suggestion deserving of further investigation. The present paper – and the existing literature – concerns only belief-forming processes and we therefore only discuss e-integration in the context of this kind of process.

[10] For a discussion on how blackboxness may affect how we rely on AI systems, see Dahl (2018) and Vaassen (2022).

operating these AI and the ones creating them do not know if — and in what contexts —they are reliable and, even if they do know that they are reliable, they may not know why this is the case. This may stand in the way of the reflective route to e-integration.

Luckily, a perspective on, or reflective access to, the reliability of one's belief-forming process is not necessary for e-integration (Greco, 2010, Pritchard, 2018a, 2018b; Palermos, 2014). Even in the absence of such a perspective, an agent may obtain counterfactual sensitivity to her process's reliability. This non-reflective route to e-integration requires the agent to employ her process to form a variety of beliefs. These beliefs will cohere with her existing beliefs and become inputs for other belief-forming processes. 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 be alerted (Palermos 2014). 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.[11]

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 can responsibly employ them.

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 reflective route to e-integration – in the position to become aware when her integrated process fails.

It's no surprise that the non-reflective route to e-integration tends to take considerably longer than the reflective route to e-integration. Having a perspective on whether (and, ideally, 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 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

[11] As a reviewer has pointed out to us, Pritchard (2010) also allows for another, more passive, non-reflective kind of e-integration. On this view, a process may e-integrate simply because it has been employed for a sufficiently long time during which it has given rise to a string of true beliefs. The agent doesn't need to develop any sensitivity to the reliability of the process (although Pritchard also requires that the agent would become aware did any of the string of beliefs turn out to be problematic). Note that such e-integration is premised entirely on the relevant process never turning unreliable. The agent doesn't acquire any sensitivity to the reliability of her process, which means that such e-integration cannot allow the agent to spot when her (AI-extended) process falters. Because of this, we do not discuss this view in the present paper and focus instead on the more popular, active, kind of non-reflective e-integration (which Pritchard also endorses in his later writings).

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 the reflective route to e-integration may be 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. We may receive expert testimony that these AI systems are reliable or we may observe a track record of beliefs that confirm their reliability.[12] Similarly, just because the non-reflective route to 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 reflective and non-reflective 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 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.[13] 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) sen-

[12] We would like to thank an anonymous reviewer for this point.

[13] In footnote 11, we noted that Pritchard (2010) allows for a passive non-reflective form of e-integration that merely requires unproblematic use over some extended period of time. If one agrees with this suggestion, then merely making sure that AI technologies be reliable (and used over some period of time) can be enough for e-integration.

sitive to its reliability. As discussed previously, this is achieved when the agent is in the position to become aware of her process becoming problematic. 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 reflective route to e-integration discussed in the previous section. Recall that the reflective route to 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).[14] 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

[14] Here, Heersmink (2018) and Schwengerer (2021)'s discussion on cultivating intellectual virtues to responsibly extend into smart technologies resonates with us. Also, see (Krügel, Ostermaier, & Uhl 2022) for the importance of digital literacy to learn to use AI responsibly.

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 that AIs that are prone to a sufficiently large number of subtle errors should be designed to render themselves present in experience while being used.

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 (Szegedy 2014; 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 design that resists transparent use – 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 problematic 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.[15] 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,

[15] The Panda example is inspired by a case in Goodfellow et al. (2015).

enforcing laws 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.[16] 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 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 giving rise to problematic beliefs. Such responsible extension means an agent may transparently use an AI without automatically incorporating its strange errors.

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

[16] 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.

experts for expert systems, fostering the general public's AI literacy, and making information about the technologies publicly available.

Acknowledgements Nothing to report yet.

Author Contributions Both authors (HN and JH) contributed to the conception, design, and writing of the manuscript, and they both read and approved the final draft.

Funding Hadeel Naeem's research is a part of her fellowship at the Käte Hamburger Kolleg: Cultures of Research, funded by the Germany Ministry of Education and Research (project number: BMBF 01UK2104). Julian Hauser's work is funded by the Swiss National Research Foundation (project number: P500PH_202829 / 1).
Open Access funding enabled and organized by Projekt DEAL.

Data Availability Not applicable.

Declarations

Ethics Approval and Consent to Participate not applicable.

Consent for Publication not applicable.

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

References

Alston, W. P. (1995). How to think about reliability. *Philosophical Topics*, *23*(1), 1–29.
Andrada, G. (2020). Transparency and the phenomenology of extended cognition. *Límite: Revista De Filosofía Y Psicología* 15 (0). https://philarchive.org/rec/ANDTAT-11.
Andrada, G., Clowes, R. W., & Smart, P. R. (2022). Varieties of transparency: Exploring agency within AI systems. *AI & SOCIETY*. https://doi.org/10.1007/s00146-021-01326-6. January.
Bergmann, M. (2005). Defeaters and higher-level requirements. *The Philosophical Quarterly*, *55*(220), 419–436. https://doi.org/10.1111/j.0031-8094.2005.00408.x.
Carter, J. A., & Kallestrup, J. (2020). Varieties of cognitive integration. *Noûs*, *54*(4), 867–890. https://doi.org/10.1111/nous.12288.
Carter, J. A., Clark, A., Kallestrup, J., Orestis Palermos, S., & Pritchard, D. (Eds.). (2018). *Extended epistemology* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oso/9780198769811.001.0001.
Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford: Oxford University Press.
Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. Philosophy of Mind. New York: Oxford University Press

Clark, A. (2022). Extending the predictive mind. *Australasian Journal of Philosophy*, *0*(0), 1–12. https://doi.org/10.1080/00048402.2022.2122523.
Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19. http://www.jstor.org/stable/3328150.
Clowes, R. W. (2019). Immaterial engagement: Human agency and the cognitive ecology of the internet. *Phenomenology and the Cognitive Sciences*, *18*(1), 259–279. https://doi.org/10.1007/s11097-018-9560-4.
Dahl, E. S. (2018). Appraising black-boxed technology: The positive prospects. *Philosophy & Technology*, *31*(4), 571–591. https://doi.org/10.1007/s13347-017-0275-1.
Dreyfus, H. L., 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, M. (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.
Farina, M., and Lavazza, A. (2022). Incorporation, transparency and cognitive extension: Why the distinction between embedded and extended might be more important to ethics than to metaphysics. *Philosophy & Technology*, *35*(1), 10. https://doi.org/10.1007/s13347-022-00508-4.
Freiesleben, T. (2021). The intriguing relation between counterfactual explanations and adversarial examples. *Minds and Machines*, 77–109. https://doi.org/10.1007/s11023-021-09580-9.
Gallagher, S. (2005). How the Body Shapes the Mind (Oxford, 2005; online edn, Oxford Academic, 1 Feb. 2006), https://doi.org/10.1093/0199271941.001.0001.
Goldberg, S. C. (2010). *Relying on others: An essay in Epistemology*. Oxford University Press.
Goldman, A. I. (1979). What is justified belief? In G. S. Pappas (Ed.), *Justification and Knowledge: New Studies in Epistemology* (pp. 1–23). Philosophical Studies Series in Philosophy. Springer Netherlands. https://doi.org/10.1007/978-94-009-9493-5_1.
Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. March 20, 2015. https://doi.org/10.48550/arXiv.1412.6572.
Greco, J. (2010). *Achieving knowledge: A virtue-theoretic account of epistemic normativity*. Cambridge University Press.
Grush, R., & Springle, A. (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, R. (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, M. (1976). *Sein und Zeit*. 13. unveränd. Aufl. Tübingen: Niemeyer.
Hernández-Orallo, J. & Vold, K. (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*(pp. 507–513). Honolulu HI USA: ACM. https://doi.org/10.1145/3306618.3314238.
Krügel, S., Ostermaier, A., & Uhl, M. (2022). Zombies in the loop? Humans trust untrustworthy AI-Advisors for ethical decisions. *Philosophy & Technology*, *35*(1), 17. https://doi.org/10.1007/s13347-022-00511-9.
Menary, R. (2007). *Cognitive integration: Mind and cognition unbounded*. Palgrave-Macmillan.
Menary, ed. (2010). *The extended mind*. Life and Mind. Cambridge, Mass: MIT Press.
Merleau-Ponty, M., & Landes, D. A. (2012). *Phenomenology of perception*. London: Routledge.
Novet, J. (2017). Microsoft Has a new app that tells the visually impaired What's in front of them. CNBC, July 18, https://www.cnbc.com/2017/07/12/microsoft-launches-seeing-ai-app-for-ios.html.
Palermos, S. O. (2011). Belief-forming processes, extended. *Review of Philosophy and Psychology*, *2*(4), 741–765. https://doi.org/10.1007/s13164-011-0075-y.
Palermos, S. O. (2014). Knowledge and cognitive integration. *Synthese*, *191*(8), 1931–1951. https://doi.org/10.1007/s11229-013-0383-0.
Palermos, S. O. (2021). System reliabilism and basic beliefs: Defeasible, undefeated and likely to be true. *Synthese*, *199*(3–4), 6733–6759. https://doi.org/10.1007/s11229-021-03090-y.
Pavlitska, S., Lambing, N., Marius, J., & Zöllner (2023). Adversarial attacks on traffic sign recognition: A survey. July 17, 2023. https://doi.org/10.48550/arXiv.2307.08278.
Petrick, E. R. (2020). Building the black box: Cyberneticians and complex systems. *Science Technology & Human Values*, *45*(4), 575–595. https://doi.org/10.1177/0162243919881212.
Pritchard, D. (2007). Anti-Luck Epistemology, Synthese, Vol. 158, No. 3 (Oct., 2007), pp. 277-297. https://www.jstor.org/stable/27653595
Pritchard, D. (2010). Cognitive ability and the extended cognition thesis. *Synthese*, *175*(S1)), 133–151. https://doi.org/10.1007/s11229-010-9738-y.

Pritchard, D. (2018b). Extended virtue epistemology. *Inquiry: A Journal of Medical Care Organization, Provision and Financing*, *61*, 632–647. https://doi.org/10.1080/0020174x.2017.1355842.
Pritchard, D. (2018a). Extended epistemology. In J. Adam, A. Carter, J. Clark, S. Kallestrup, Orestis Palermos, & D. Pritchard (Eds.), *Extended epistemology* (pp. 90–104). Oxford University Press.
Proust, J. (2014). Epistemic action, extended knowledge, and metacognition. *Philosophical Issues*, *24*(1), 364–392. https://doi.org/10.1111/phis.12038.
Schwengerer, L. (2021). Online intellectual virtues and the extended mind. *Social Epistemology*, *35*(3), 312–322. https://doi.org/10.1080/02691728.2020.1815095.
Smart, P. R., Andrada, G., & Clowes, R. W. (2022). Phenomenal transparency and the extended mind. *Synthese*, *200*(4), 335. https://doi.org/10.1007/s11229-022-03824-6.
Sosa, E. (1992). Generic reliabilism and virtue epistemology. *Philosophical Issues*, *2*, 79–92. https://doi.org/10.2307/1522856.
Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. & Fergus, R. (2014). Intriguing properties of neural networks. February 19, 2014. https://doi.org/10.48550/arXiv.1312.6199.
Thompson, E., & Stapleton, M. (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.
Vaassen, B. (2022). AI, opacity, and personal autonomy. *Philosophy & Technology*, *35*(4), 88. https://doi.org/10.1007/s13347-022-00577-5.
Wheeler, M. (2019). The reappearing tool: Transparency, smart technology, and the extended mind. *AI and Society*, *34*(4), 857–866. https://doi.org/10.1007/s00146-018-0824-x.
Wheeler, M. (2021). Between transparency and intrusion in smart machines. *Perspectives Interdisciplinaires Sur Le Travail Et La Santé (PISTES)*.