AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI

José Hernández-Orallo*
DSIC
Universitat Politècnica de València
Spain
jorallo@dsic.upv.es

Karina Vold
Leverhulme Centre for the Future of Intelligence,
University of Cambridge
UK
kvv22@cam.ac.uk

ABSTRACT

Humans and AI systems are usually portrayed as separate systems that we need to align in values and goals. However, there is a great deal of AI technology found in non-autonomous systems that are used as *cognitive tools* by humans. Under the *extended mind thesis*, the functional contributions of these tools become as essential to our cognition as our brains. But AI can take cognitive extension towards totally new capabilities, posing new philosophical, ethical and technical challenges. To analyse these challenges better, we define and place *AI extenders* in a continuum between fully-externalized systems, loosely coupled with humans, and fully-internalized processes, with operations ultimately performed by the brain, making the tool redundant. We dissect the landscape of cognitive capabilities that can foreseeably be extended by AI and examine their ethical implications. We suggest that cognitive extenders using AI be treated as distinct from other cognitive enhancers by all relevant stakeholders, including developers, policy makers, and human users.

CCS CONCEPTS

• **Social and professional topics** → **Codes of ethics**; *Computing / technology policy*; • **Computing methodologies** → **Artificial intelligence**; **Theory of mind**; *Cognitive science.*

KEYWORDS

Extended mind; AI extenders; cognitive assistants; ethics of AI; societal impact of AI; cognitive augmentation

ACM Reference Format:

José Hernández-Orallo and Karina Vold. 2019. AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), January 27–28, 2019, Honolulu, HI, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3306618.3314238

1 INTRODUCTION

Many societal and ethical issues about AI are seen under the perspective of autonomous systems that can replace humans, such as a self-driving car, a robotic delivery system or a fully-fledged diagnosis system. While this is often a useful perspective, other AI systems are more interactive and coupled, such as a language translator assistant that relies on human interaction. These systems are not autonomous in the traditional sense as they do not perform tasks on their own. Instead their purpose is to help humans complete tasks. Some would say that these systems are work-in-progress AI, because the task has just been semi-automated or assisted. But many sophisticated AI techniques, including machine learning techniques, can just as well be applied in interactive and coupled systems. One very interesting feature of these interactions is the way the user changes their reasoning processes: it is not that part of the process has been replaced; rather the whole task has been redesigned, and the skills of the human user often co-evolve with the technology.

Non-autonomous AI systems offer a repertoire of services and tasks that will become more abstract and general in the future, and carry the potential to offer humans entirely new cognitive abilities: they can be a *cognitive tool*. This is aligned with –but goes beyond– what is sometimes called "cognition as a service" [34] or the rise of "cognitive assistants" [23]. The key difference is that cognitive tools are more tightly coupled with our biological cognitive system and, at the same time, more ubiquitous, such that viewing them as an *external* service or an application no longer makes sense, especially when we have to account for how these technologies are perceived and assimilated by humans, and how their impact and ethical consequences are assessed.

In contemporary philosophy there are views that can shed light on this phenomenon. The extended cognition thesis is a popular emerging view in the fields of philosophy of mind and cognition that claims that tools in an agent's environment (i.e., beyond their biological organism) can serve as partially constitutive 'extensions' of their cognitive states and processes. The view thereby rejects the longstanding tradition of understanding the mind as nothing more than the operations of the brain. Perhaps the most influential argument for the extended cognition thesis comes from Clark and Chalmers [10], but in the last two decades many other arguments have been proffered, and the view has become a widely debated topic. In general the thesis maintains that the tools we use to help us complete cognitive tasks can become seamlessly integrated into our biological capacities, being on a par with our brains in so far as both play an indispensable functional role in bringing about our cognitive abilities. A standard example is the indispensable role that a pen and paper play for a mathematician in solving complex equations. These 'cognitive tools' are more than just *tools*, they are

*Also at the Leverhulme Centre for the Future of Intelligence, UK.

AIES '19, January 27–28, 2019, Honolulu, HI, USA

ACM ISBN 978-1-4503-6324-2/19/01.
https://doi.org/10.1145/3306618.3314238

incorporated as part of the mind (it likewise would not make sense to call the brain a 'tool' for the mind).

In this paper we explore what new abilities could be extended in the future using AI, and what their ethical and societal implications might be. Within the philosophical literature on extended cognition, these possibilities remain entirely unexplored. Meanwhile the development of non-autonomous AI systems remains second-class to the 'genuine' AI that is explicitly devoted to the development of autonomous agency. Hence in this paper, we aim to tie these two fields together, to the mutual benefit of both disciplines.

These different objectives will likely translate to different conceptions of AI, and its role for the future. While early computers were aimed at automating tedious number-crunching tasks, a different vision emerged of the computer as a real-time interactive system that could support and expand human capacities [13, 20]. Now a similar dilemma emerges between AI systems that are autonomous and can replace humans in certain tasks and those that are designed to augment human intelligence in different ways [29]. Instead of AI systems functioning as autonomous agents, one can conceive the creation of "intelligent" or "cognitive extenders". Indeed, systems can vary quite significantly in terms of their autonomy and degree of coupling with humans. This space of possibilities merits consideration. In the end, *cognitive extenders* could be a standalone paradigm inside AI:

(1) because of the tight coupling of cognitive extension, an accurate modeling of the user is needed to know what the best way is of compensating or enhancing some skills,
(2) with the introduction of new capabilities, the resulting extended humans can behave in less predictable ways, with unprecedented cognitive power and personalities,
(3) as these cognitive extenders can be poorly designed, can fail or can change, the implications are more direct in terms of undesirable effects such as miserliness (cognitive idleness), dependency and atrophy; and
(4) due to an integration where humans are able to remain in control, the responsibility for the AI developer may be diluted. If the humans are just seen as extended or enhanced by the tool, they would plausibly remain responsible.

The impact of cognitive extenders is sufficiently different from autonomous AI agents that the ethical and social implications need to be analyzed separately. This paper includes a series of contributions. We look at AI under the perspective of extended cognition and place it in a spectrum of present and future AI systems ranging from autonomous systems that substitute for human cognition to systems that can help humans internalize new concepts and ideas. We enumerate a range of possible cognitive capabilities that can extend human cognition in the future, and what their impact might be. From here, we analyze their ethical and societal consequences through the specifics of cognitive extension. Finally, we address AI researchers and developers, psychologists and philosophers, regulators and policy-makers with some recommendations about cognitive extenders, such as how they should be built in the first place, offered to users, monitored in a non-intrusive way, and able to be removed (or 'decoupled') without disabling the user.

2 COGNITIVE EXTENSION

Many philosophers now argue that our use of technology has enabled us to expand beyond our biologically-bound cognitive capacities, making us smarter and more capable agents. This is the thesis of extended cognition, popularized by Clark and Chalmers [10]. The view pushes back against, and in some cases outright rejects, some of the core commitments of traditional cognitive science and the field of AI [2, 36, 37].

Most importantly, extended cognition theorists universally reject intracranialism, the view that all cognition is instantiated in the brain. They instead maintain that cognitive states and processes can be partially constituted by mechanisms beyond one's brain, that is, the vehicles of our cognitive representations need not be instantiated by sets of neurons in the brain. Their argument is typically motivated by a version of the computational theory of mind, a core commitment of cognitive science, which (put crudely) maintains that the mind is the 'software' that runs on the 'hardware' of the brain. This view allows for the possibility that the same mental type could be 'multiply realized' or instantiated (just as a Turing machine could) by heterogeneous physical types [25, 26].

Clark and Chalmers appealed to this view to argue that cognition can sometimes be partially instantiated by extra-bodily, non-biological elements, so long as those pieces of 'hardware' are able to instantiate the right kinds of computations. Accordingly, their argument for the extended cognition thesis has been dubbed the 'parity argument', as it relies on the claim that we should treat computationally equivalent processes with "the parity they deserve", irrespective of whether they are internal or external to the skull [10, p.8]. They then further argue that, while human minds might require brains, cognition can be partially instantiated by external objects, such as our notebooks and smartphones. It is key that the external resource be appropriately integrated; it must be a constant in one's life, highly accessible, and reliable (as the brain typically is). It is important to emphasize the strength of the extended cognition thesis; it does not merely claim that cognition causally depends on wider processes in the environment, rather it claims that these wider processes can *constitute* cognition. Hence, they can be on a par with the brain.

While the parity argument helped popularize the extended cognition thesis in contemporary philosophy of cognition, it is not without limitations. One important consideration for our discussion is whether cognitive extenders can give humans novel cognitive capacities or merely be cognitive prosthetics. One of the limitations with the parity argument is the extent to which it relies on an appeal to functional similarities between inner and outer resources. In cases where the outer resource is completely novel and enables cognitive capacities that are beyond anything that could be done internally (i.e., by the brain), the argument fails. In an attempt to overcome this limitation, as well as others, several philosophers have argued for the extended cognition view in distinct ways (e.g., [12, 21, 30, 35]).

In what follows, we will adopt the view that cognitive extenders can not only replace or substitute for functions of the brain, they can also move beyond the brain to enhance our biological functions. We will also adopt and adapt some definitions in the literature. For instance, the physical objects that we use to extend our cognition

are typically designed for the purpose of helping us complete some cognitive task. Hutchins calls these 'cognitive artifacts', which he characterizes as "physical objects made by humans for the purpose of aiding, enhancing, or improving cognition" [18, p.199]. In general, if we include non-artifacts, i.e., natural objects (e.g., the sun as an orientation extender), we can use the term 'cognitive extender'.

The previous definition also blurs the distinction between enhancement and extension, but we need to make the differences explicit. For instance, cognitive enhancement through drugs or other kinds of neural interventions do not count as cognitive extenders. What is important for cognitive extension is that the vehicles of our mental representations are located outside of the head. Whereas nootropics might influence and enhance brain activity, they do not challenge intracranialism. Similarly, the assimilation of a rule of thumb, a new word and any other case of learning or psychological *internalization* [40], enhances cognition, but it is not an extension. On the other extreme, a complete *externalization* (or "cognitive outsourcing"), where the process is delegated to another person or system in a batched, decoupled fashion (e.g., translate a novel between two languages) does not count as extension either[1].

In order to make the notion of cognitive extension more precise, we slightly refine Hutchins's definition as follows:

> *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.*

In the definition, it is important to emphasize that the effect is lost when the element is removed. We thereby understand terms such as "cognitive atrophy" and "cognitive prosthetics", emphasizing that the effect ceases when the extender is removed. In the case of atrophy, more than the given positive effect is lost. Furthermore, a cognitive extender can be virtual –this is to include software and augmented reality– and does not need to be an object in any strict sense.

Let us introduce further terminology and some notation. The system or human A that is extended can be referred to as the *extendee*, and the result will be denoted by $A[E]$, where E is the *extender*. It is only through a very coupled interaction between A and E where we can actually talk about A being extended, and consider E as part of its cognitive resources. Of course the lines are blurry sometimes, in the same way extended cognition has been observed in social contexts with relatives (parent-child), highly interdependent couples or very close friends, where A (e.g., child) is actually extended by E (e.g., parent). But note that in the context of extension we are not interested in the collective capabilities or the social aspect (the collective, denoted by $A + E$), but the way A operates as an individual, extended by E, i.e., $A[E]$. Occasionally, we will use the notation $A \leftarrow E$ for A being replaced (or overridden) by E and A^E to illustrate when A has internalized E, so E is no longer necessary.

Note that the definition above is not anthropomorphic. It could be applied to non-human animals and to AI systems, which could be enhanced by other objects, AI systems, and humans, through 'human computation' [39] or 'human-extended machine cognition' [32].

[1] "Cognitive offloading" [27] is a related term usually referring to particular ways of aiding and/or improving cognition by gestures or manipulation. We will avoid the term here, as it is not always clear if some examples of cognitive offloading are really extenders or are just metacognition strategies, e.g., tilting one's head to read a rotated text.

3 AI: EXTERNALIZE, INTERNALIZE OR EXTEND?

AI can be used for cognitive externalization (i.e., outsourcing), for cognitive internalization and for cognitive extension. Let us examine each case in detail.

The traditional view of AI is typically externalization. Minsky defined AI as the "science of making machines capable of performing tasks that would require intelligence if done by [humans]" [22]. An AI system should solve tasks independently, with limited or no human assistance or manipulation. The term "autonomous agent" was introduced to represent this goal of AI, and the perspective from outside the field often reinforces this view, by use of the related concept of automation. Whenever humans are still needed it is because AI is not capable enough or because humans must control or supervise what machines are doing. Humans and machines can even be *synergetic*, where the whole is more than the sum of its parts. However, even in this $A + E$ situation, AI is not designed to induce a change in the way the human individual (user) performs cognition.

The narrative of externalization, associated with a view of replacement $A \leftarrow E$, especially in the workplace, becomes more complex as machines are able to do some tasks better than humans, e.g., memory and calculation. Machines perform computations faster than humans, deal with larger amounts of data, and so forth. However, more recently we see machines doing kinds of cognition that look very different from the way humans think, exploiting, e.g., the divergence between biological and artificial neural networks.

On the other hand, internalization in the AI domain implies the acquisition of processes that are observed on a machine. For instance, a human can see how an AI system solves a problem and internalize the procedure. That does not mean that the machine is necessarily redundant, but that the human can reproduce (at least approximately) what the machine is doing. The potential of internalization with AI, and generative models in machine learning, has recently been explored by Carter and Nielssen [7]: "Rather than outsourcing cognition, it's about changing the operations and representations we use to think; it's about changing the substrate of thought itself. And so while cognitive outsourcing is important, this cognitive transformation view offers a much more profound model of intelligence augmentation. It's a view in which computers are a means to change and expand human thought itself". Under this view, AI becomes the creator of new concepts and representations, which we can then use. AI becomes a teacher or discoverer, a contributor to the conceptual baggage of human culture.

Internalization seems more empowering than other ways of augmenting cognition. However, as AI becomes more powerful, humans may not be able to internalize many of the concepts created by AI, because of their different capacities and representations, even with huge progress in the area of explainable AI [31].

Finally, cognitive extension, as defined in the previous section, is not fully externalized, as the tight coupling remains, and not fully internalized, as the cognitive extender is needed for the functionality. For AI, the design of a system E to work as $A[E]$ is different

from a whole autonomous system E, but also from $A + E$ or an internalized A^E, *after* interacting and learning from E. The flexibility for extension is much higher, as only the interface needs to be internalized (e.g., in order to use a calculator, we only need to internalize the use of the buttons), but many other things do not need to be understood by the user, in the same way one can drive a car without knowing all its mechanics. Cognitive extenders fuelled with AI (henceforth 'AI extenders') bridge the area of human-computer interaction with AI.

This perspective puts the emphasis on an AI that is more human-centered, but less human-like. If AI systems were just designed to mimic or replace human behavior or being internalized by them, the possibilities of cognitive extension would be limited to cognitive prosthetics, applicable when the effects of pathologies or aging require to recover the 'standard human cognition'.

As an example of how the same functionality can be externalized, internalized or extended – and the sometimes blurry lines between them – let us consider translation. A batch machine translator that takes a text and converts it into another language is an externalized translation service. It is difficult to go beyond the capabilities that the translator provides (translation tasks) if just used occasionally and in a batch mode. A human can look at the result and learn elegant translation transformations. In this way, the human internalizes new methods of translation. However, with a more interactive version and a regular use, the translation system starts being used in a different way (delegating the easy cases, and reserving those that the person deems more difficult for the machine). As a result the translation quality of the coupled system $A[E]$ can increase significantly, as the user can understand where E fails, or correct some translations that do not make sense. If a human A knows a little bit about the languages to be translated, this assisted translation using E will become much better than any of A or E could do independently. In this latter case, the human is using the extender, controlling and integrating it for the solution of the task.

Note that the lack of autonomy of E is crucial to see this as an extension rather than a collaboration. This detachment between cognition (or even intelligence) and autonomy is well-aligned with the view of cognition as a service [34], where several facilities for visual perception, speech and language processing are provided, as well as other inference and reasoning solutions, independent of any task. For instance, an online translation system can be provided as a service, which can be integrated into many kinds of applications and goals. Whether it is fully externalized or seen as an extension will depend on the degree of coupling, integration, and interaction.

4 WAYS AI CAN EXTEND COGNITION

The crucial aspect of AI extenders is that they themselves implement kinds of cognitive processes; they are not mere static or interactive tools. This makes cognitive extension far more powerful and complex than they were when the extended cognition thesis was introduced – at that time the standard example was a notebook. Before making an analysis of the future implications of AI extenders, we need to have a better understanding of the kinds of extensions that are envisaged by current and future AI.

Not only must we understand the different areas of cognition, but we have to realise that cognitive extenders are designed to be tightly coupled. With AI extenders, machine learning can be used to model human cognition, spot our cognitive limitations, and exploit our capabilities in full. As a result, one new trait of the next generation of cognitive extenders is that they will model the user and change their behavior by learning from the extendee, thereby allowing them to personalize the best cognitive aid depending on the situation. For instance, an AI extender can learn that a person usually utters a particular word or intonation before introducing a compelling argument. The system can suggest these words when writing or speaking in order to produce more of these arguments. In the end, future AI extenders can become cognitive coaches or therapists, if properly devised to do so. However, for reasons we will discuss in the next section (e.g. users' miserliness and companies' profit), AI extenders may be designed in such a way that extensions increase faster than the cognitive augmentation can be gained, such that the user loses control of the symbiotic situation.

In order to see the possibilities of AI extenders, we could look at recent breakthroughs produced by machine learning in areas such as voice recognition, emotion understanding, social interaction, game playing, etc., but that might give us a shortsighted view of what is coming ahead. A richer approach can be based on analyzing all areas in which cognition can be extended. Of course, no standard catalogue of cognitive abilities exists, but the hierarchical theories of intelligence in psychology, animal cognition and the textbooks in AI usually agree (at least partially) on the following abilities, or at least, in this way of organizing the vast space of cognition. For our purposes, we integrate from several sources[2] and analyze AI extenders under these categories:

- Memory processes: cognitive extension happens when writing is introduced, and our memories (individually and collectively) were enhanced. In the future, we will surely see more of the Google effect [33] and cognitive assistants reminding us of an event or appointment. But AI extenders can also introduce new customized mnemonics to improve long-term memory, or tag our experiences with related people, concepts and other situations to improve episodic memory.
- Sensorimotor interaction: AI extenders can perceive, recognize and manipulate patterns in different ways, not only through new sensory and actuator modalities but in terms of mixing representations through generative models [7]. This means that new sensors can be intelligently translated into ways that are properly understood by our senses, and new actuators can be integrated as interactive possibilities.
- Visual processing: many of the most striking new applications of machine learning have been around ways in which images and videos can be processed and generated. Coupled with augmented reality and other ways of transforming the input through intelligent filters, the possibilities of seeing things we cannot usually see –and in different ways– are endless.
- Auditory processing: this includes systems that highlight those parts of the speech that might be missed by the user,

[2] We take a mixture from figures 3.1 and 3.2 (human psychometrics, from Thurstone's primary mental abilities and Cattell-Horn-Carroll hierarchical model), table 4.1 (animal cognition research, from Wasserman and Zentall's book) and tables 5.2, 5.3 and figure 5.3 (AI, AGI and benchmarks, from AI Journal and Adams et al.), all sources found in [17].

following different conversations and prompting the user when an interesting topic is raised by any of them. AI extenders can do this in more powerful (and less stressful) ways than we are used to.
- Attention and search: examples such as "the invisible Gorilla" [9] show how easily humans overlook things. This can be seen as a feature rather than a bug, but success is sometimes only known in hindsight. If an AI extender models our interests or goals, however, it can search through information or focus our attention on things that we would otherwise overlook.
- Planning, decision-making and acting: agendas and other daily tasks will be planned by our cognitive assistants, rather than just being passive systems where we write things up that are reminded later. This will entail better use of our time, including what times of the day we are ready for different kinds of cognitive tasks.
- Comprehension and expression: AI extenders will help us with understanding information. This will go beyond floating annotations, including rewriting or re-rendering to improve interpretability. This can also be applied to reading, watching films, listening music, other arts, etc.
- Communication: AI systems could reply to our emails and write our tweets, but they could also be used in more executive ways, such as telling an assistant to communicate X to Y (e.g., that Y is fired) in the best possible way.
- Emotion and self-control: with AI becoming better than humans at what is called 'emotional intelligence', we could have systems that will inform us of our emotions and those of others, detect when emotions are fake and help us trigger the right emotional reactions.
- Navigation: this goes much beyond the use of GPS devices, to include the association of places and routes with cognitive processes, including memories, people, images, etc., with a new sense of location and time, related to the new kinds of episodic memory mentioned above.
- Conceptualization, learning and abstraction: machine learning and knowledge representation can find categories that humans could have never found. These abstractions can be inferred from data or knowledge and explained to us so that we operationalize (but not fully internalize) the new concept as part of our reasoning processes.
- Quantitative and logical reasoning: there is already much potential in ways uncertainty can be processed by cognitive extenders in terms of probabilities and frequencies. For instance, many systems monitoring us can report probabilities of events (e.g., risk of accidents) or quantities (people in a room) in real time.
- Mind modeling and social interaction: AI extenders can model the extendee's network of contacts, and anticipate decisions, actions and interests of other people. In the end, they can report the BDI (beliefs, desires and intentions) of others (enhancing social capabilities).
- Metacognition: the metacognition of A evolves into the metacognition of $A[E]$. AI extenders can identify the potential and limitations of E, who will be more aware of them, using the capabilities and strategies of both A and E more optimally.

The previous characterization in terms of breadth (range of abilities that can be extended) and depth (how the extender adapts to and intervenes on the user) gives us a more grounded position to understand the impact of AI extenders.

5 IMPLICATIONS

The list in the previous section looks highly empowering, giving a generally positive view that is shared with some other takes on human augmentation [4]. Other sources, however, focus directly on the negative effects [5, 6, 11, 15]. Some of these concerns extend over the range between externalized and internalized processes, but do not properly situate the analysis under the extended mind thesis [38]. For instance, Danaher [11] revisits the issue of social interaction as deceptive if done by AI tools, but this is seen differently if we consider that the 'persona' is *actually* the whole $A[E]$ and not an "outsourced" E "on A's behalf". Surveying previous work from a full extended mind perspective, we classify the ethical issues into five groups:

- Atrophy and safety: The first main issue about extension that differs from augmentation through externalization or internalization –but is shared with augmentation through drugs (nootropics)– is that *all –or more than– its positive effect is lost when the element is not present*, as for our definition of cognitive extender. There seems to be nothing wrong about making humans' life easier, but taking into account the miserliness of human cognition [14], as AI provides more cognitive possibilities, the risk of humans being dumbed down increases (Carr's "degeneration" effect, [5]. Recent research [1] shows that individuals have different predispositions to cognitive miserliness and delegation on devices such as smartphones. Cognitive atrophy also generates many safety issues, as people may become vulnerable when the extender is removed (or if it suddenly fails to work). It is not E, $A[E]$, $A \leftarrow E$, A^E or $A + E$ that become unsafe, but $A[\emptyset]$.
- Moral status and personal identity: cognitive extension is particularly sensitive since the "degree of dependency and integration [is] proportional to the artifact's moral status" [16]. This means that regulation should go beyond ownership and compensation for damage [6]: our extended minds should not be interfered with without permission, or removed if the user can no longer afford it. Similarly, privacy should reach all the elements that play a role in an extended mind, including intellectual property, as whatever $A[E]$ creates should be owned by A and not shared, as in an $A + E$ situation. Frischmann and Selinger [15],for example, warn that we should be concerned about external parties coming to own E, and what is created by E, rather than A.
- Responsibility and trust: it is not always clear who is responsible when autonomous AI systems fail or behave in an unfair way, but AI extenders can well take the perspective of software licenses, where the manufacturer usually puts responsibility on the user, with trust being compromised. If this is the case, companies will even be encouraged to use extended humans $A[E]$ when autonomous systems on their

own (E) are forbidden (e.g., for some jobs) or a human-in-the-loop is needed to supervise or correct E. For instance, who is responsible when a doctor extended by a diagnosis system makes a mistake compared to a fully automated diagnostic system?
- Interference and control: many of the issues about the impact of AI on independence and manipulation [11] become more complex when AI is tightly coupled through cognitive extension. The main problem will come with extenders that model and monitor human behavior to find the *targeted interventions* that optimize some metric of cognitive enhancement, and can degenerate into sophisticated ways of surveillance and manipulation, well beyond the relatively decoupled smartphones and 'nudging' personal assistants of today.
- Education and assessment: as AI extenders become more powerful and integrated, it will be more difficult (perhaps even unethical) to remove them whenever humans are in education or evaluation contexts. Should college exams be performed with ($A[E]$) or without ($A[\emptyset]$) the cognitive extenders? A job interview? Should cognitive evaluation, including IQ tests, be modified or compared to the situations with and without the cognitive extenders?

All the specifics above suggest that many issues about the impact of AI must be reconsidered in the particular context of AI extenders, including reliability, moral status, value alignment, evaluation, regulation, manipulation and other cultural, political and social impacts.

6 RECOMMENDATIONS

Our general recommendation is that AI extenders must be seen distinctively from other cognitive extenders not using AI and especially from AI systems that are externalized, autonomous, or decoupled. This is the case inasmuch as the questions about the impact of AI extenders become much more subtle and specific than in the other cases.

For instance, do we need to understand what an AI system does? Most technological tools and processes that become strongly integrated with human cognition (e.g., writing, driving, typing, etc.) are not "explained" to humans. There is no need to explain how a pen works, or a keyboard, in order to have a smooth integration between the human and the tool. The role of human-computer interfaces is not explaining how the subsystem works, but creating a reliable interface such that some cognitive processes are created and internalized to take the most of the tool [3, 7, 19, 24].

There are some more specific recommendations that we can direct to AI developers in particular. First, and most importantly, AI developers should take care to distinguish when they are developing an autonomous system, a decoupled system or a system that is meant to be fully coupled. They can inherit the experience of user interfaces and human-computer interaction, such as how the experience of the user will be affected, how comfortable it will be or what the secondary effects are for humans [28]. This means that these systems require not only AI prowess but a strong expertise in human cognition, especially at the level of capabilities and personality traits.

Second, there are lessons for philosophers of mind and psychologists under the view of AI extending human cognition. How do we ensure that these possibilities improve the mind? How do we measure whether they are really 'extensions' and not simply tools? How do we analyze the differences in adaptation before, during and after the cognitive extenders are used? What kinds of innate general abilities, cognitive styles or personalities make extension more powerful? Theories of development may be revised, as some sequences of cognitive extensions may be more effective, beneficial or risky than others. As cognitive extenders are going to modify humans at a profound psychological level, it is important to determine what kinds of interventions and intrusions are potentially dangerous or unethical.

Third, while the notions of autonomy and agency are crucial, regulators and policy-makers should be careful about *only* regulating autonomous systems (e.g., a ban on autonomous weapons) as this still leaves humans vulnerable to being used to circumvent these regulations, creating extended 'zombies', a new way of "undermining responsibility" [8]. Certifications should be given not only after analyzing when the cognitive extender is operating (for a diversity of human users) but most especially when the cognitive extender changes or ceases to operate.

In general, AI is going to present a diverse range of options for achieving some functionalities, with different degrees of coupling. Given the future potential of going beyond human capabilities, a relevant part of AI must focus on cognitive extenders. These *AI extenders* must be designed taking full awareness of the capabilities and autonomy of the extended human jointly with some objective function of what the resulting symbiosis will be able to do, and all the implications of that coupling.

ACKNOWLEDGMENTS

JHO was supported by the EU Spanish MINECO under grant TIN 2015-69175-C4-1-R, by Generalitat Valenciana (GVA) under grants PROMETEOII/2015/013 and PROMETEO/2019/098, the Future of Life Institute under grant RFP2-152, and a Salvador de Madariaga grant (PRX17/00467) from the Spanish MECD and a BEST grant (BEST/2017/045) from GVA. KV was supported by the Leverhulme Centre for the Future of Intelligence, Leverhulme Trust, under Grant RC-2015-067.

REFERENCES

[1] Nathaniel Barr, Gordon Pennycook, Jennifer A Stolz, and Jonathan A Fugelsang. 2015. The brain in your pocket: Evidence that Smartphones are used to supplant thinking. *Computers in Human Behavior* 48 (2015), 473–480.
[2] W Bechtel and G Graham. 1999. A compendium to cognitive science.
[3] Nathalie Bonnardel and Franck Zenasni. 2010. The impact of technology on creativity in design: an enhancement? *Creativity & innov. management* 19, 2 (2010), 180–191.
[4] Nick Bostrom and Anders Sandberg. 2009. Cognitive enhancement: methods, ethics, regulatory challenges. *Science and Engineering Ethics* 15, 3 (2009), 311–341.
[5] Nicholas Carr. 2015. *The glass cage: Where automation is taking us.* Random House.
[6] J Adam Carter and S Orestis Palermos. 2019. The ethics of extended cognition: Is having your computer compromised a personal assault? *Journal of the American Philosophical Association (forthcoming)* (2019).
[7] Shan Carter and Michael Nielsen. 2017. Using Artificial Intelligence to Augment Human Intelligence. *Distill* (2017). https://doi.org/10.23915/distill.00009 https://distill.pub/2017/aia.
[8] Stephen J Cave, Rune Nyrup, Karina Vold, and Adrian Weller. 2018. Motivations and Risks of Machine Ethics [40pt]. *Proc. IEEE* 99 (2018), 1–13.

[9] Christopher Chabris and Daniel Simons. 2010. *The invisible gorilla: And other ways our intuitions deceive us.* Harmony.

[10] Andy Clark and David Chalmers. 1998. The extended mind. *Analysis* 58, 1 (1998), 7–19.

[11] John Danaher. 2018. Toward an Ethics of AI Assistants: an Initial Framework. *Philosophy & Technology* (2018), 1–25.

[12] Ezequiel Di Paolo. 2009. Extended life. *Topoi* 28, 1 (2009), 9–21.

[13] D. C. Engelbart. 1962. Augmenting human intellect: a conceptual framework. Summary Report AFOSR-3233, Stanford Research Institute, Menlo Park, CA.

[14] Susan T Fiske and Shelley E Taylor. 1984. *Social cognition: From brains to culture.* McGraw-Hill.

[15] Brett Frischmann and Evan Selinger. 2018. *Re-engineering humanity.* Cambridge Univ. Press.

[16] Richard Heersmink. 2017. Extended mind and cognitive enhancement: Moral aspects of cognitive artifacts. *Phenomenology and the Cognitive Sciences* 16, 1 (2017), 17–32.

[17] J. Hernández-Orallo. 2017. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence.* Cambridge University Press.

[18] Edwin Hutchins. 1999. Cognitive artifacts. *The MIT encyclopedia of the cognitive sciences* 126 (1999), 127.

[19] Julie A Jacko. 2012. *Human computer interaction handbook: Fundamentals, evolving technologies, and emerging applications.* CRC press.

[20] Joseph CR Licklider. 1960. Man-computer symbiosis. *IRE transactions on human factors in electronics* 1 (1960), 4–11.

[21] Richard Menary. 2007. *Cognitive integration: Mind and cognition unbounded.* Palgrave Macmillan.

[22] M. L. Minsky (Ed.). 1968. *Semantic Information Processing.* MIT Press.

[23] Grzegorz J. Nalepa, Angelo Costa, Paulo Novais, and Vicente Julian. 2018. Cognitive Assistants. *International Journal of Human-Computer Studies* 117 (2018), 1–68.

[24] Allen Newell and Stuart K Card. 1985. The prospects for psychological science in human-computer interaction. *Human-computer interaction* 1, 3 (1985), 209–242.

[25] H. Putnam. 1960. Minds and Machines. In *Dimensions of Mind*, S. Hook (Ed.). New York University Press.

[26] Hilary Putnam. 1967. The Nature of Mental States. (Originally published as 'Psychological Predicates'). Vol. 1. University of Pittsburgh Press, Reprinted in Rosenthal (1971): 150-61, 37–48.

[27] Evan F Risko and Sam J Gilbert. 2016. Cognitive offloading. *Trends in Cog. Sciences* 20, 9 (2016), 676–688.

[28] Yvonne Rogers, Helen Sharp, and Jenny Preece. 2011. *Interaction design: beyond human-computer interaction.* John Wiley & Sons.

[29] William B Rouse and James C Spohrer. 2018. Automating versus augmenting intelligence. *Journal of Enterprise Transformation* (2018), 1–21.

[30] Mark Rowlands. 2010. *The new science of the mind: From extended mind to embodied phenomenology.* MIT.

[31] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* (2017).

[32] Paul R Smart. 2018. Human-extended machine cognition. *Cognitive Systems Research* 49 (2018), 9–23.

[33] Betsy Sparrow, Jenny Liu, and Daniel M Wegner. 2011. Google effects on memory: Cognitive consequences of having information at our fingertips. *Science* 333, 6043 (2011), 776–778.

[34] Jim Spohrer and Guruduth Banavar. 2015. Cognition as a service: an industry perspective. *AI Magazine* 36, 4 (2015), 71–86.

[35] John Sutton. 2006. Distributed cognition: Domains and dimensions. *Pragmatics & Cognition* 14, 2 (2006), 235–247.

[36] P Thagard. 1992. *Conceptual revolutions.*

[37] Paul Thagard. 1996. *Mind: Introduction to cognitive science.* Vol. 4. MIT press Cambridge, MA.

[38] Karina Vold. 2018. Overcoming Deadlock: Scientific and Ethical Reasons to Embrace the Extended Mind Thesis. *Philosophy and Society* 29 (2018), 475. Issue 4.

[39] L. von Ahn. 2005. *Human Computation.* Ph.D. Dissertation. Carnegie Mellon University.

[40] L Vygotsky. 1978. *Mind in society.* Harvard University Press.