Final Draft - Please cite published version, forthcoming in M. Ienca & F. Jotterand (eds.), Ethics of Artificial Intelligence in Brain and Mental Health. Springer.

AI Extenders and the Ethics of Mental Health

Karina Vold and José Hernández-Orallo

Abstract: The extended mind thesis maintains that the functional contributions of tools and artefacts can become so essential for our cognition that they can be constitutive parts of our minds. In other words, our tools can be on a par with our brains: our minds and cognitive processes can literally 'extend' into the tools. Several extended mind theorists have argued that this 'extended' view of the mind offers unique insights into how we understand, assess, and treat certain cognitive conditions. In this chapter we suggest that using AI extenders, i.e., tightly coupled cognitive extenders that are imbued with machine learning and other 'artificially intelligent' tools, presents both new ethical challenges and opportunities for mental health. We focus on several mental health conditions that can develop differently by the use of AI extenders for people with cognitive disorders and then discuss some of the related opportunities and challenges.

Keywords: cognitive extension; extended mind; enhancement; AI ethics; mental health; cognitive disorder; cognitive capability; Alzheimer's disease; memory; function

1 Introduction

Consider two scenarios. Helen is an 80-year-old woman who was diagnosed with Alzheimer's a few years ago. She used to use physical labels and other cues at home to help support her memory. But she now wears augmented

reality glasses that label objects in her vision range, detect hazards when she is manipulating objects, indicate where things are, keep her agenda, recognize the faces of family and friends, and help her while navigating beyond her home and in many other everyday scenarios. Helen's grandson, Lewis, is a 10-year-old boy with ADHD (attention deficit hyperactivity disorder). He uses a special device that monitors his activity (movements, speech, gaze, etc.), issues recommendations through a gamified scoring system, sends indicators to teachers and family, and performs other well-thought smooth interventions. The device has boosted Lewis's self-confidence and focus, and academic results are improving. Lewis is also allowed to use his device for exams, becoming the envy of his classmates.

While scenarios such as those above are not yet possible with current clinical technologies, with current trends moving towards digital health applications they may become commonplace in the future. The distinctive features of both these cases are (1) the person is *tightly coupled* with a tool (a device, such as a smartphone, or a wearable, such as an augmented reality headset) and (2) the tool integrates artificial intelligence (AI) capabilities (image recognition, face detection, navigation, event recognition, natural language processing, speech recognition, prediction, etc.). Together these features put tension on an "internalist" view of the mind, according to which any process outside the brain is considered as subsidiary to human cognition. In contrast, these two characteristics are in perfect alignment with the Extended Mind Thesis (EMT), the view that the human mind and cognition is sometimes constituted by more than just the brain.[1]

In Hernández-Orallo & Vold, we introduced the notion of "AI extenders", as any external tool that uses AI capabilities and is sufficiently tightly coupled with a person's cognitive system that it should be considered a cognitive extender more broadly.[2] While in contemporary philosophy there has been quite a lot of discussion of the theory and potential of EMT, these discussions focus on relatively simple technologies. There has been almost no consideration of what more sophisticated emerging AI and data-enabled technologies can do qua cognitive extenders. We argue that with the use of artificial intelligence, there is a strong case that some of our cognition is taking place outside our brains and having deeper effects for replacing,

enhancing or regulating parts of our cognitive activity. In the examples above, we have two cases of AI extenders being used to help people with very different mental conditions. How are "AI extenders" going to serve and affect mental health, as well as our philosophical and ethical interpretation of treatments and interventions? Answering this question is the goal of this chapter.

The range of mental conditions is structured in well-known classifications, such as the ICD-10 Classification of Mental and Behavioral Disorders published by the World Health Organization (WHO).[3] We will analyze some of these conditions under the perspective of the EMT and will investigate how current and future AI extenders, according to the capabilities they provide, may lead us to new therapeutic possibilities, new ethical challenges and a philosophical re-understanding of some aspects of mental health.

The rest of the chapter is organized as follows. Section 2 reviews the EMT, and its evolution towards more flexible interpretations. Section 3 gives a definition of AI extenders, as a particular case of cognitive extenders, making it distinctive from other uses of technology for mental health. Section 4 selects a few mental conditions and interprets them under the EMT. Section 5 is more explicit about the capabilities that AI tools can extend, how these tools can be applied to a diversity of mental conditions, and, how this can change in the future, especially if the EMT is accepted by clinicians and patients. Section 6 explores some of the positive and negative impacts of AI extenders, when used either with or without a therapeutic motivation. Finally, section 7 gives a series of recommendations to AI designers and clinicians, and open questions for future research.

2 What is the Extended Mind Thesis?

For a long time now most scientific investigation into the mind, e.g. in neuroscience and cognitive science, has considered the brain to be the sole physical locus of the mind. According to these brain sciences, the mind is an information processing system that sits in between sensory inputs and

motor outputs, and which functions by performing computations on inner representations of the world.[4] This 'internalist' view is 'neuro-centric' in the sense that all of the relevant inner representations and computations are thought to be instantiated in neural networks in the brain, while everything beyond the brain is considered an input source, or an arena for outputs.[5] A result of this demarcation of the mind is that mental disorders have also tended to be demarcated based on this assumed boundary of the mind, that is, mental disorders are thought to be brain disorders.

Over the past two decades, however, a new picture of the mind has gained popularity, which, if true, would challenge this orthodox view. The EMT maintains that human thought and reason are not entirely 'in the head'. Instead, the effective circuits of human thought and reason sometimes crucially involve the technologies we use and even our social networks and institutional structures, such that the physical locus of the mind is 'extended' beyond the brain.[1] The technologies that are often cited as examples of 'extenders', range from humble writing utensils, such as pens and pencils, and the external symbols they create[1], to more sophisticated technologies, such as smartphones, as well as many things in between, including Scrabble tiles and Venn diagrams. We can state the EMT as follows:

> Extended Mind Thesis (EMT) = Representational vehicles (or information-bearing structures) located beyond the brain can be partly constitutive of an agent's mental states and processes.

The EMT accepts the claim that the mind is an information processing system—a core commitment of cognitive science—but maintains that the relevant information-bearing structures, that is, the vehicles of mental representations, can sometimes be instantiated by non-biological elements, beyond the brain. To put it simply, if the EMT is correct, then there is more to the mind than the brain. And, accordingly, at least in some cases, in order to explain and treat mental disorders, we may need to look beyond the brain. Indeed, a number of defenders of the EMT have argued that certain disorders, such as Alzheimer's, borderline personality disorder, and autism, can be better understood, assessed, and treated by taking a wider lens on

physical locus of the mind—we will discuss these examples in section 4. But first we will discuss some of the arguments that have been used to support the EMT.

In their seminal paper titled, 'The Extended Mind', Clark and Chalmers motivate the EMT by considerations of parity between external representational vehicles and internal cognitive parts.[1] They describe a scenario intended to motivate their view which involves two people—Otto and Inga. Inga has a well-functioning biological memory that allows her to recall the location of a museum she wishes to visit and to successfully navigate her way there. Despite having Alzheimer's, Otto also performs this task quite well, but he does so by relying on a notebook, which he uses as an external memory tool—recording important information and consulting his notes whenever needed. Clark and Chalmers argue that "in all the relevant respects", Otto's notebook plays the same functional role in guiding his behavior as Inga's biological memory does for her, and so the information stored in Otto's notebook should count as a part of the constitutive machinery of Otto's mind just as the information stored in Inga's brain does for her. Hence, their argument is based on the idea that external resources can make equivalent functional contributions to one's cognitive processes as internal resources can. They write:

> "If, as we confront some task, a part of the world functions as a process which, *were it done in the head*, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world *is* (so we claim) part of the cognitive process." [1] (p. 8; emphasis in the original)

This idea has come to be known as the 'parity principle'. Arguments based on considerations of functional parity, like Clark and Chalmers', represent the first-wave of arguments for the EMT. They have come under criticism on several fronts, which eventually led to a second-wave of arguments for the EMT. While we will not rehearse all of these criticisms here, there are a few worth mentioning because they are particularly relevant to the discussion at hand.

One long-standing debate around the appeal to functional parity has been over how to characterize the relevant functional role that Clark and Chalmers appeal to (this amounts to a commitment to the view that philosophers call 'role functionalism'). This functional role determines in what ways external components must be similar to internal ones, in order to count as a genuine part of the mind. Clark and Chalmers, for example, mention three features they think are important for capturing the ways in which the information in Otto's notebook is on par with the information in Inga's brain: (1) both are a constant in the agent's life, (2) both are readily accessible when the agent needs them, and (3) both are relied upon, such that the agent trusts and endorses the information without hesitation. Among other objections, these features have come under criticism as being too coarse, for at least two reasons. Some argue that certain important features are missing and that if the particular nuances of human biological cognitive functions were included, then it would be unlikely that Otto's notebook, or indeed any external resources could really count as extensions.[6] While others have argued that these three conditions alone are too coarse as they would allow for all sorts of external objects to count as extenders, leading to absurd scenarios where everything that one reads on the internet, for example, or the entire library that one visits, are a part of their mind.[7]

The parity argument has further been criticized for relying on an overly normative picture of the mind: by letting some notion of a 'healthy' biological mind set the baseline for what could count as a possible extension of the mind. For example, imagine a 'healthy' biologically mind, like Inga's, started to rely on external objects to enhance her memory. Extended mind theorists are generally quite supportive of the idea that cognitive extenders enhance healthy minds—they allow us to go beyond what the naked brain can do. But in order for the parity principle to support these cases, one would have to imagine a case that involved someone (e.g. a Martian with superhuman cognitive capacities) who (purely internally) had mental capacities that go beyond those of a 'normal' human. It is a roundabout way of arguing for what should be a straightforward commitment of the EMT.

Ultimately these limitations, along with other philosophical challenges facing the parity principle, were in part what motivated 'second-wave'

extended mind theorists to instead appeal to a criterion of functional *contributions* rather than functional *parity*. Second-wave arguments focus on the different but complementary contributions that external and internal resources make to bring about cognitive functions.[8,9] This style of argument can straightforwardly defend the possibility of enhancing 'normal' or 'healthy' minds, and hence is able to overcome the first-wave focus on compensation for biological deficits. For this reason, it is likely that AI extenders, as we will define them in the next section, will need to appeal to complementarity arguments. Part of what makes AI applications so useful for humans is how they can go beyond our own cognitive capacities: processing more information, at faster speeds, and in new ways.

Complementarity arguments tend to focus on *how* external resources can be appropriately integrated with internal resources such that they can jointly govern cognitive activities and behavior, even though their functional contributions are not strictly analogous (as parity had demanded). One of the central issues for these views is articulating exactly how inner and outer resources need to be integrated. Some version and selection of the features that Clark and Chalmers defend—constancy, accessibility, and reliability—are often endorsed, though rarely the exact set.[9-11] Heersmink, for example, has recently defended a multidimensional framework, including the dimensions of information flow, reliability, durability, trust, procedural transparency, informational transparency, individualization, and transformation.[11] We will not engage this debate here, but for the purposes of this chapter we broadly endorse Heersmink's definition of 'coupling' as well as his view that each of these dimensions are a matter of degree. Perhaps it is most important to note that all of these dimensions are relational. That is, they depend on how a particular agent stands in relation to the tool—Does the agent rely on the tool in order to complete a cognitive task? Does the agent trust the information provided by the tool (rarely questioning its veracity?) How individualized is the tool for that particular agent (i.e. how difficult would it be for another agent to use it)? This would have to be assessed on a case by case basis.

The EMT also includes cases beyond non-biological artifacts; sometimes people rely on the minds of others as cognitive extenders. Such cases are

known as 'socially extension'. Clark and Chalmers, for example, discuss the possibility of Otto relying on Inga's mind, rather than on his notebook.[1] So long as Inga is constantly in his life, the information in her mind (about where the museum is located) is accessible to him when he needs it, and he relies on the information (trusting its veracity), thereby we could say that Otto's mind extends into Inga's. Supporting this idea is a growing body of research on the social distribution of cognition, which tends to focus on the psychology of memory and group decision-making. The theory of transactive memory developed by Daniel Wegner, for example, maintains that memory processes (including encoding, storage, and retrieval) are sometimes shared across stable dyads (with a particular focus on highly interdependent couples) and groups.[12,13] Wegner explains that transactive memory systems involve a "set of individual memory systems in combination with the communication that takes place between individuals" (p. 186).[12] Memory processes are, thus, not reducible to the internal processes of any particular individual in the system, as the communication between individuals is an essential part of the process. Extended mind theorists have appealed to this paradigm as one example of how a single agent's cognitive process can be distributed across multiple agents.[14]

One reason that AI extenders present a paradigm shift, as we will argue in the next section, is that in some important respects they are more like cases of social extension than extension into static artifacts. For one, AI extenders have some sophisticated cognitive capacities that have previously only existed in humans, e.g. speech recognition, facial recognition, pattern recognition, and so forth. This means that while traditionally there may have been some cognitive tasks that could have only been completed through socially distributed systems, in the future there may be opportunities for individuals to rely on AI extenders rather than other people. This could have plenty of upside for people who are deficient in certain capacities, like executive control functions, and who have traditionally had to rely on members of their social network for help—we will discuss the condition of borderline personality disorder as an example of this in the next section.

There are several advantages to relying on an AI extender over another person. For one, in the case of highly interdependent couples it is often

found that each individual relies on the other in various ways (e.g. Inga remembers where the museum is located, while Otto remembers the best place to park), so there is typically some degree of reciprocity, or symmetry.[13] But asymmetric socially extended beliefs might also be possible. For example, Clark and Chalmers discuss the possibility that someone relies on his regular waitress at the restaurant he frequents to determine what food he should order, thereby offloading his decision-making.[1] These asymmetric cases seem to imply, however, that one person is always being paid or 'used' for cognitive labour. AI extenders present the opportunity to asymmetrically rely on a cognitively sophisticated device. Furthermore, other agents may be less stable than an AI extender device, which one can own and carry with them everywhere they go.

3 What is an 'AI extender'? How this differs from standard kinds of cognitive extension

As mentioned, the tools that are often cited as examples of cognitive extenders include both simple technologies, such as Otto and his notebook[1], as well as more advanced tools, such as smartphones[15]. Both of these technologies had transformative effects on human cognition. The use of writing tools to create external symbols and write words down represented a major shift in human intellectual history: a move from the oral tradition to literacy. The smartphone has likewise been transformative—a democratized and powerful personal computer that travels with users wherever they go. While extended mind theorists have discussed smartphones[15], it rarely gets mentioned how the computational power of smartphones has grown dramatically over the last few years, not only because of their processors but also because of the connected use of cloud services, with many apps running or refining pre-trained deep neural networks and other technologies. We argue that this increased use of machine learning, and other functionalities brought by artificial intelligence, is qualitatively different from the kinds of cognitive extension that preceded it in several ways: these systems can perceive, navigate, make complex decisions, recognize and produce language, plan, identify emotions, etc., all in complex and changing situations.

To more precisely characterize what an AI extender is, we start from a definition of *cognitive extender* as given by Hernández-Orallo & Vold, which was adapted from Hutchins:[2,16]

> A cognitive extender is an external physical or virtual element that is coupled to enable, aid, enhance, or improve cognition, such that all – or more than – its positive effect is lost when the element is not present.

Again, crucially, the external physical or virtual element must be *appropriately* 'coupled' to be considered a cognitive extender. That is, the right conditions must be met in order for an artifact to count as a literal part of an agent's mind (i.e. on a par with the brain); and only when these conditions are met is the mind extended.

In contrast, AI extenders are an increasingly important and distinctive subkind of cognitive extender that are distinguished by their use of a particular kind of technology – AI – and their distinct implications (to be discussed in later sections). Here is a more precise definition:

> An AI extender is a cognitive extender that is "fueled" by AI. This means that some AI technology is directly responsible for the cognitive capability that the extender is able to deploy, in conjunction with its user.

With the above two definitions, what counts as an AI extender is precise, as far as what count as AI is precise. Today, we associate AI with a range of possibilities such as image and speech recognition, machine translation and planning, many of them realized through machine learning. However, AI will cover more and more *cognitive* capabilities, as we further characterize in section 5.

Those working in psychiatry and clinical neuroscience may be familiar with related terminologies such as "cognitive assistants" or "cognitive prosthetics". It is therefore important to be clear about what distinguishes

AI extenders from these well-entrenched applications and what justifies the introduction of this new category. The crucial distinction here is that, as a subspecies of cognitive extenders, AI extenders challenge the internalist view of the mind: AI extenders are a part of an agent's wide existing cognitive system, they perform cognitive functions for a human agent—just as the brain does. It is this rather strong metaphysical claim that distinguishes them from mere "cognitive assistants" or "cognitive prosthetics". This metaphysical claim is important for many ethical and policy implications that we will discuss in the next sections, and how these systems have to be regulated and built. For instance, following the examples we gave in the introduction, the definition above includes Helen's augmented reality glasses, which means that they have to be designed, in the first place, by considering how Helen's cognitive processing is going to change with their use. This may be an important, but secondary design principle, when considering some other devices that the above definition excludes, such as autonomous vehicles or cognitive robots that interact with humans occasionally (but are not tightly coupled), or an exoskeleton or a "smart" shoe that use ML algorithms to stabilize the body, but are not really providing cognitive functions to the user.

To be more precise, we might say that AI extenders (and cognitive extenders more broadly) are themselves a subset of cognitive assistants, while not all cognitive assistants are extenders. Certainly, AI extenders provide cognitive assistance or "cognitive services".[17] For example, Helen's augmented reality glasses and Lewis's monitor system perform sophisticated cognitive processes: from perception to planning, and as such they both serve as cognitive assistants. But AI extenders are distinguished from the broader category of cognitive assistants by the degree of *coupling* (following, e.g., the dimensions from Heersmink, as mentioned above).[11] It is this tight coupling that, so the argument for the EMT maintains, warrants us calling them a part of the agent's mind and thereby challenges the commitment to internalism. A cognitive extender is not merely processing inputs and producing outputs (such as an online machine translator); it is the locus of states that are created and accessed at any time by the human agent.

AI extenders should similarly be distinguished from cognitive prosthetics (sometimes also referred to as "orthosis"), a term that comprises many systems that can be attached to (and detached from) humans and can help or completely replace some lost or nonexistent cognitive human function. Originally, a cognitive prosthesis was defined as "a compensatory strategy that changes the environment and focuses on functional activities [...] designed specifically for rehabilitation purposes" (p. 41).[18] However, many so-called cognitive prosthetics are simple software or hardware devices, that are not tightly coupled and that are not "fueled" by AI; in some cases there is no information processing or otherwise intelligent processing happening on them, like a stick compared to a leg. In other cases, no attachment (or coupling) takes place. In learning environments, for example, any device in a classroom is said to be cognitive prosthetics.[19] In our view, even if the trend today is to use the notion of cognitive prosthetics for interventions that involve some computing technology[20], many of these cannot be considered AI extenders due to a lack of appropriate coupling. As a result, many cognitive prosthetics do not carry with them any interesting metaphysical claim. They do not challenge the internalist picture of the mind, and as a result they come with a distinct (though perhaps overlapping) set of risks and opportunities from those we will discuss around AI extenders.

The categories of AI extenders and cognitive prostheses may be overlapping at times (i.e. they are not mutually exclusive) but they are also not identical. Some cognitive prostheses really are appropriately coupled to an agent and do make use of AI technologies. For instance, one of the early AI extenders was COACH (Cognitive Orthosis for Assisting aCtivities in the Home), a device that uses AI to observe, supervise and assist people with dementia, "learn from his or her actions, and issue prerecorded cues of varying detail"[21] or Solo, another prototype that used planning and other AI techniques to help "cognitively impaired clients and their caregivers in managing their daily activities".[22] These intelligent assistance devices for people with dementia are perhaps the best current clinical examples of AI extenders. Many of these research prototypes are now superseded by commercial products, and in cases targeting the general public, such as Ellie, Woebot and Tess, with some of them known as virtual cognitive behavior therapists[23] a term borrowed from the early days of ELIZA, the famous

computer therapist.[24] Because the relevant dimensions that characterize the appropriate coupling necessary for cognitive extension are relational in nature, these AI-driven cognitive prosthetics may in some cases also count as AI extenders.

Our definition of AI extenders also suggests why second-wave arguments for the extended mind thesis are better able to support the possibility of AI extenders. This is because the way that machine learning systems process information is likely to be relevantly dissimilar from the ways that humans do. Furthermore, as we have seen, the performance capacities of AI extenders far exceed what a notebook or a calculator can do, or even (in some respects) what a human mind can do. Indeed, AI can do much more than analogous functions (as a parity-driven argument for the EMT would require). As covered by this and other volumes, AI can lead to better diagnosis, prognosis and treatments in mental health[25], and robotic and virtual systems are treating people with dementia, autism, and other conditions, educating children with developmental disabilities, on top of a range of possibilities for training, consultation, healthcare management. Meanwhile, the area of affective computing is making machines able to detect and react to emotional states, where machine learning can create high-level representations from sensors on the body and brain-computer interfaces, detecting normal and abnormal situations.

Under the scope of AI extenders we consider all these possibilities, with the condition that the system must be appropriately coupled with the person such that the effect is lost without the extender. An occasional or detached use of a robotic therapist is not an AI extender (not coupled). The use of augmented reality to treat a phobia (so that the patient is "cured") is not either (the effect is permanent). Though both of these technologies might be considered as cognitive assistants.[26]

Finally, this range of examples of AI extenders is indicative of just how broad and inclusive the category is intended to be. It can include a rather heterogenous set of technological applications. A companion robot could be an AI extender for the same reasons that we consider social extenders to be, for example. Some kinds of ambient intelligence (beyond smart homes and

buildings), such as the Persuasive Mirror, could also count.[27] Probably the most obvious cases will involve software tools, such as decision-making support systems, or tools that are designed in ways that easily satisfy the relevant dimensions of coupling: applications on one's smartphone, for example, are well-suited to fit these criteria, because of how portable our phones are, how much personal information they track, how readily accessible their applications are to us, how likely we are to trust and rely on the information they provide us, and so forth. Even though AI extenders can be heterogenous in terms of their physical properties and instantiations, what makes them a cohesive category (distinct from cognitive assistances, cognitive prosthetics, and even other kinds of cognitive extenders), worthy of discussion is the role they play in the cognitive lives of humans and the ethical and design considerations that emerge from this context. This is true even though we are still in early stages of developing AI for use in clinical settings (especially systems that interact directly with the user). For this reason, our chapter focuses more on future possibilities around how AI could be used to extend cognition, in the context of mental health, exploring the risks, and the design and policy implications around how we might deal with these future scenarios.

4 How the extended mind can change our understanding, assessment, and treatment of cognitive disorders

By now, numerous authors have described how the EMT can improve either how we understand, assess, or treat mental and behavioral disorders though few have focused specifically on the kinds of intelligent assistive technologies that we have termed 'AI extenders'.[28-33] In this section we will review some of this work.

When it comes to *understanding* disorders, the central point that extended mind theorists tend to make is that there can be constitutive factors that lie outside the brain, and hence to fully understand a disorder one cannot look to the brain alone. Simply put, there are cases of cognitive impairment that do not involve impairments of the brain. The issue of *assessment* is related to this point. Some of the standardized tests for cognitive function assume

an internalist picture, focusing only on what the brain of a patient is capable of (for example, by testing them without tools or assistive technologies). In doing so, these tests often disregard the real-life circumstances of the individual, which may involve the use of tools that make essential contributions to their cognitive functioning.[31-33] As a result, test scores can skew the picture of how 'well' a patient is really doing and what they are really capable of. Hence, even if a patient has a cognitive impairment with a neuro-explanation, this might not impair their functioning in everyday life.

Finally, the matter of *treatment* is about how to view the different techniques available for rehabilitation. Researchers working on cognitive impairment in various domains have drawn distinctions between different kinds of rehabilitative strategies,[34,35] which several extended mind theorists have employed to help draw out the difference between the internalist and externalist views on treatment.[31,33] "Restorative" strategies aim to directly address an individual's cognitive impairment by restoring their ability to perform tasks *in just the same way* that a non-impaired individual would. "Compensatory" strategies, on the other hand, attempt to circumvent impairment by helping the individual perform the same tasks but in different ways, namely by using assistive technologies.[31] Cognitive prosthetics and cognitive orthosis tools, like COACH and SOLO discussed above, were built as compensatory strategies—ways of substituting for biological deficits that could not be directly addressed.[18,21] Because the internalist picture says that all cognition is a function of the brain alone, on this view restorative strategies must involve repairing one's internal neuro-capacities, as this is the only 'true' way to improve a person's cognition. King explains that an internalist would view compensatory strategies as a second-best option; while assistive technologies might help an individual *compensate* for impairment, they do not actually fix the problem.[31] In contrast, because the EMT sees cognition as constitutively involving more than just the brain, it can view both rehabilitative strategies as genuinely restoring cognitive capabilities.

In what follows we will discuss five illustrative examples of cognitive disorders that extended mind theorists have argued can be understood in the light of the EMT.

4.1 Alzheimer's disease

In their now much-discussed example, Clark and Chalmers describe Otto as suffering from Alzheimer's disease, a degenerative cerebral condition characterized by a slow deterioration of multiple higher cognitive functions.[1,3] They describe Otto as being able to function normally, despite his deteriorating biological memory, by relying on his 'extended' memory, namely, the information that he records in his notebook. The example suggests that by taking this wider view of the mind we might develop new ways of assessing and treating Alzheimer's, as well as other kinds of dementia.

Drayson and Clark discuss a compelling real-life case that brings this point to life.[33] An inner-city group of Alzheimer's sufferers had scored so dismally on standard tests for Alzheimer's, such as the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) protocol, that doctors had anticipated the patients would need to be relocated to full-care hospitals. Yet, the patients perplexed doctors as they continued to be able to cope with the demands of daily life and to successfully live alone in a complex urban environment. Upon making home visits, doctors found that these patients had each transformed their home environments, creating ingenious personalized cognitive tools, props, and aids that supported their memory: from open storage cabinets, to notes and labels indicating what to do and when, as well as who each person was in their family photos.

Because the CERAD protocol only tests biological memory it could not explain how these patients continued to effectively function in the world. What is more, if doctors had taken the tests of internal memory as the only standard, these patients might have been forcefully relocated to controlled hospital settings much sooner than necessary. Drayson and Clark note that the relocation of Alzheimer's patients is often a fateful turning point in which their conditions become more severe.[33] From the EMT perspective, this is not surprising; one would predict that removing patients from their supportive environments would have a detrimental effect on their

functioning. The same could be said about the example of Helen, who we discussed in our introduction. Helen relied on augmented reality glasses that had been designed to support her memory and perception. Her case is an evolution where most of the physical cognitive tools, props, and aids in her home are replaced by a single tightly coupled device, an AI extender, going much beyond what Drayson and Clark found in the real case. Hence, by taking the wider 'extended' view of memory and other cognitive functions not only are we able to explain how these patients could continue to function, despite low test scores, we might also rethink how we assess and treat sufferers of Alzheimer's.

4.2 Learning disabilities and disorders

A learning disability affects the way a person is able to problem-solve, plan, and acquire new knowledge and skills.[3] King argues that adopting the neuro-centric internalist view of the mind commits one to problematic views about the cognitive capabilities of learning-disabled individuals, whereas the EMT allows us to more accurately assess and treat them.[31] King describes the fictional case of learning-disabled woman named Dana. Without any tools, Dana struggles to compare and evaluate various relevant factors for making decisions. She is, however, perfectly capable of making good complex decisions when she uses a graphic organizer such as a Venn diagram. This visuo-spatial way of representing information allows her to evaluate relevant factors, and to reach a good conclusion. Indeed, when using this assistive technology, Dana's decision-making skills are just as good as anyone else's (p. 49).[31] So is Dana capable of making decisions or not? How should we assess her capacities?

Notice that Dana's use of an assistive technology counts as a compensatory rehabilitative strategy—much like the Alzheimer's patients who had relied on external resources in their environments. Hence, internalists would have to maintain that Dana has not really been 'cured' because her capacities have not been restored, and so, even with the assistive technology, she cannot really make good complex decisions. Indeed, King nicely explains that the internalist is committed to an inverse relation between the extent to which

an individual relies on external tools and the extent to which *she,* as an agent, is really engaging in some cognitive process. This means that Dana is only "doing" as much, cognitively speaking, as her neurons are doing (p. 56).[31] And therefore, she only merits "cognitive credit" for what her neurons do, not the cognitive work that is done by whatever assistive devices she employs. The internalist is, thus, forced to say that Dana has less cognitive capacities and deserves less cognitive credit than someone who could perform the same task 'intracranially'. This carries practical implications, as by adopting the stricter restorative conception of "cure" we may give preference for alternative internalist treatments, even if they are less effective (or have undesirable side-effects) for the wellbeing of the patient.

King argues that we should resist internalism in favor of the EMT, which instead allows us to say that while Dana may need to rely on graphic organizers, she is quite capable of making complex decisions. On this view, teaching Dana how to effectively use assistive technologies (that will be readily available in her everyday life) is as good as any restorative strategy. Furthermore, Dana should get the same "cognitive credit" as someone who is able to complete a similar task without the assistive technology (i.e. internally).[31] Heersmink and Knight have similarly argued that education and assessment should take into account how agents are able to assemble and use tools in their environment as extended cognitive systems, discussing in particular students' use of the internet during exams.[32] We can draw a similar lesson for the 10-year old, Lewis, who (as we described in our introduction) relies on an AI extender to help aid some of the learning-related symptoms of ADHD. If one takes the extended approach, Lewis should get recognition and credit for his improved academic results, even though he could not achieve these results without his assistive device.

4.3 Addiction

The ICD-10 clinical definition of dependence syndrome, henceforth 'addiction', includes a cluster of physiological, behavioral, and cognitive phenomena in which the use of a substance takes on a much higher priority or value in one's life than is usual (p. 69).[3] Diagnostic symptoms include a

compulsion to take the substance, difficulties in controlling substance-taking behavior, and neglect of alternative pleasures or interests. Levy argues that adopting the EMT is useful in treating addiction because of how it can help support self-control interests.[i, 28]

Internalism promotes the idea that the only way to recover from addiction is to change one's mind: addiction is entirely a matter of 'will-power' and the addict just needs to 'say no' to their cravings (pp. 219–220).[28] On this view, addicts tend to be held more responsible for not overcoming their addictions. But Levy cites research on 'ego-depletion'—the idea that self-control draws on a limited pool of (internal) mental resources—which suggests that addicts have depleted self-control and thus they experience more difficulty in resisting their cravings than one who craves but is not addicted. If this is true, then it may be "literally impossible" for an addict to resist taking the substance they crave when it is available to them and their will-power is depleted (p. 219).[28] the EMT is useful here as it points us to the agent's wider environment to look for new methods of treatment. Levy suggests, for example, that environmental modifications, including the use of technology and social support systems, can be quite effective at helping overcome addiction (other work in 'positive computing' also supports this).[36] Social support is perhaps closer to the use of AI extenders insofar as they can detect when a person is feeling weak in controlling their impulses and needs coaching or nudging.

As just one example, a system could learn what kind of situations make self-control more difficult for a particular individual. For instance, smokers usually associate tobacco with some situations (e.g., pubs, coffee, etc.) and less with others (e.g., going for a walk). So for a particular person, a machine learning system could detect that the person is likely to have depleted self-control when meeting with certain friends or going to certain places where they used to smoke. Suggesting walking routes that avoid smoking zones, or even reminding the agent that these situations may be challenging (and

[i] This is not an exhaustive list, but rather a selection from the ICD-10 Diagnostic guidelines for Dependence syndrome.

perhaps directing them to resources) might help them in controlling their impulses.

4.4 Borderline Personality Disorder

Bray argues that the EMT can offer a better understanding of certain personality disorders, such as Borderline Personality Disorder (BPD).[29] Personality disorders usually involve lasting and inflexible adverse patterns of thinking and feeling about oneself and others that impair how an individual functions in many aspects of life.[3] Because they can include affective dysregulation, cognitive and perceptual distortions, and impulsive behavior, they tend to be thought of as a subclass of mental and behavioral disorders.[ii,3] BPD is characterized by emotional instability and disorders of mood that affect how a person relates with others. Its symptoms include a deficit in one's ability to perform certain high-level cognitive tasks, such as emotional regulation and impulse control, disturbed patterns of thinking or self-perception, and "a liability to become involved in intense but unstable relationships" (p. 160).[3]

Bray suggests that because people with BPD have a meta-cognitive deficit, it is possible that they are more likely to rely on those around them as a way to help supplement their internal, biological deficit.[29] Recall the possibility of social cognitive extension—where one agent relies on the information and abilities of another agent as an 'extension' of her cognitive system. Bray argues that people with BPD may form particularly close dyads with others, especially romantic partners, friends, or family members, in an attempt to make use of their executive functions, as this is the only kind of coupling that can fill the deficit in their own biological cognitive system.

This could explain why people with BPD tend to form "unusually intense" relationships, why they suffer from a fear of abandonment, and why they are typically "devastated" when these relationships end.[29] As we say above, in order to cognitively extend, one needs a tight coupling with the external

[ii] They are categorized as mental disorders in the ICD-10.

element, e.g. a stable, reliable, and high-bandwidth connection with the 'extender' (among, perhaps, other features).[11] But reliance on others has certain inherent drawbacks, Bray explains: "Imagine what it would be like if important parts of your own brain were able to detach themselves at will and wander away for indeterminate periods, perhaps never to return."[29] For the BPD sufferer, this is how he views the people in his life, with whom he has formed close 'couplings'. If these people up and leave, he would be left without the ability to self-regulate, to control his emotions or impulses. This also points to one way in which AI extenders could be used for treatment—AI extenders could replicate some of these metacognitive skills while at the same time being more reliable than social extenders. We will pick up on this below where we discuss the benefits of AI extenders.

4.5 Autistic Disorders

Autistic disorders are 'pervasive developmental disorders' characterized by deficits in social interactions and social communication, and restricted and repetitive patterns of behaviors, interests, or activities (p. 198).[3] Krueger and Maiese maintain that while there is no current consensus on the cause of autism, the most popular explanations over the last several decades appeal to a Theory of Mind deficit.[37] The result of this has been to think of the disorder as a disturbance confined to the head of the individual—that is, to assume an internalist perspective. This in turn has shaped the typical treatment and intervention strategies, which are generally aimed at helping individuals develop their mind-modelling capacities. While this may be one helpful technique, Krueger and Maiese argue this perspective overlooks the fundamentally embodied and relational factors which contribute to autism and in doing so also overlooks potential treatment strategies.[37]

Krueger and Maiese argue, for example, that people with autism typically suffer from "style blindness"—they have "a perceptual inability to extract socially salient information from the qualitative kinematics of others' actions" (p. 24).[37] This explains why they often cannot pick up on subtle social cues (e.g. non-verbal communicative behaviors, such as gestures and facial expressions) and why they struggle to understand figurative language.

But there is evidence that this lack of access to social norms is not "in principle", as people with autism can access and abide by social norms of their own group (i.e. other autistics) and when norms are made explicit (p. 23).[37] Hence, instead of only employing restorative strategies with the expectation that people with autism need to develop their internal mind-modelling capacities, we might also use AI extenders for compensatory strategies aimed at making social cues (such as categorizing the tone of voice or body language for the user) and the meaning of figurative language more 'visible'.

Another example is how we think about and treat the characteristic movements and behaviors of people with autism, which can consist of "hand-flapping, finger-snapping, tapping objects, repetitive vocalizations, or rocking back and forth" (p. 27).[37] According to Krueger and Maiese, these are typically viewed as meaningless reflexes or nervous tics, but in fact these behaviors (sometimes called "self-stimulations" or "self-stims") may be strategic deployments used to organize incoming sensory information—for example, to occlude signal noise when incoming information threatens to be overwhelming, or to heighten arousal in order to better access salient information (ibid). But standard 'internalist' treatment programs have traditionally tried to eliminate or suppress self-stims, whereas a wider approach could recognize their important role as embodied cognitive coping strategies, and even try to foster these strategies. For instance, an AI extender could be designed to find and produce appropriate stims (the most effective and least visible for other people).

5 The specific effects of AI extenders on mental health

An AI extender can come in different forms: a device (e.g., a tablet), a wearable (e.g., a watch), or an app or service that is available across different platforms (physical personal assistant and computer). These tools can be generic (e.g., a navigation system or an agenda), can be addressed to a range of mental health issues (e.g., a monitoring system) or can be devised for a particular condition (e.g., an anti-stress app).

If we start with generic AI extenders, they are usually devised to improve or compensate for one or more cognitive abilities. In Hernández-Orallo & Vold, we identified 14 cognitive abilities in which AI can extend cognition (the full definition can be found in that paper).[2] These capabilities are reproduced in Tab. 1. The table also includes examples of AI extenders for each ability, either in general (non-clinical applications, second column) or for mental health (clinical application, third column). General applications may be motivated by comfort or efficiency; e.g., most people do not use GPS navigation devices to compensate for any limitation, but as an enhancement. Clinical applications of AI extenders aim at—although not exclusively—compensatory uses. The capabilities in Tab. 1have effects on many daily tasks and are expected to generate a range of applications as soon as AI can enhance or compensate them.

Capability	General Examples	Clinical Examples
MP: Memory processes	Automated reminders or prompts; new customized mnemonics to improve long-term memory, or tag our experiences with related people, concepts and other situations to improve episodic memory.	Apps telling an Alzheimer's patient whether something has already been done, said or visited before.
SI: Sensorimotor interaction	Pattern-recognition systems; mixing representations through generative models; intelligent sensors and actuators	Haptic/robotic clothing aides for people with Parkinson's disease.
VP: Visual processing	Object and scene recognition or colour-recognition tools for visually-impaired; facial recognition; augmented reality; intelligent filters and lens	Scene sketchers to contrast visual hallucinations in patients with schizophrenia
AP: Auditory processing	Voice-to-text applications for hearing-impaired; highlighting parts of speech that might be missed; following multiple conversations and prompting the user based on modeled interests; music apps.	Ambient and speech recognition systems to contrast auditory hallucinations in patients with schizophrenia
AS: Attention and search	Modelling user interests and goals to focus our attention; e.g. through text search or summaries, web search engines, or with object recognition	Attention focusing devices for patients with attention deficit disorders
PI: Planning	Automated agendas, daily task planners, and prompts based on modelling of user's goals and interests	Daily task organizers for a patient with moderate mental retardation
CE: Comprehension and expression	Digital writing assistance tools using natural language processing (e.g. Grammarly); automated re-writing or re-rendering to improve interpretability (for reading, watching films, listening to music)	Vocabulary and grammar assistance tools for a patient with some language disorders

CO: Communication	Automated emails, social media posts; improved intelligence in communication; effective spreading memes or ideas.	Effective communication tools for people with Asperger's
ES: Emotion and self-control	Systems that predict and inform us of our emotional states and those of others, help us detect fake emotions in others, or trigger our emotional responses.	Systems recognizing emotional states of people for patients with autism.
NV: Navigation	GPS apps (e.g. Waze), building associations between places, routes, and our cognitive states to help with route-finding, safe-walking, or orientation	Route assistants to safely navigate surroundings for people with dementia
CL: Conceptualization, learning and abstraction	Machine learning apps helping find new categories, concepts or possibilities, e.g. new patterns about daily or public events; new personalized learning strategies.	Personalized learning assistants for people with learning disorders or disabilities.
QL: Quantitative and logical reasoning	AI systems that process uncertainty (e.g. risk or number of accidents), or quantities (e.g. people in a room) in real time	Diet analyzers and estimators for patients suffering from anorexia.
MS: Mind modeling and social interaction	Modelling social networks to help anticipate decisions, actions, and interests of other people	Apps determining double meanings in conversations, for people with Asperger's.
MC: Metacognition	Self-tracking and analysis can help identify the potential and limitations of users, making users more aware of their own capacities	Systems monitoring self-esteem and confidence in depression episodes.

Table 1 The left-hand column indicates fourteen cognitive capabilities that can be extended by AI, the middle column provides examples of the kind of AI application that can achieve this (full account in Hernández-Orallo & Vold 2019). The right-hand column shows particular clinical examples in mental health.

Having limitations for one or more of them can indirectly (i.e. as a side-effect) cause many mental conditions, such as stress, depression or loss of self-esteem, as the subject feels unable to do things that other people do easily. All these capabilities have effects on cognition and development, so they are linked to mental and behavioral disorders in one way or another (i.e. the consequences and side-effects can be numerous). In other words, any extender using AI that can alleviate or compensate some cognitive limitations could have, in principle, positive effects on some of these conditions. This is sometimes referred to as the "mental capital — the cognitive and emotional resources that influence how well an individual is able to contribute to society and experience a high quality of life", and increasing this capital could "mitigate the risk of disorders such as

depression, substance-use disorders, bipolar disorder and dementia" (Sahakian, p.c.).[38]

In order to determine the way AI extenders can impact on various mental conditions, we analyzed one standard classification of mental and behavioral conditions, the WHO's ICD-10 Classification of Mental and Behavioral Disorders.[3] This classification (as explained in the ICD-10 "blue book") is a comprehensive list including the clinical descriptions of the conditions. One shortcoming is that it excludes some related conditions, such as Parkinson's disease, which are classified as diseases of the nervous system, but whose symptoms might nonetheless be improved by AI extenders addressing the 14 capacities in Tab. 1 (e.g. changes in communication, or sensorimotor function).

From this list of conditions (each with an ICD assigned code starting with "F"), we identified those that are associated with each of the cognitive capabilities shown in Table 1. By 'associated' we mean that a variation in the cognitive capacity, i.e. either an increase or decrease, will have a direct effect on the mental or behavioral condition. For each capability we searched through the ICD-10 for a series of tokens[iii]. For instance, for sensorimotor interaction (SI), we looked for "sensor*" and "moto*", and checked (manually) whether the reference made sense (e.g. was it actually talking about an association in the form of a cause or a symptom?). Table 2 shows the result of this analysis. It highlights which of the 14 cognitive capabilities have a direct effect on mental health conditions.

[iii] The full list of items used for each capability can be found in the Appendix A.

ICD-10 Condition	Cognitive Capability												
	MP	SI	VP	AP	AS	PI	CE	CO	ES	NV	CL	QL	MS	MC
F00-F03 Dementia (Alzheimer's, Vascular, Others, Unspecified)	■	■	■		■				■	■	■	■		
F04 Amnesic syndrome	■										■			
F05 Delirium	■		■	■	■		■			■	■			
F06 Other disorders due to brain issues	2		0	0							7			
F07 Personality and behavioural disorder due to brain issues	2			0		0	0		0		2			2
F1x Mental and behavioural disorders due to brain issues			0,5	0,5					2		6			
F20 Schizophrenia			■	■			■	5	■				■	
F21 Schizotypal disorder				■										
F22 Persistent delusional disorders				■										
F23 Acute and transient psychotic disorders			0	0,3										
F25 Schizoaffective disorders														0
F30 Mania			1	1										1,2
F32 Depressive episode				3	■				■					■
F40 Phobic anxiety disorders													■	1
F43 Reaction to severe stress, and adjustment disorders					0				23	0			■	
F44 Dissociative [conversion] disorders		6	6		3				■	0				
F51 Nonorganic sleep disorders										3				5
F60 Specific personality disorders						4	4		■				■	■
F63 Habit and impulse disorders									■					
F71 Moderate mental retardation							■				■			
F73 Profound mental retardation			■				■	■			■			
F80 Specific developmental disorders of speech and language				0			0,1,3						2	
F81 Specific developmental disorders of scholastic skills	0		■	0			0				2	2		0
F82 Specific developmental disorders of motor function		■												
F84 Pervasive developmental disorders (e.g., ASD, Asperger's)					4		0	■	■		■		■	
F90 Hyperkinetic disorders (or "attention deficit disorder"					■									
F91 Conduct disorders													■	■
F94 Disorders of social functioning (children and adolescents)													■	

Table 2 Association between mental conditions (rows) and the capabilities that can be extended by AI (columns, as per Table 1). A black solid cell means that the whole category of mental conditions is affected by the capability, a grey cell means that this is the case only for some of the subcategories (the code of the affected subcategory or subcategories appears in the cell), and empty cells mean no association.

There are a few things to note about Tab. 2. First, it must be understood as showing 'direct' effects of AI extenders rather than any potential side effects they might have on mental conditions. For instance, a visual processing system (so providing VP capabilities) that objectively describes what is on the patient's scene might help discard the false perceptions that come from hallucinations (as captured by "F05 Delirium"). This applies not only to visual and auditory inputs, but also some other misperceptions (e.g., "that person is looking at me all the time", or "he is following me", etc.). In this way, AI systems can be an alternative source to perceive reality, which can help, in some cases, discard those false perceptions (sensory, emotional, etc.) that are common in many mental conditions. Note that in this example, the AI extender has a direct effect on mental conditions that involve

hallucinations. These same systems might also have the side-effect of improving conditions that involve one's reasoning or planning, e.g. personality disorders, but we do not include these in the corresponding right-hand column.

Second, for many of these conditions the ICD-10 explicitly states that the cause is unknown, and the clinical descriptions include lists of symptoms and diagnostic guidelines that are based on an assessment of the presence or absence of certain features or characteristics. Hence, in many cases the best that we can predict is that AI extenders will have a direct effect on alleviating the symptoms (rather than addressing the causes) of the conditions listed.

Finally, Tab. 2 can be used to recognize the potential of an AI extender featuring a capability (or research in one particular area of AI) for a range of mental conditions. For instance, it is no surprise that MS (mind modeling and social interaction) is associated with many conditions, but it is was less expected perhaps that AP (auditory processing) had such a number of repercussions. This is especially interesting as the state of the art of AI in auditory perception has improved significantly during the years, and its integration with hearing aids may be on its way. Table 2 can also be read in the other direction. If we want to treat or improve the state of patients having some particular condition, we must look at the matrix and see what cognitive capabilities we need to imbue on a system. For instance, sleepwalking (or 'somnambulism') could be treated with some device that, through the use of AI, could follow where the patient is moving and check for obstacles and hazards. The table is a first approximation, but it can be valuable to have a first understanding of the many possibilities of AI (and AI extenders in particular) for mental health.

From all the abilities in Tab. 1, metacognition is perhaps the most critical one to discuss. This is for two reasons, first because of the methodology we had to employ in searching the ICD-10 bluebook for associated conditions in Tab. 2, and second because of how it is (we believe) associated with so many different disorders. In the first case, we note that the term

'metacognition' does not appear at all in the ICD-10 bluebook.[iv] Nonetheless we believe that metacognition has wide associations with many of the conditions listed because of how a patient must realize their own limitations related to their particular condition. Many patients, for example, improve simply by being diagnosed ("Now I understand what is happening to me"). Relatedly, treatment and care are much easier when this is known.[39] In the context of cognitive extension, however, it is very important that the person realizes how the added AI extenders change the person's capabilities, so what the person does and what the person *thinks* she or he does—which the diagnosis clarified—are kept aligned with the use of AI extenders. A planned and temporary removal of an AI extender can be very helpful for this alignment, in the same way that hearing-impaired people realize how bad their condition is when they compare hearing with and without their hearing aids. This is also related to a placebo effect that may appear with the use of AI extenders, simply because the person thinks that he or she now has "superpowers" or a subsystem to rely on, which may boost his or her confidence.

Other AI extenders may be more focused towards monitoring and intervention rather than enhancing or replacing some cognitive capabilities. A monitoring system using machine learning to determine when a person is more likely to have an outburst or a crisis can be considered an AI extender as much as it extends our self-awareness, in the sense of an internal perception of indicators in our bodies that we can understand and react to accordingly. If the tool also makes recommendations or interventions, we can still consider it an AI extender, which helps the patient with self-control, awareness of the situation, or simply suggesting the best actions, and so forth. In other words, monitoring and recommendations can be seen as extenders at the metacognition and decision-making levels.

6 Potentialities and Challenges of AI Extenders

[iv] We instead had to search for related terms such as awareness, capabilities, limitations, consciousness, self-confidence, etc. (see Appendix A for a complete list).

There are many potential benefits of the use of AI extenders for mental health—both for helping those who are cognitively impaired (which is our focus in this chapter), but also for healthy users, who rely on digital devices as cognitive enhancers. In this section we will focus on five benefits, followed by five risks.

1. AI extenders inherit all the benefits of using non-invasive treatments, something that is shared with (physical) orthopedics, in terms of flexibility, updates, repair and removal. The use of machine learning can (a) improve the degree of personalization, as systems can learn and improve their behaviour from the information they collect, and (b) make sense of a wider data set about one's lifestyle (i.e. one that looks beyond the biological individual) than a doctor ever could—including information about one's social life, screen time, the environments one spends time in, etc. Collecting and analyzing this wider data set could eventually allow for a better understanding, assessment and treatment of mental health conditions. These benefits are available for both cognitively impaired and cognitively health users.

2. Under the lens of the EMT, we can consider the use of an AI extender as a genuine cure provided the system is reliable and integrated. Traditionally, a "cure" is some intervention that aims to directly address the cognitive deficit by making underlying mechanistic functions work better or by limiting their negative effects (what we describe as 'restorative' strategies above). With an AI extender, however, we instead aim to design a system that is able to compensate for those malfunctioning underlying mechanisms. With the right kind of device integration (or 'coupling'), if a person gets used to giving a description of a scene or determining false memories or perceptions from true ones using a device, this could ultimately be incorporated as part of their cognition, and help cancel or replace those malfunctioning biological processes. When this happens, we argue, the new situation can be assimilated to being "cured" or "back to a safe condition".

3. AI extenders may be a good option for those cases where restorative strategies through internal interventions, such as medications (e.g. antipsychotic drugs) for improving conditions like meta-cognition or mind modelling, may not yet be available. There is no known restorative strategy for dementia, for example, but Ienca et al. note that a wide range of intelligent assistive technologies are being developed to provide general cognitive support aimed at "empowering" adults with dementia.[40]

4. The tight coupling of AI extenders makes it easier to give 'cognitive credit' to the person for their accomplishments. Returning to the case of Lewis, who relies on a device to help him cope with symptoms of ADHD that affect his learning, the EMT allows us to still credit Lewis as learning, while the regular presence of the extender makes it easier (like a pair of glasses). We described Lewis as being allowed to use his assistive device during examination, for example, which also makes sense under the EMT, as the device is really part of the substrate of his cognition. Under the EMT we would have to give a similar analysis of the cognitive accomplishments of cognitively healthy users of AI extenders as well: they deserve credit for what they achieve with their device.

5. Finally, AI extenders can provide more sophisticated resources than regular extenders. Consider again the case of BPD discussed above. Bray had hypothesized that people with BPD tend to rely on others in order to compensate for their internal deficits of executive functions because this is the only available option to them, and that this explained their characteristic fears of abandonment, and losing their autonomy.[29] But AI extenders could potentially provide the same support for meta-cognitive deficits as other people could, only with increased stability and reliability. Again, this can be a benefit in both clinical settings, but also for the cognitively healthy, looking to enhance their abilities.

The negative side-effects of AI extenders for mental health can be varied. Some of them are also shared by other extenders or cognitive enhancers and

are related to the four basic principles of medical ethics—respect for autonomy, justice, beneficence and non-maleficence,[41,42] but others are more specific to AI extenders (when used both in clinical settings by the cognitively impaired, and for enhancement purposes, by the cognitively healthy). The reason is that the use of AI technology and the tight coupling of an extender can make interactions less predictable. We will focus on five areas of concern:

1. The first consideration is *autonomy*. In medical ethics, the principle of autonomy includes respect for both an individual's right to decide and for the freedom of whether to decide.[43] One risk is that, for the sake of having the patient under control, some AI extenders will make use of interventions and nudges that effectively bypass the agent's right to decide. By encouraging actions without appealing to the agent's rationality (e.g. by presenting them with reasons to act), these devices could risk becoming *manipulative*. These scenarios become particularly concerning when we consider the technology to be a genuine part of the person's mind—the innermost space of private information, where one's intentions are formed and decisions are made.[44] As such, any manipulative interventions would clearly deviate from the maxim of non-maleficence. Indeed, in the worst case, some of these systems could be hacked and used with malicious purposes.

2. In another important sense, autonomy should also protect one's ability to safely act in the ways one decides, ensuring short-term and long-term *reliability*. We can imagine cases of over-reliance in which a person is put in risky situations (in terms of mental health), by becoming overly dependent on an AI extender which is liable to unexpectedly fail, as any technology can. This goes beyond the classical problems of cognitive laziness and atrophy that may be caused by the use of AI extenders.[45,46] A somewhat related concern is a scenario in which patients feel so integrated with the extender that they resist changes to the system, as these would imply a change of personality and cognitive capabilities.

3. A third problem derives from an *unregulated or recreational use* of these AI extenders for mental health, where the appropriate validation and certification of procedures and tools do not follow the standards of medical practice with some other treatments, putting beneficence (good practice) at risk. This is particularly worrisome when mentally healthy people experiment with AI extenders, leading to some pathological mental situations (e.g., similar to either substance abuse or dependence syndrome), but with some technological and AI components that could be new to the analysis. This is also related to the above points on autonomy, as an overreliance would negatively affect one's autonomy.

4. A fourth problem regards *moral status and privacy*. This goes beyond the risk that an extender may be stolen or accessed by a third party (or by the clinician or family beyond some established parameters)—a risk that applies to essentially any medical device. Under the EMT, and AI extender really is a part of the person's mind, and hence gaining access to the personal information stored in a device would be like reading the brain of a person, especially as these extenders may contain memories, experiences, decisions and other very sensitive information.[47,48] This is the classic double-edged sword in AI: while collecting more information about the individual can fuel powerful and highly personalised predictions (a benefit we discuss above), it also threatens personal privacy.

5. Finally, there may be problems with their *allowance in the public space* caused by a misunderstanding (or strong disagreement) of the EMT. This may lead to limitations on when and where these devices are allowed (exams, recruitment, etc.), and for how long they can be removed (airport security, other hospital treatments, etc.). This is of course related to the medical ethical principle of justice, and to the question (discussed above in 'benefits') of whether we should consider AI extenders as cures.

There is a broader concern worth being mentioned, which is common to many kinds of enhancement. A widespread use of AI extenders can change

our conception of what humans are capable of, and in the particular case of mental health, our notion of "mental normality". As more and more capabilities can be enhanced or modified with these devices, the diversity of behaviors and capabilities may change as people can increasingly choose what cognitive profile they prefer for themselves. The principle of justice also demands that we consider future scenarios that could arise for society—such as a moment when everybody has access to enhancements. Determining what profiles are safe for the person (typically in the long term) and for society is going to require a deeper understanding of what mental health is, to what degree mental conditions are pathological, and what enhancements people should be allowed to make. Such considerations are well beyond the scope of this chapter, but what is clear is that the notion of a "standard" or "normal person", only comprising what the brain can do, if it ever made sense, will likely have to be completely discarded, especially when associated with a goal of being "cured".

7 Recommendations

In the previous sections, we have argued that a widespread use of AI extenders, and their understanding as such, may have important implications in the analysis and practice of mental health. For instance, the attachment of a patient with their AI extender can be so close, that any change on the device or its software may require a deeper consideration for which the professionals involved may not be used to yet. It is then these professionals—the designers of AI extenders, coming from different areas of engineering and especially AI, and the clinicians, from physicians to nurses and other careers—who need a re-understanding of what these AI extenders mean for the evolution of the mental health and all the possible side-effects on a patient.

The most urgent recommendations should be addressed to AI designers. The regulations and expectations that are put on an app or another kind of "software" or "hardware" extender should be no less stringent than those put on drugs or other kinds of treatment. The reference to take here is similar to the area of orthopedics, where manufacturers must include diverse research

and development teams, including clinicians, and perform careful tests. Likewise for any digital monitoring app, development teams must determine an ethically acceptable way of designing these systems so that we can avoid these potentially negative effects.[40] But AI extenders must be more reliable than physical orthopedics. If Helen or Lewis's AI extenders fail, the consequences may be serious and even dangerous; hence, designing for safety and reliability is essential. But, on top of this, from the point of view of cognitive extensions, the manufacturer must understand that the software and the hardware become part of the mind, so no updates, discontinuations or access to the data can be done without informed consent. Under a strict interpretation of the EM thesis, modifying an AI extender should be compared to modifying the brain.

Clinicians, too, must be aware that new gadgets imbued with obscure AI are going to become a regular part of their repertoire of diagnostic, treatment and monitoring tools. They need to understand their basics, and how they couple with the human mind in order to create some new behaviors unseen in their careers. A good starting point for training and information for clinicians could be based on the six issues raised by Bauer et al.: (1) decide when to recommend an extender, (2) observe what other extenders the patients use (and consider how different extenders might potentially interact), (3) understand how their monitoring works, (4) explain the effects to the patients, (5) keep themselves informed about the state of the art of AI extenders, and (6) scrutinize and validate them.[49] With the inception of technology, and especially AI, human minds are changing, and mental health must change too, in terms of categories and the consideration of normality. Even if clinicians are not familiar with the philosophical underpinnings of the EMT, they know well what orthopedics is, and understand the feeling of many patients that an artificial arm, say, is a real arm. A similar analogy can be used for AI extenders, but going beyond the idea of mimicking the original functions exactly, in the same way that a titanium leg may be more effective and elegant than a more realistic plastic prosthetic.

Finally, there are many future directions for research for a better understanding of AI extenders in the context of mental health, for which this

chapter is just a beginning. Tab. 2, and future refinements, can be used to spot gaps and limitations, or ways in which some devices can be used for some other conditions. But beyond each particular set of capabilities and conditions, we need more general guidelines, methodologies and well-designed experiments to help in the development of the future AI extenders used for mental health. The EMT can leverage this research, but we also need better structural incentives to create intelligent assistive health technologies, rather than focusing only on biological causes and cures.

Acknowledgements

We thank Richard Heersmink and Jacopo Domenicucci for their valuable comments and suggestions. JHO was supported by the EU (FEDER), and the Spanish MINECO under grant RTI2018- 094403-B-C32, the Generalitat Valenciana PROMETEO/2019/098, and the Future of Life Institute under grant RFP2-152. KV was supported by the Leverhulme Centre for the Future of Intelligence, Leverhulme Trust, under Grant RC-2015-067.

References

1. Clark A, Chalmers, D. The Extended Mind. Analysis. 1998;58:7-19.
2. Hernández-Orallo J, Vold K. AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI. AAAI /ACM Conference on Artificial Intelligence, Ethics, and Society (AIES 2018), Honolulu, Hawaii, USA. January 27-28, 2019. p. 507-513.
3. World Health Organization. ICD-10: international statistical classification of diseases and related health problems. 2nd ed. World Health Organization; 2004.
4. Hurley SL. Vehicles, Contents, Conceptual Structure and Externalism. Analysis. 1998;58:1-6.

5. Wheeler M. Reconstructing the cognitive world. Cambridge, MA: MIT Press; 2005.

6. Rupert RD. Challenges to the Hypothesis of Extended Cognition. Journal of Philosophy. 2004;101(8):389-428.

7. Sprevak M. Extended Cognition and Functionalism. Journal of Philosophy. 2009;106:503-27.

8. Sutton J. Exograms and interdisciplinarity: History, the extended mind, and the civilizing process. In Menary R, editor. The extended mind. Cambridge, MA: MIT Press; 2010. p. 189-225.

9. Menary R. Cognitive integration and the extended mind. In: Menary R, editor. The extended mind. Cambridge, MA: MIT Press; 2010. pp. 267-88. https://doi.org/10.7551/mitpress/9780262014038.003.0010

10. Rowlands M. The New Science of the Mind. Cambridge, M.A.: MIT Press; 2010.

11. Heersmink R. Dimensions of integration in embedded and extended cognitive systems. Phenomenology and the Cognitive Sciences. 2015;14:577-598. https://doi.org/10.1007/s11097-014-9355-1

12. Wegner DM. Transactive memory: A contemporary analysis of group mind. In: Mullen B, Goethals GR, editors. Theories of group behavior. New York: Springer-Verlag; 1987. p. 185-208.

13. Wegner D, Raymond P, Erber R. Transactive memory in Close Relationships. Journal of Personality and Social Psychology. 1991;61(6):923-929.

14. Sutton J, Harris CB, Keil PG, Barnier A. The Psychology of memory, extended cognition, and socially distributed remembering. Phenomenology of Cognitive Sciences. 2010;9(4):521-60. https://doi.org/10.1007/s11097-010-9182-y

15. Chalmers D. Foreword to Andy Clark's Supersizing the Mind: Embodiment, Action, and Cognitive Extension. Oxford: Oxford University Press; 2008. p. ix-xvi.

16. Hutchins E. Cognitive artifacts. In: Wilson RA, Keil FC, editors. The MIT Encyclopedia of the Cognitive Sciences (MITECS). New ed. Cambridge, MA: MIT Press; 1999. p. 126-7.

17. Spohrer J, Banavar G. Cognition as a service: an industry perspective. AI Magazine. 2015;36(4):71-86.

18. Cole E. Cognitive prosthetics: an overview to a method of treatment. NeuroRehabilitation. 1999;12:39-51.

19. Kolodner JL. Cognitive Prosthetics for Fostering Learning: A View from the Learning Sciences. AI Magazine. 2015;36(4):34-50.

20. Derian M. Cognitive Prosthetics. Oxford: Elsevier; 2019.

21. Mihailidis A, Fernie GR, Barbenel. JC. The use of artificial intelligence in the design of an intelligent cognitive orthosis for people with dementia. Assistive Technology. 2001;13(1):23-39.

22. Simpson RC, LoPresti EF, Schreckenghost D, Kirsch N, Hayashi S. Solo: a cognitive orthosis. AAAI Spring Symposium: Persistent Assistants: Living and Working with AI. 2005.

23. Fulmer R. Artificial intelligence and counseling: Four levels of implementation. Theory & Psychology. 2019. https://doi.org/10.1177/0959354319853045.

24. Weizenbaum J. ELIZA---a computer program for the study of natural language communication between man and machine. Communications of the ACM. 1966;9(1):36-45.

25. Luxton DD, editor. Artificial Intelligence in Behavioral and Mental Health Care. 1st ed. Academic Press; 2015.

26. Juan MC, Alcaniz M, Monserrat C, Botella C, Baños RM, Guerrero B. Using augmented reality to treat phobias. IEEE computer graphics and applications. 2005;25(6):31-37.

27. del Valle ACA, Opalach A. The Persuasive Mirror: computerized persuasion for healthy living. Proceedings of the 11th International Conference on Human-Computer Interaction; 2005.

28. Levy N. Neuroethics: Challenges for the 21st Century. Cambridge University Press, 2007.

29. Bray A. The Extended Mind and Borderline Personality Disorder. Australasian Psychiatry. 2008;16:8-12.

30. Vold K. Overcoming deadlock: Scientific and ethical reasons to embrace the extended mind thesis. Philosophy and Society. 2018;29(4):489-504.

31. King C. Learning Disability and the Extended Mind. Essays in Philosophy. 2016;17(2):38-68.

32. Heersmink R, Knight S. Distributed learning: Educating and assessing extended cognitive systems. Philosophical Psychology. 2017;31(6):969-990.

33. Drayson Z, Clark A. Cognitive disability and embodied, extended minds. In: Wasserman D, Cureton A, editors. Oxford Handbook of Philosophy and Disability. Oxford: Oxford University Press; 2019.

34. Garner JB, Campbell PH. Technology For Persons with Severe Disabilities: Practical and Ethical Considerations. The Journal of Special Education. 1987;21(3):122-32.

35. Kirsch NL, Levine SP, Fallon-Krueger M, Jaros LA. Focus on clinical research: The microcomputer as an "orthotic" device for patients with cognitive deficits. The Journal of Head Trauma Rehabilitation. 1987;2(4):77-86.

36. Calvo R, Peters, D. Positive Computing: Technology for Wellbeing and Human Potential. MIT Press; 2014.

37. Krueger J, Maiese M. Mental institutions, habits of mind, and an extended approach to autism. Thaumàzein. 2018;6:10-41.

38. Cooper C, Goswami U, Sahakian BJ. Mental capital and wellbeing. Wiley-Blackwell; 2009.

39. Shergill SS, Barker D, Greenberg M. Communication of psychiatric diagnosis. Social psychiatry and psychiatric epidemiology. 1997;33(1):32-8.

40. Ienca M, Wangmo T, Jotterand F, Kressig, RW, Elger, B. Ethical Design of Intelligent Assistive Technologies for Dementia. Science and Engineering Ethics. 2018;24(4):1035-55. https://doi.org/10.1007/s11948-017-9976-1

41. Beauchamp TL, Childress JF. Principles of biomedical ethics. 7th ed. New York: Oxford University Press; 2013.

42. Gillon R. Medical ethics: four principles plus attention to scope. British Medical Journal. 1994;309:184-8. https://doi.org/10.1136/bmj.309.6948.184.

43. Burr C, Morley J. Empowerment or Engagement? Digital Health Technologies for Mental Healthcare. 2019. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3393534 Accessed 5 Sep 2019.

44. Reiner P, Nagel S. Technologies of the extended mind: defining the issues. In Illes J, Hossain S, editors. Neuroethics: Anticipating the future. Oxford: Oxford University Press; 2017. p. 108-122.

45. Carr N. The glass cage: Where automation is taking us. London: Random House; 2015.

46. Barr N, Pennycook G, Stolz JA, Fugelsang, JA. The brain in your pocket: Evidence that smartphones are used to supplant thinking. Computers in Human Behavior. 2015;48:473-80.

47. Blitz MJ. Freedom of Thought for the Extended Mind: Cognitive Enhancement and the Constitution. Wisconsin Law Review. 2010;4: 1049-1119.

48. Søraker J. The moral status of information and information technology: A relational theory of moral status. In Hongladarom S, Ess C, editors. Information technology ethics: Cultural perspectives. Hershey: Idea Group Publishing; 2007. p. 1-19.

49. Bauer M, Glenn T, Monteith S, Bauer R, Whybrow PC, Geddes J. Ethical perspectives on recommending digital technology for patients with mental illness. International Journal of Bipolar Disorders. 2017;5:1-14.

APPENDIX A

The following table includes the search tokens that we used to look for the conditions in the ICD-10 blue book for each capability, in order to construct Table 2 in the chapter.

Cognitive capability	**Tokens used for the search in the ICD-10**
Memory processes	*memor**
Sensorimotor interaction	*sensor*, moto**
Visual processing	*visu*, percept**
Auditory processing	*audi*, percept**
Attention and search	*atten**
Planning	*plan*, organiz**
Comprehension and expression	*expres*, compre**
Communication	*commun*, lang**
Emotion and self-control	*emot*, control* and affect**
Navigation	*orient*, naviga**
Conceptualization, learning and abstraction	*learn*, conceptual**
Quantitative and logical reasoning	*calculat*, mathemat**
Mind modeling and social interaction	*social**
Metacognition	*aware*, capab*, limitations, conscio*, self-*, incompetent*