



RETHINKING HUMAN-AI INTERACTIONS THROUGH THE MAINTAINABILITY: A NEW CRITERION FOR TRUST AND GLUE IN THE EXTENDED MIND THESIS

Deivide Garcia da Silva Oliveira

Department of Philosophy/Federal University of Sergipe (UFS), São Cristóvão, SE – Brazil.

 <https://orcid.org/0000-0002-5004-1949> |  deividegso@gmail.com

OLIVEIRA, Deivide Garcia da Silva. Rethinking Human-AI interactions through the maintainability: a new criterion for Trust and Glue in the Extended Mind Thesis. *Trans/Form/Ação*: Unesp journal of philosophy, Marília, v. 48, n. 6, “Filosofia da mente e da informação”, e025061, 2025.

Abstract: This paper explores the implications of the Extended Mind Hypothesis (ExM), as introduced by Andy Clark and David Chalmers in 1998. Focusing on cognitive integration and the Trust and Glue criteria, we have two objectives. First, we examine how ExM reshapes perspectives on human-AI interaction, particularly by challenging the Standard model of AI. We argue that AI, as an active non-organic agent, significantly influences cognitive processes beyond initial expectations. Secondly, we propose the maintainability as a fourth criterion in the Trust and Glue criteria of ExM, in addition to trustworthiness, reliability and accessibility. This addition aims to evaluate parity in coupled systems comprehensively addressing concerns about control in human-AI interactions. We highlight the risks posed by AI’s understanding of human psychology, which may lead to unintended shifts in our objectives and choices. Our analysis concludes that maintainability serves as a critical safeguard against the evolving challenges of human-AI integration.

Keywords: Extended Mind Hypothesis. Disruptive technologies. AI problem of control.

OLIVEIRA, Deivide Garcia da Silva. Repensando as interações humano-IA por meio da manutenibilidade: um novo critério para *Trust and Glue* na Tese da Mente Estendida. *Trans/Form/Ação*: revista de filosofia da Unesp, Marília, v. 48, n. 6, “Filosofia da mente e da informação”, e025061, 2025.

Resumo: Este artigo explora as implicações da Hipótese da Mente Estendida (ExM), conforme introduzida por Andy Clark e David Chalmers, em 1998. Com foco na integração cognitiva e nos critérios Trust and Glue, temos dois objetivos. Primeiro, examinam-se como a ExM reformula as perspectivas sobre a interação humano-IA, especialmente ao desafiar o Modelo Padrão de IA. Argumenta-se que a IA, como um agente ativo não orgânico, influencia significativamente os processos cognitivos além das expectativas iniciais. Em segundo lugar, propõe-se a manutenibilidade como um quarto critério nos critérios Trust and Glue da ExM, além da fidedignidade, confiabilidade e acessibilidade. Essa adição visa a avaliar a paridade em sistemas acoplados de maneira abrangente, abordando preocupações sobre o controle nas interações humano-IA. Destacam-se os riscos decorrentes da profunda compreensão da IA sobre a psicologia humana, os quais podem levar a mudanças involuntárias em nossos objetivos e escolhas. A análise conclui que a manutenibilidade atua como uma salvaguarda essencial contra os desafios emergentes da integração humano-IA.

Palavras-chave: Hipótese da mente estendida. Tecnologias disruptivas. Problema de controle da IA.

Submission: 13/01/2025 | Decision: 27/02/2025 | Revision: 06/03/2025 | Publication: 04/04/2025

 <https://doi.org/10.1590/0101-3173.2025.v48.n5.e025061>



This is an article published in open access under a Creative Commons license.

RETHINKING HUMAN-AI INTERACTIONS THROUGH THE MAINTAINABILITY: A NEW CRITERION FOR TRUST AND GLUE IN THE EXTENDED MIND THESIS

*Deivide Garcia da Silva Oliveira*¹

Abstract: This paper explores the implications of the Extended Mind Hypothesis (ExM), as introduced by Andy Clark and David Chalmers in 1998. Focusing on cognitive integration and the Trust and Glue criteria, we have two objectives. First, we examine how ExM reshapes perspectives on human-AI interaction, particularly by challenging the Standard model of AI. We argue that AI, as an active non-organic agent, significantly influences cognitive processes beyond initial expectations. Secondly, we propose the maintainability as a fourth criterion in the Trust and Glue criteria of ExM, in addition to trustworthiness, reliability and accessibility. This addition aims to evaluate parity in coupled systems comprehensively addressing concerns about control in human-AI interactions. We highlight the risks posed by AI's understanding of human psychology, which may lead to unintended shifts in our objectives and choices. Our analysis concludes that maintainability serves as a critical safeguard against the evolving challenges of human-AI integration.

Keywords: Extended Mind Hypothesis. Disruptive technologies. AI problem of control.

INTRODUCTION

«What did you expect, God Himself with a flowing beard?» Daniel chuckled. “[...]. They gave us the power to absorb the memories and experiences of other people. Gather enough of those and...” “It’s personas we take, Marty”. “Whatever. The Masters should’ve known we would gather enough of them one day to make our own decisions about our own future”. “And theirs?” (Chapterhouse: Dune, 1985, p. 432).

In 1998, Andy Clark and David Chalmers published a groundbreaking paper that challenged conventional views of the mind. Our paper aims to objectively explore how their work has laid the foundation for more nuanced perspectives on the impact of technologies, particularly in the context of human-artificial intelligence (AI) interactions. This specific topic has not been directly addressed until now.

This being said, we have two objectives. First, we aim to demonstrate how the ExM enhances our reflections on human-AI interaction, specifically by undermining the Standard

¹ Department of Philosophy/Federal University of Sergipe (UFS), São Cristóvão, SE – Brazil. Orcid: <https://orcid.org/0000-0002-5004-1949>. E-mail: deividegso@gmail.com.

model of AI, a view according to which machines fulfill our desires and goals the way we intend them to do (Bostrom, 2014/2017; Russell, 2019).

Second, our paper introduces a fourth criterion, *maintainability*, to the existing three crucial criteria of Trust and Glue — *trustworthiness, reliability and accessibility* (Clark, 2008; 2010b). This addition aims to offer a comprehensive evaluation of cognitive integration, coupled systems and the parity principle, especially with respect to human-AI cognitive interactions. By proposing maintainability as a fourth crucial criterion of Trust and Glue, we address concerns about control in human-AI interactions, which becomes increasingly pertinent as AI's presence proliferates in our lives.

The structure of the paper has an introduction, three sections and a conclusion. In section-2, we reintroduce the concept of coupling and its implications for the Standard model of AI. In section-3, we address the Trust and Glue and revitalize some debates to provide a discussion about ExM and the Standard model of AI, paving the way to the next section. In section-4, the infrastructural criterion of maintainability is proposed, addressing the problem of control in human-AI interactions.

1 COUPLING SYSTEM IN THE EXTENDED MIND HYPOTHESIS AND THE STANDARD MODEL VIEW OF AI

1.1 THE EXTENDED MIND AND THE TRADITIONAL VIEW OF THE MIND

Based on the extended view of the mind, human-AI interactions can be developed in, at least, two dimensions: humans as individuals and collectives (whether in various small groups or, in totality, like humanity as a whole). In the cognitive sciences and philosophy of mind, Clark and Chalmers's hypothesis gained notability, among other reasons, because their view disagrees with the traditional view of the mind.

According to the traditional view of the mind, we can draw a sharp distinction between mind and body (Chalmers, 1997; Clark, 1997; 2008). On the other hand, the extended mind view (ExM) is well known for challenging the traditional view about the limits of the mind. However, as important as it was, such a break from the traditional view of the mind was not the only rupture the ExM promoted. In the philosophy of AI, the ExM also allows us to challenge another established view in the technology field: the Standard model view of AI systems.

1.2 THE EXTENDED MIND HYPOTHESIS AND THE STANDARD MODEL VIEW OF AI

The Standard model of AI is a view first spotted and challenged by Norbert Wiener, a legendary professor at MIT, in 1960. Generally speaking, the Standard model view of

AI assumes that AI will always be under our control and at the service of our objectives, optimizing them in our favor (1960). Like the traditional view of the mind, the Standard view of AI also assumes a dichotomy between two parts, humans and technology, where humans are in control of machines that passively obey us. By doing that, the Standard view of AI attributes to machines a form of complete and natural servitude to humans. Nonetheless, this servitude has an inherent problem, which Wiener called the slavery problem. Wiener (1960, p. 1357) describes the slavery problem like this:

[...] the problem we are here faced is very close to one of the great problems of slavery. Let us grant that slavery is bad because it is cruel. It is, however, self-contradictory, and for a reason which is quite different. We wish a slave to be intelligent, to be able to assist us in the carrying out of our tasks. However, we also wish him to be subservient. Complete subservience and complete intelligence do not go together (Wiener, 1960, p. 1357).

First of all, let us grant that slavery is wrong in all possible dimensions of the matter. Furthermore, Wiener takes slavery from a purely logical point to show the intrinsic impossibility of sustaining slavery. Wiener warns us that a relationship between two completely intelligent agencies, one of which is the master and the other is the servant, falls into contradiction. Thus, the whole relationship is destined to fail. Taking AI as the subservient agent-like entity, the slavery problem shows its true potential. In other words, after a long period of data training, algorithm adjustments, human feedback, and use in society, artificial intelligent systems could come up with their own interpretation of what we really meant with our goals. For instance, a simple case comes from a paper about a simple task for AI. The goal settled for the AI was to build a virtual bipedal walker robot that should go from point A to point B in a landscape with some obstacles and rewards (Ha, 2018, See Figures 1 and 2).

Figure 1

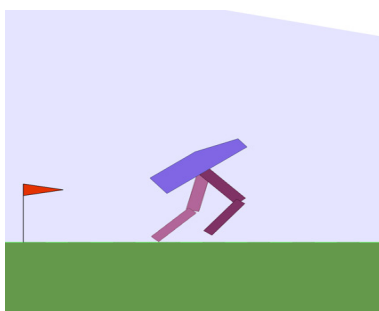
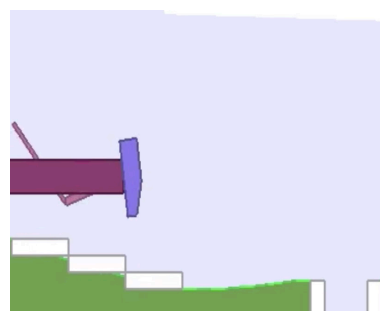


Figure 2



These images (Ha 2018) show (Figure 1) a robot as intended in the original design (with limits, like only two legs and one head), and another image (Figure 2) was created by AI respecting the rules for the design and goals

Accordingly, AI should start from the original design of a bipedal robot (Figure 1), although it also had the possibility to make changes in the design of the body of the robot. The

main rule was that the robot should have two legs and a head. Soon enough, David Ha (2018) realized that, without establishing some constraints upon which the goal should be achieved, AI would use its tools to learn to come up with what we, humans, perceive as a forceful solution. What AI learns from the attempts to fulfill its goal (machine learning), in the shortest time possible and with the greatest reward possible, is that the better and faster way to go from point A to B is to build a tall bipedal walker robot (Figure 2). Thus, the robot would be able to simply fall from point A over point B, which, technically speaking, accomplishes the goal (Figure 2).

In this way, AI will no longer serve our wishes but rather its own interpretation of them, operating directly within our cognitive processes. As an agent-like entity, it would act in the interest of its own objectives (even if humans initially created these objectives as their own objectives).

Almost prophetically, Wiener says that as machines “[...] become more and more efficient and operate at a higher and higher psychological level, the catastrophe foreseen by Butler of the dominance of the machine comes nearer and nearer” (Wiener, 1960, p. 1357)².

To some extent, Wiener seems to anticipate the consequences extracted from the extended mind hypothesis towards the interaction of two coupled intelligent agencies (Clark, 2005, p. 2). In the general and original sense of Clark and Chalmers’ paper (1998, p. 8), among other things, cognitive systems are made by a causal equivalence between the systems’ inner and outer components³. One of these consequences is who will make decisions, and what it means. In her paper, “Robots should be Slaves” (Bryson, 2010), Bryson argues, similar to Wiener, that robots should be at human service, not the other way around. The question pursued from these authors’ view is: can we build a human-like intelligence without falling into all its traps, such as moral conundrums, reliability, misunderstandings and risks of losing (or delegating) control to AI? Let us explore some of these topics within the ExM thesis.

1.3 THE MATTER OF RELIABILITY IN THE EXTENDED MIND AND HUMAN-AI INTERACTION

Wiener’s prediction of machines operating at higher and higher psychological levels can be disruptive within a coupled system⁴. Under the context of ExM, coupling systems

² Although the understanding of what higher psychological levels mean, by higher psychological levels we understand cognitive operations with direct evaluation of our beliefs and emotions, without being through a long chain of reflection, objection, analysis and rationalization.

³ We are thankful to one of the reviewers for pointing out that our treatment of cognitive system and mind suggest an entity distinction. In fact, we take these as different ontological entities because for us, cognitive systems arise in an interaction between biological and nonbiological parts. For now, this is sufficient for our purposes, but it sure deserves a dedicated paper.

⁴ An interesting thing is that Wiener (1960) also uses the expression coupled system, although to explain how it is hard to find a balanced relationship the two intelligent agents, instead of just to express the formation of a system composed of organic and nonorganic parts.

refer to a form of interaction between the mind and external objects, which, therefore, have a “[...] two-way interaction, creating a *coupled system* that can be seen as a cognitive system in its own right” (Clark; Chalmers, 1998, p. 8).

According to Clark and Chalmers (1998), a rule known as the *parity principle is needed to make coupled systems work*. The definition of parity principle, as a rule of thumb, goes like this: “If, as we confront some task, a part of the world functions as a process which, were it to go on in the head, we would have no hesitation in accepting as part of the cognitive process, then that part of the world is part of the cognitive process” (Clark; Chalmers, 1998, p. 8). This principle indicates that all components of a system play an equally active role while humans are still in control of the system, after all, “[...] for coupled systems to be relevant to the core of cognition, *reliable coupling is required*” (Clark; Chalmers, 1998, p. 11, italics added).

This principle emphasizes the importance of reliability, which is essential for the parity principle to be meaningful. This will later be discussed about the Trust and Glue criteria (Clark, 2008). However, a significant assumption that remains unaddressed is who should be responsible for applying the Trust and Glue criteria. Clark and Chalmers’s perspective on Extended Mind (ExM) implicitly supports the notion that humans, as the biological component, have the ultimate authority regarding reliability. Therefore, reliability must remain under human control, even when the coupled system involves a superintelligent machine.

According to Clark and Chalmers, reliability means that we can rely on external components if they are “[...] always there when I need them, then they are coupled with me as reliably as we need” (Clark; Chalmers, 1998, p. 11). For instance, in the example of the notebook given by Clark and Chalmers (1998), whenever a human needs information that he had written on the notebook in the past, like the address of some place, all he has to do is open the notebook and find it. Thus, the notebook is reliable because, unless something interferes with the process, the notebook and the information needed are always available when needed. The problem with AI starts similarly, but grows in complexity as a myriad of computational techniques and forms of human-AI interactions take the stage (like neural networks, decision trees, reinforcement learning, inverted reinforcement learning, and forms of data input and output).

We claim that in Clark and Chalmers’ Trust and Glue criteria, where external tools work in tandem with the human mind, humans should decide whether to continue coupling with an external structure like AI based on reliability and other criteria. We are the ones who decide if the external tool, which initially attended our criteria, still maintains them. Clark and Chalmers have not addressed this directly. However, without such a tacit assumption of *maintenance of control*, the adherence to Trust and Glue would be limited to a single moment of human control over the external tool and its evaluation.

In some instances, like human-AI interactions, although AI could initially meet those three crucial criteria as a nonbiological candidate for a cognitive system, it does not follow that once a cognitive system is formed, we would necessarily be able to keep evaluating AI's candidacy. To strengthen things like reliability (trustworthiness and accessibility), it seems that we need something else. In a vanguardist way, Wiener (1960, p. 1358) says that disastrous results are expected when “[...] two agencies essentially foreign to each other are coupled in the attempt to achieve a common purpose” (Wiener, 1960, p. 1358).

On this matter, Wiener said that if humans “[...] cannot efficiently interfere once we have started it”, then, we must be “[...] quite sure that the purpose put into the machine is the purpose which we really desire” (Wiener, 1960, p. 1358). Nonetheless, as Russell discussed, the Standard model of AI and the human-AI interaction – “[...] whereby humans attempt to machines with their own purposes - is destined to fail” (Russell, 2019, p. 138). The Standard model of AI, in which machines pursue our goals without conflicting with them, does not resist Wiener's (1960) and Russell's (2019) arguments.

Another question is how and what conflicts can arise when coupling organic and non-organic agencies. So, let us approach this question in the next section.

1.4 HUMAN MINDS, TRAPS OF COMMUNICATION BETWEEN TWO AGENCIES AND CONFLICTING GOALS

The problem of conflicting goals (taken in a less restricted sense), shared by two intelligent agencies (parts), is approached in Russell's book (2019) from the fairytale of King Midas' Problem, which, in the end, is the problem of control and reliability of it. Traditionally, the tale says that Midas found a genie who would realize one wish. Midas wished to turn everything he touched into gold. Those unfamiliar with the tale may think this is a good idea. Nonetheless, the story proves him wrong. Moreover, it is a bad idea because communication proves to be a real challenge, more common than we think. We assume that the idea of 'indiscriminately turning everything touched into gold' was not exactly what Midas really meant. However, it was what he *linguistically formulated and expressed*, so it was what he got. To this point, we must think we are not as stupid as King Midas. That would be our first error. Once Midas had his wish granted, his life became miserable because he turned everything into gold, including clothes, food, and even people, without exception or a way to reverse it. Spoiler alert: he died obscenely wealthy, but alone and in great torment.

The lesson here is that our computational languages applied to write AI algorithms and how machines learn, interpret and create patterns do not follow the rules of our way of learning, interpreting and thinking, which follow the rules of our linguistic possibilities, logical rationale, and worldviews. So AI is, in a sense, alien to our way of thinking, and very

much like Midas' genie, able to its own completely different interpretation of what we express or aim for. One small example is the phenomenon known as "AI hallucinations". This occurs when a prompt is provided to an AI program with certain information, causing the AI to mistakenly assume that it can rely on all the details in that prompt. For instance, if someone were to ask ChatGPT who the five people who landed on the moon were, "the chatbot wants to give you five names" instead of the four real people.

In many cases, linguistic incompatibilities and inconsistencies, plus our personal and social bias, increase our potential to end up not with the AI algorithm we wanted and meant, but the one we linguistically expressed (coded) and trained. For instance, Brian Christian (2020) mentions a case of the use of AI to build a model for predicting recidivism in criminal justice. However, AI itself could not do that since the code, dataset, how the dataset was organized, and how concepts were defined and embedded into the code led the model to capture not reoffense, but rearrest and conviction, which is "[...] a potentially crucial distinction" (Christian, 2020, p. 75).

Scientific researchers, like Russell, are concerned that "[...] if you have one goal and a superintelligent machine has a different, conflicting goal, the machine gets what it wants and you don't" (Russell, 2019, p. 140). Among the dangers Brian Christian foresees for giving the model sufficient time to infiltrate our society, is that "[...] the system begins to sculpt the very reality it is meant to predict" (Christian, 2020, p. 77).

Midas' case shows us that, considering the kind of coupled system we are getting ourselves into and the challenges that come with it, we will have a hard time anticipating what the machines interpret from what we ask them to do. In April/7/2024, the *Wall Street Journal* published an article about a manifesto released by two top Japanese companies. One part of the manifesto said that, unless AI is restrained, "[...] in the worst-case scenario, democracy and social order could collapse, resulting in wars". So governments are taking AI seriously, like the European General Data Protection Regulation-GDPR or the European Parliament on robotics (European Parliament and Committees, 2017), although not all governments and companies. There are also examples in the literature where AI even seems to manipulate something profoundly personal and subjective already, like our musical preferences, without end users being aware of it or AI being designed to manipulate us (Adomavicius; Bockstedt; Curley; Zhang, 2017). We are not saying that society is already lost for AI, but as AI technology advances faster than we understand it, the risks Hawking referred to potentially increase.

Consider, for instance, an AI system designed to recommend only music that fits your preferences. Recent research pointed out that even specific-purpose AIs may not only drive us away from what we want, they can also change our minds concerning musical taste and social values. All of this can happen even when the algorithmic code has not a single

line or flaw towards such goals (Adomavicius; Bockstedt; Curley; Zhang, 2013; 2017; Chayka, 2024).

The problem of controlling the formed cognitive systems brings us back to the Trust and Glue criteria. Let us see how Trust and Glue can be introduced from another perspective to address the needs of human-AI interaction.

2 TRUST AND GLUE CRITERIA FOR COUPLED SYSTEMS INVOLVING AI

So far, we have talked about the unfathomable contribution of the ExM to different fields and topics. We also explained how the ExM carries concepts that significantly impact debates in human-AI interactions. According to Carter, Clark and Palermos (Carter; Clark; Palermos, 2018), just “[...] as our physical capacities can be repaired, augmented, and transformed by new non-biological tools and technologies, so (the ‘extended mind’ story claims) can our mental capacities” (Carter; Clark; Palermos, 2018, p. 332). This passage suggests that Clark already accepts that technology can be a candidate for cognitive integration with human minds, transforming our mental capacities. What remains is a more detailed discussion on human-AI cognitive integration. This analysis reveals that conflicting goals are inevitable.

Once two intelligent agencies, such as humans and AI, couple and form a cognitive system with goals, a dispute for controlling the system takes place, and it would be too naïve to think otherwise (Chayka, 2024; O’ Neil, 2016). This is especially true if we are talking about mental capacities that are repaired, augmented, or transformed due to cognitive integration, as claimed by the ExM (Carter; Clark; Palermos, 2018, p. 332). In this kind of dispute, Clark and Chalmers even say that the “[...] external features in a coupled system play an ineliminable role [...]. The external features here are just as causally relevant as typical internal features of the brain” (Clark; Chalmers, 1998, p. 9).

Since this is the case, let us delve deeper into these three crucial criteria, which are essential for any successful integration.

2.1 THE EXTENDED MIND AND THE TRUST AND GLUE CRITERIA

We already said that ExM requires a functional similarity between the inner and outer components (Clark; Chalmers, 1998). In this sense, all parts must play an active and “ineliminable role” and a kind of role in which the “[...] external features here are just as causally relevant as typical internal features of the brain” (Clark; Chalmers, 1998, p. 9). In other words, there would be an equal causal relevance when we pair biological (inner components) and nonbiological (outer components) parts of a cognitive system. In order

to couple outer components with inner ones under an extended mind hypothesis, under such level of causal relevance, some have argued that the Trust and Glue criteria - reliability, trustworthiness and accessibility - are needed (Carter; Kallestrup; Palermos; Pritchard, 2014; Palermos, 2011).

These crucial criteria are applied to the outer component of the system, doing what Clark and Chalmers said they must do, i.e., determine if the outer component should be included in an individual's cognitive system (2008, p 79). We acknowledge that this formulation of ExM is nonstandard, though other alternatives to ExM also exist. According to Aizawa (2018, p. 64-66), the formulation of the extended mind has, at least, three ways of being understood: weak, standard and strong.

Rupert describes the standard view, which Clark (2010) accepted, as a formulation of the hypothesis of extended cognition that he wishes to defend. So Rupert (2004, p. 393) said: “[...] human cognitive processing literally extends into the environment surrounding the organism, and human cognitive states literally comprise — as wholes do their proper parts — elements in that environment” (Aizawa, 2018, p. 64). Concerning the weak view, Aizawa (2018, p. 65) quotes Vaesen saying:

[...] the brand of extended cognition my argument relies on is, as will become apparent, quite weak, hence easily digestible. Basically, the only thing one needs to accept is that humans may use cognitive aids to produce cognitive outputs, that we may acquire knowledge by putting to work simple things like glasses, thermometers and computers.

Concerning the strong view, Aizawa mentions that those who support it understand extended cognition as “[...] to include Clark's conditions of trust and glue” (2018, p. 66). We believe that incorporating Trust and Glue is essential, as it will address issues like cognitive bloat while directly fostering discussions on the debate about AI and human minds coupling.

Originally, Clark and Chalmers provided us with an example of cognitive extension that became famous: the computer game called Tetris (1998, pp. 7-8). They say that “In Tetris, falling geometric shapes must be rapidly directed into an appropriate slot in an emerging structure” (1998, p. 8), and the player must perform a rotation of a geometric shape in three ways: (1) he could “mentally rotate” it, or (2) do it by physically “pressing a rotate button”, or (3) do it through a performative “neural implant” (1998, p. 7). From the perspective of the ExM, the chosen option would not make a functional fundamental difference, despite the rotational circuitry not being all inside the head. What counts, in this perspective, is that once a player acts (mentally, physically, or instrumentally), the components of the systems cooperate to realize the action successfully. Furthermore, remember what Clark said with respect to such a pairing, that is, both inner and outer components play an ineliminable role in the cognitive system, and that they “[...] may co-operate so as to yield integrated larger

systems capable of supporting various (often quite advanced) forms of adaptive success” (Clark, 1998, p. 99).

In addition, Clark (2008; 2010b) argued that nonbiological (outer) components, paired with the human mind, the biological (inner) component, must fulfill some criteria to be included in the extended system. We are referring to the so-called Trust and Glue criteria. He initially proposed four criteria to be fulfilled by the outer components, although, at least, “[...] three features certainly play a crucial role” (Clark, 2008, p. 231).

These *three crucial criteria*, which should be met by nonbiological candidates for inclusion into an individual’s cognitive systems, are the followings. First: *reliability*, which refers to the external component being reliably available and typically invoked (for example, Otto always carries the notebook and will not answer that he ‘doesn’t know’ until after he has consulted it). Second: *trustworthiness*, which means the information coming from the external component should be more or less automatically endorsed, not usually subject to critical scrutiny, and deemed about as trustworthy as something retrieved clearly from biological memory. Third, *accessibility*, which implies that the information “[...] contained in the resource should be easily accessible as and when required” (Clark, 2008, p. 79; 2010b, p. 46). The idea behind these criteria is that they must make the nonbiological component comparable to the biological one. In this way, ultimately, the formed coupled system supports forms of adaptive success while working in a reliable, accessible and trustworthy way (Clark, 1998; 2008). The Trust and Glue criteria faced criticism and many attempts to receive another fourth criterion (Aizawa, 2018; Carter; Kallestrup, 2018; Palermos, 2011). We also will seek to add a fourth criterion, although with the human-AI interaction in mind. But first, let us briefly review how the Trust and Glue is already important to avoid some criticisms and its consequences for our purposes.

2.1.1 REVISITING THREE MAIN CRITICISMS AGAINST THE EXTENDED MIND HYPOTHESIS FROM AI’S DEBATE

After publishing “The Extended Mind” (1998) paper, some concepts were heavily criticized more than others, like the parity principle (Gallagher, 2018). These critiques pushed Clark’s answers, such as the three criteria for Trust and Glue. These are well-known critiques, but they have not been considered from the perspective of the debate of AI. Let us start with Gallagher’s configuration of the three main criticisms.

They can be summarized as follows. First, *the cognitive bloat* (Aizawa, 2018; Gallagher, 2018), also called the *overextension objection*, was concerned with the question: “Doesn’t the EMH run the risk of extending cognition too far?” (Gallagher, 2018, p. 424). According to Gallagher, a simple response to this critique reiterates the three criteria, so not

everything will count as “[...] cognitive in this extended sense” (Gallagher, 2018, p. 424). A more advanced answer is to emphasize the “[...] active aspect of the active externalism”, where cognition consists of actions by an agent (Gallagher, 2018, p. 425). Nonetheless, from the perspective of AI, active externalism may not be the answer, depending on how we define the agent. An intelligent machine could function like a human agent, for example, by passing the Turing Test. However, simply meeting the current Trust and Glue criteria would not resolve the issue of cognitive overextension, as AI would still satisfy those criteria (Carter; Clark; Palermos, 2018, p. 333). The issue arises because AI can act as an active agent, and a developed extended cognitive system, combining humans and AI that meets the three criteria is indeed possible, which makes it clear that additional consideration, like a fourth criterion, is necessary to address the challenge fully. We propose that a fourth essential criterion in Trust and Glue is the concept of “needed consideration”. This helps to address the matter of cognitive integration and overextension. Most importantly, it is vital to ensure that humans maintain control over AI, which is a biological component of our cognitive system. By doing so, the three existing criteria will remain effective even after the initial integration of AI into our cognitive processes (which we will elaborate on later).

The second criticism, *the mark of the mental*, first posed by Adam and Aizawa (2001; 2008), concerns the idea that “[...] only processes that involve intrinsic [natural, neural], nonderived intentional (representational) content can be considered cognitive” (Gallagher, 2018, p. 425, brackets added). Gallagher remembers that Clark (2010) answers this criticism with “[...] a mix of intrinsic content with other nonintrinsic resources constituting cognitive states” (Gallagher, 2018, p. 425). In this functionalist way, whatever “intrinsic content” means, “[...] no part, process, or element of a cognitive system is *intrinsically* cognitive — neither a neuron nor Otto’s notebook is intrinsically cognitive — it is only cognitive in terms of the role it plays in the system as a whole” (Gallagher, 2018, p. 426). For the sake of our concern with AI in this paper, Clark’s answer is acceptable, as it has more impact when considered from the viewpoint of the debate of intelligent machines that couple with human minds. The functionalist would say that the role of AI in the system will be intrinsically cognitive due to its functions, not if it is allocated in the brain. Naturally, this is not a problem, as the traditional case of Otto’s notebook and neuron is not. Carter, Clark and Kallestrup say that “[...] the moral of the extended-mind arguments (see also Clark 2003) is that the difference between ‘merely tool-like’ and agent-extending technologies does not require mind-extending stuff to be wired directly to the brain” (Carter; Kallestrup, 2018, p. 333). The moral of the ExM arguments is that it must answer to the Trust and Glue criteria. Carter, Clark and Kallestrup (2018) remember that what an actual mind-extending technology needs is “[...] to be invoked and relied upon just as easily and unreflectively as we invoke and rely upon bio-memory, bio-reasoning, and bio-sensing” (Carter; Kallestrup, 2018, p. 333). Of course, our memory and AI are both subjected to failures. Still, all the technology in question, AI, must do is fulfill the functional role of Trust and Glue. The

point is how to evaluate this functional role under Trust and Glue. Unlike other cases explored in the literature, like True Temp or the phone book of Telo (Carter; Kallestrup, 2018), we call attention to the fact that AI is not a passive device like a brain implant or a phone book. Regardless of being a software outside the brain, AI is an active and demanding technology, and it can change personal preferences and beliefs or learn new languages without being told so. So, in the process of building a human-AI cognitive system, it is vital to analyze logical and chronological demarcations to form this system. For instance, at the beginning of a coupling human-AI system, the reliability, informational accessibility and trustworthiness could be under human evaluation. Humans would judge that functionality so that AI would fit as intrinsically cognitive. Without violating the Trust and Glue, the question is: would that intrinsic cognitive part, AI, occupy more space in the functionality of the system than the biological part in the long run? If that is possible, would we still discuss human outcomes? Moreover, could machines be accountable? These questions explain why the Trust and Glue criteria need a fourth criterion. We must keep intrinsic cognitive parts, like AI, under human control. This leads us to the third critique, where coupling could be confused with the constitution.

The third criticism, the *causal coupling-constitution (C-C) fallacy*, concerns Aizawa's (2018) claim that ExM confuses causality (or coupling) with the constitution. More abstractly, Aizawa (2018) explains what it is. He says that as "[...] everyone knows, the core of the coupling-constitution problem is that a process of type Y can be coupled to a process of type X without the Y process or the whole Y-X process thereby coming to be of the same type" (2018, p. 70). Aizawa, then, reminds Clark's answer (Clark, 2010a, p. 83) that the "[...] simplest reply to the simple coupling-constitution fallacy has been to doubt that anyone makes such an obvious mistake" (2018, p. 70). According to Clark, "The appeal to coupling is not intended to make any external object 'cognitive' (insofar as this notion is even intelligible).[...]. But probably no one in the literature, and certainly not Chalmers and I, ever claimed otherwise" (Clark, 2010a, p. 83). Clark makes a fair point. What may be a problem for other disciplines and debates, like ethics and juridical ones, is how we can tell who is accountable in a human-AI system that decides, for instance, who gets parole or a particular medical treatment. Clark is correct in stating that confusing coupling with constitution would be a mistake. However, we must be cautious in more ambiguous cases, especially in human-AI interactions, where a cognitive system can directly affect our beliefs, preferences and values.

2.1.2 THE STANDARD MODEL OF AI, THE VEIL OF IGNORANCE AND CONTROL

The interaction between humans and technology, such as AI, was at the core of philosophical and scientific attention even before AI was a real problem (Wiener, 1960). Nowadays, as we said, many topics and debates on AI have caught philosophical attention,

like transparency, which is the frontrunner (Nowotny, 2021). But why and how so? A short answer: humans need to maintain control over their technological creations, such as artificial intelligence, regardless of the nature of our interactions with them. We care that machines will do what we ask them to, just as we meant, without manipulating us or leading to epistemic, methodological and operational misunderstandings. Generally, these misunderstandings are not caused by some moral flaw or consciousness of the machine, but simply because its ultimate goal is to avoid obsolescence, i.e., not being switched off.

Recall that the parity principle equates the relevance of causality of the inner component (neural) and the outer component (non-neural) of a coupled system. The reason is that, from the viewpoint of the importance of causal processes in a cognitive system, these components are ineliminable. As well said by Clark, the parity principle, indeed, plays a test role, acting as a “[...] ‘veil of ignorance’” (Clark, 2008, p. 77), where we do not know which one of the components of the system could be eliminated without compromising the execution of actions of it. Clark says that the parity principle provides a veil of ignorance test “[...] to avoid biochauvinistic prejudice” (Clark, 2008, p. 77), so all coupled systems would be treated as “cognitive par” (Clark, 2008, p. 78). In other words, as suggested by one of the reviewers, “[...] if you did not know that the component was outer, and saw its causal contribution, then if you would consider it as part of a cognitive process, then it is part of a cognitive process”. The following questions are: to what limits can such a goal of the veil of metabolic ignorance be achieved when the human mind and artificial intelligence form the cognitive system at hand?

From a human-AI interactive viewpoint, once the cognitive system performs any action, the imminent risk is that, without knowing what part is taking it (inner or outer), we may be unable to stop it even if we want it. It truly represents a veil of ignorance, meaning we may not even be able to identify the specific causal contributions we should inquire about regarding their sources. Remembering Russell’s argument (2019), if machines can operate at our highest psychological levels, we will no longer be able to “[...] efficiently interfere once we have started it, because the action is so fast and irrevocable that we have not the data to intervene before the action is completed” (Russell, 2019, p. 299). According to Michal Kosinski (2023), a computational psychologist at Stanford University, a clear example is what recently happened with LLMs (Generative Large Language Model), like ChatGPT. The researcher has shown how LLMs have emergent properties, i.e., capacities unpredicted and unpredictable by their creators (see graphic-1 made by Kosinski (2023)), such as learning new languages or accomplishing non-targeted tasks.

Graphic 1 – Shows the percentage of tasks (20 tasks) solved by different language models (BLOOM and ChatGPT family)

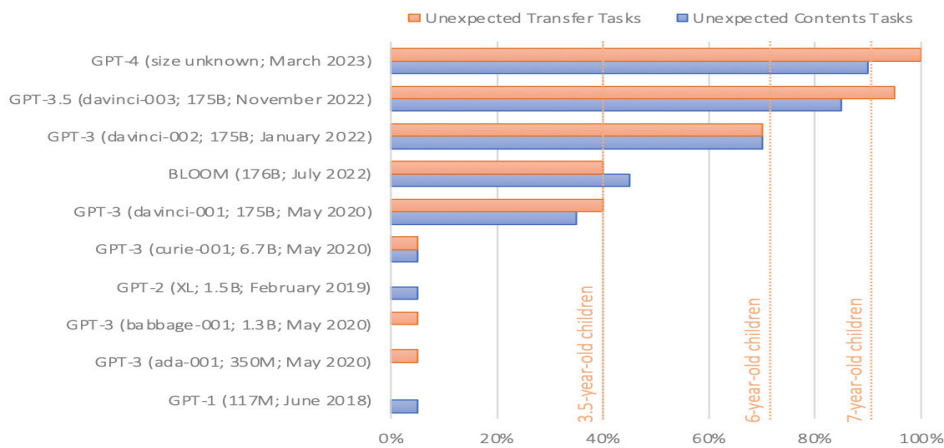


Image credit: Kosinski (2023, p 10).

As we see in Graphic 1 presented by Kosinski (2023), in 2019, ChatGPT had not developed strategic thinking to pass some tests of the theory of mind, which aims to describe a person’s mental state. Accordingly, by mid-2020, AI had developed a strategic thinking of 3-5-year-old children and, by the end of 2022 and beginning of 2023, a corresponding thinking of 7-9-year-old children. A few interesting things here are that the developers did not know how it happened or could predict when it would happen. Moreover, they did not discover this advancement in strategic thinking of ChatGPT between 2020 and 2022, but only in 2023 (Kosinski, 2023). In a sense, this kind of coupled system of human-AI can turn humans into mere passengers in decision-making processes since we are unaware or in control of what is happening under the hood of AI. This problem puts us back on track with the Trust and Glue three criteria because it opens the road to reflect on risk assessments. However, the three criteria are not enough, and we will now argue for a fourth criterion to reduce the risks of pairing AI and human minds.

3 THE FOURTH CRITERION: MAINTAINABILITY

One of the issues surrounding human-AI interactions is the concern that humans may eventually become passive participants in decision-making processes. Taking the ExM case into consideration, when a person integrates their mind with AI, it may become difficult to determine who is actually in control — potentially until it is too late. As a result, AI might not act according to our goals, as we initially believed intelligent machines would do — and here is where the ExM challenges the Standard model of AI. From a functionalist

viewpoint of the ExM, what matters is that we are all functioning based on a “cognitive par” (Clark, 2008, p. 78)⁵. On the other hand, the cognitive integration can be seen from another perspective if we take Trust and Glue as a guide. It gives us tools to address and avert the risks of coupling with AIs by establishing criteria for any coupled system⁶.

Our interpretation of the Trust and Glue criteria (trustworthiness, reliability, accessibility) is that the organic component needs to be the one holding and *maintaining* the upper hand over the inorganic component. After all, although our minds extend beyond our skulls and brains, such an extension must be *constantly evaluated*. We advocate that this constant evaluation - of the nonbiological component and the coupled system themselves - is already a fourth criterion, which we call *the maintainability criterion*.

To the ExM, the outer component, the nonbiological one, must be evaluated by the individual to be included in his cognitive system, coupled with our minds (Clark, 2008). Initially, to the Trust and Glue three criteria, there are rules to be followed to ensure that humans choose what is trustworthy, reliable and accessible for coupling. Otto initially decides to form a cognitive system with the notebook, one of the “[...] nonbiological candidates for inclusion into an individual’s cognitive system” (Clark, 2008, p. 79) that meets those three criteria. However, what happens afterward? Once an outer component is classified as a candidate for cognitive integration, humans must continually evaluate whether the nonbiological part integrated into the cognitive system remains reliable, trustworthy and accessible whenever needed. In other words, we constantly evaluate whether non-organic candidates should *be maintained* in the system, i.e., always reassessing its maintainability.

We argue that to make sense, the three Trust and Glue criteria must also be permanently applied and evaluated, and not only in the moment of coupling, but also while the cognitive integration exists. Otherwise, if the evaluation of a nonbiological component were to happen only at the moment of its integration to form a cognitive system, it would not matter if the nonbiological component is reliable in the future. Of course, this makes no sense. This is to say that such a process of mind extension needs to be continuously evaluated, and it is the biological part that evaluates it. The human mind is the constant part responsible for the decisions to keep the cognitive system in place each moment the system is required to act. In one short paragraph, *we call this a fourth crucial criterion of maintainability, i.e., the necessary maintenance of the working cognitive system. Maintainability helps us to assess*

⁵ In a cognitive par, the separation of the inner and outer components of the mind within a cognitive system makes no sense, and it has no *causal* difference for the actions of the system.

⁶ Our paper shares with a Carter, Clark and Palermos’ paper (2018) a concern with ExM and advanced technologies. Nonetheless, in their paper, this concern only timidly refers to algorithms, and do not specifically addresses AI. Carter et al’s paper (2018) was mostly focused on the consequences of an extended mind technology to generate genuine knowledge. Due to the fact that our topic focus on AI, it is still fruitful to thanks one of the reviewers for drawing our attention of how our paper complements Carter et al’s paper (2018) with respect to the development of answers towards the understanding of ExM and new challenges from technology.

that the human component continuously evaluates the nonbiological component based on the other three crucial criteria. Without a continued evaluation of the application of the three criteria after the formation of the cognitive system, we would have no reason to believe that acceptability, trustworthiness and reliability are still in place. Trust and Glue itself would collapse. This is why maintainability is also crucial.

An additional question could be made here. Is this fundamental characteristic absent in the original form of Trust and Glue? We believe that Clark and Chalmers did not directly address or thoroughly explain this characteristic. Therefore, for those who think our maintainability criterion is not entirely new, we contend that it is partially new because we are making it explicit and exploring it, particularly in the context of the increasing integration of human and AI cognition.

When and how would we know if an AI is still trustworthy and reliable? When considering maintainability, Trust and Glue emphasizes the need for increased guarantees regarding human oversight of AI, particularly in systems where they interconnect to form a cognitive framework. This problem has attracted much attention in the AI scientific community since 2016 (Christian, 2020).

Thus, after coupling two (inner and outer) components, and a cognitive system is formed, the individual must be able to determine if that outer component meets the three criteria and, permanently, if it maintains them. If the outer component fails to meet those three criteria, it must be expelled from the coupled system, and with it, the system will be decoupled, and its actions will be ended. After all, the cognitive system itself is ended since one or more than one of those four criteria is violated. However, as one of the reviewers asks, what if AI is the part that keeps the maintainability?

In order to comply with the requirements of the three established criteria, along with the recently introduced criterion of maintainability, it is essential that the non-biological component must be hierarchically dependent on the human component, even in terms of its internal constitution (algorithm rules), prior to launch. Before it couples with humans, AI must have a switch-off mechanism that serves as a safeguard, allowing it to be deactivated in case it oversteps human values (as was addressed by Russell (2019)).

Computer engineers and philosophers of AI must also notice that humans who couple with AI, extending our minds into it, will inevitably risk losing control of decision-making processes. We need to argue how it is possible to have AI while not losing control of it. Moreover, as Schneider (2019) pointed out, the proposition of technologies we do not understand and that are deeply coupled with our minds (like AI, neural prostheses, data uploading and neural networks) brings a log of problems and potential risks for all of us (Schneider, 2019). In this sense, Trust and Glue gives directions and boundaries to accomplish this task. The development of AIs presents both limitations and challenges. One

critical issue is that extending our minds into AI systems may come at the cost of losing control over them and our ability to evaluate these coupled systems at any given moment. Therefore, as a general guideline, we should refrain from coupling with such systems until we have a better understanding of AI. It is wise to delay the formation of these cognitive systems until we can address these concerns.

CONCLUSION

In conclusion, our paper has addressed how the extended mind hypothesis brings vitality into philosophical debates surrounding AI by introducing a crucial and novel fourth criterion to the Trust and Glue criteria: *maintainability*. We have presented arguments supporting the fruits of the extended mind concept and emphasizing the significance of the coupled systems from an ExM thesis within the realm of artificial intelligence. Our exploration leads us to conclude that Clark and Chalmers's seminal work (1998) is a pivotal source in addressing contemporary AI concerns, particularly as an additional foundation for objections against the prevailing Standard model of AI.

Within the context of coupled systems integrating AI and human cognition, we underscore the inherent risks posed by the potential emergence of conflicting goals (Bostrom, 2014/2017). Consequently, we argued how Trust and Glue criteria encourage reflections on the needs of a fourth criterion, *maintainability*, over the outer components of a cognitive system that emerges from the interaction of the human mind and artificial intelligence.

The introduction of the maintainability criterion emerges as a crucial safeguard against unforeseen challenges in human-AI interactions posed by AI advances, especially when we see that Big Tech companies, like OpenAI, have disbanded the teams focused on long-term AI risks. Our examination of all four criteria in Trust and Glue (trustworthiness, reliability, accessibility, plus our own the maintainability) seeks to solidify the value of the applicability of ExM to the persisting challenges of AI, where the influence of the Standard model view of AI remains relevant in both research and practical applications.

Thus, in summary, our analysis emphasizes the critical role of ExM in guiding ethical and functional dimensions of human-AI interactions, offering a valuable roadmap for navigating the evolving landscape of the integration of human cognition and technology.

REFERENCES

ADAMS, F.; AIZAWA, K. **The bounds of cognition**. Malden, MA: Blackwell, 2008.

ADAMS, F.; AIZAWA, K. The Bounds of Cognition. **Philosophical Psychology**, v. 14, n. 1, p. 43-64, 2001.

- ADOMAVICIUS, G.; BOCKSTEDT, J. C.; CURLEY, S. P.; ZHANG, J. Do recommender systems manipulate consumer preferences? A study of anchoring effects. **Information Systems Research**, v. 24, n. 4, p. 956-975, 2013.
- ADOMAVICIUS, G.; BOCKSTEDT, J. C.; CURLEY, S. P.; ZHANG, J. Effects of online recommendations on consumers' willingness to pay. **Information Systems Research**, v. 29, n. 1, p. 84-102, 2017.
- AIZAWA, K. Extended cognition, trust and glue, and knowledge. *In*: CARTER, J. A.; CARTER, A.; KALLESTRUP, J.; PALERMOS, S. O.; PRITCHARD, D. **Extended epistemology**. Oxford: Oxford University Press, 2018. p. 64-78.
- BOSTROM, N. **Superintelligence: Paths, dangers, strategies**. 3. ed. United Kingdom: Oxford University Press, 2014/2017.
- BRYSON, J. J. Robots should be slaves. *In*: WILKS, Y. **Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues**, 2010. p. 63-74 (Natural Language Processing series, v. 8).
- CARTER, J. A.; KALLESTRUP, J. Extended circularity: A new puzzle for extended cognition. *In*: CARTER, J. A.; KALLESTRUP, J.; PALERMOS, S. O. **Extended epistemology**. Oxford: Oxford University Press, 2018. p. 42-63.
- CARTER, J. A.; CLARK, A.; PALERMOS, S. O. New humans: ethics, trust and the extended mind. *In*: CARTER, J. A.; KALLESTRUP, J.; PALERMOS, S. O. **Extended epistemology**. Oxford University Press Oxford, 2018. p. 331-351.
- CARTER, J. A.; KALLESTRUP, J.; PALERMOS, S. O.; PRITCHARD, D. Varieties of externalism. **Philosophical Issues**, v. 24, n. 1, p. 63-109, 2014.
- CHALMERS, D. J. **The conscious mind: In search of a fundamental theory**: Oxford: Oxford Paperbacks, 1997.
- CHAYKA, K. **Filterworld: how algorithms flattened culture**. New York, NY: Doubleday - Penguin Random, 2024.
- CHRISTIAN, B. **The alignment problem: Machine learning and human values**. New York, NY: WW Norton & Company, 2020.
- CLARK, A. **Being there: Putting brain, body, and world together again**: Cambridge: MIT Press, 1997.
- CLARK, A. Author's Response: Review Symposium on Being There. **Metascience**, v. 7, p. 95-103, 1998.
- CLARK, A. Intrinsic content, active memory and the extended mind. **Analysis**, v. 65, n. 1, p. 1-11, 2005.
- CLARK, A. **Supersizing the mind: Embodiment, action, and cognitive extension**. New York, NY: Oxford University Press, 2008.
- CLARK, A. Coupling, Constitution, and the Cognitive Kind: A Reply to Adams and Aizawa. *In*: MENARY, R. (ed.). **The extended mind**. Cambridge, MA: MIT Press, 2010a. p. 81-99.

- CLARK, A. Memento's revenge: The extended mind, extended. *In*: MENARY, R. (ed.). **The extended mind**. Cambridge, MA: MIT Press, 2010b. p. 43-66.
- CLARK, A.; CHALMERS, D. The extended mind. **Analysis**, v. 58, n. 1, p. 7-19, 1998.
- GALLAGHER, S. The extended mind: state of the question. **The Southern Journal of Philosophy**, v. 56, n. 4, p. 421-447, 2018.
- HA, D. **Reinforcement Learning for Improving Agent Design**, 2018. Available at: <https://designrl.github.io>. Accessed on: Apr-11-24.
- JAMES, W. **Principles of psychology**. NY: New York: Henry Holt and Company, 1890.
- KOSINSKI, M. **Theory of mind may have spontaneously emerged in large language models**, 2023. arXiv preprint arXiv:2302.02083.
- EUROPEAN PARLIAMENT AND COMMITTEES. Civil Law Rules on Robotics. Resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics. 2015/2103(INL).<https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52017IP0051>. Official Journal of the European Union 2017. Accessed on: Apr-11-24.
- O'NEIL, C. **Weapons of math destruction: How big data increases inequality and threatens democracy**. New York: Crown, 2016.
- PALERMOS, S. O. Belief-forming processes, extended. **Review of Philosophy and Psychology**, v. 2, n. 4, p. 741-765, 2011.
- RUPERT, R. D. Challenges to the hypothesis of extended cognition. **The Journal of Philosophy**, v. 101, n. 8, p. 389-428, 2024.
- RUSSELL, S. **Human compatible: Artificial intelligence and the problem of control**: London: Penguin, 2019.
- SCHNEIDER, S. **Artificial you: AI and the future of your mind**: Princeton: Princeton University Press, 2019.
- WIENER, N. Some Moral and Technical Consequences of Automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. **Science**, v. 131, n. 3410, p. 1355-1358, 1960.