



Generative midtended cognition and Artificial Intelligence: *thinging with thinging things*

Xabier E. Barandiaran¹ · Marta Pérez-Verdugo¹

Received: 30 July 2024 / Accepted: 9 February 2025
© The Author(s) 2025

Abstract

This paper introduces the concept of “generative midtended cognition”, that explores the integration of generative AI technologies with human cognitive processes. The term “generative” reflects AI’s ability to iteratively produce structured outputs, while “midtended” captures the potential hybrid (human-AI) nature of the process. It stands between traditional conceptions of *intended* creation, understood as steered or directed from *within*, and *extended* processes that bring exo-biological processes into the creative process. We examine the working of current generative technologies (based on multimodal transformer architectures typical of large language models like ChatGPT) to explain how they can transform human cognitive agency beyond what the conceptual resources of standard theories of extended cognition can capture. We suggest that the type of cognitive activity typical of the coupling between a human and generative technologies is closer (but not equivalent) to social cognition than to classical extended cognitive paradigms. Yet, it deserves a specific treatment. We provide an explicit definition of *generative midtended cognition* in which we treat interventions by AI systems as constitutive of the agent’s intentional creative processes. Furthermore, we distinguish two dimensions of generative hybrid creativity: 1. Width: captures the sensitivity of the context of the generative process (from the single letter to the whole historical and surrounding data), 2. Depth: captures the granularity of iteration loops involved in the process. Generative midtended cognition stands in the middle depth between *conversational* forms of cognition in which complete utterances or creative units are exchanged, and micro-cognitive (e.g. neural) subpersonal processes. Finally, the paper discusses the potential risks and benefits of widespread generative AI adoption, including the challenges of authenticity, generative power asymmetry, and creative boost or atrophy.

Keywords Extended cognition · Generative AI · Creativity · Authorship · Cyborg intentionality · Midtention

Extended author information available on the last page of the article

“I really do think with my pen, because my head often knows nothing about what my hand is writing”.
Wittgenstein (1980, p. 106).

1 Introduction

You are surrounded by colleagues at a conference. You are about to explain your opinion about the French philosopher Gilles Deleuze: “He is very inspiring, but his writing is too...”—you can’t quite find the right word, “... abstruse” says your colleague, “yeah, his writing is too abstruse (thanks)” you continue. That is the word you needed, the one you wanted but could not find. You do in fact hold the opinion that Deleuze is abstruse, you simply could not generate the sentence fluently, and you completed it by accepting the offered suggestion. You made it yours. At the current pace of evolution of generative technologies, it is not unreasonable to suggest a scenario in which similar conversations are increasingly generated (suggested) by AI (instead of your colleagues). How would this be possible? What kind of cognitive process would this be? How are this and parallel scenarios different to any technologically or socially extended cognitive processes we experienced before? How should we characterize them?

Generative technologies, and Large Language Models in particular (systems like ChatGPT, Gemini, Llama, Mixtral, Claude, etc.), are deeply transforming agency. In a recent article (Barandiaran & Almendros, 2024), one of us introduced the concept of *midtention* to characterize the type of hybrid “intentional” agency that can result from a deep integration between human and LLM interaction:

Transformers are also bringing with them a much deeper meaning of extended agency (with deeper dialectical connotations). There is a form of extended agency that LLMs already offer that get more intentionally intimate than any previous known form. In fact, this extensional character is closer to the intentional character of the mind that deserves a proper name: *midtentional*. (...) The enormous complexity and regulatory capacity of the brain-body system (compared to that of the passive materiality of the tool and work environment) is now challenged by an ongoing activity of language automata, which are constantly reading us and writing (for) us. (...) This brings the power of transformer-human interaction closer to a proper cyborg agency, beyond any experience of instrumental, social or intersubjective agency we might have ever encountered before. (Barandiaran & Almendros, 2024, pp. 29–30).

In this paper, we expand, deepen and generalise over this basic intuition beyond text-based LLMs to generative technologies. We develop the notion of “generative *midtentioned* cognition” as a new type of so-called “extended cognition”—that is, processes that are characterised as cognitive and are constituted by factors external to the cognizer’s brain–body. The meaning will unfold along the paper, but we shall advance that we have chosen the term “generative” to mean what, in different circum-

stances, might have been called “creative”, yet devoid of the strong connotations of the term. The term generative has, of course, also been chosen to name the generative AI technologies that have emerged recently (Akhtar, 2024; Jebara, 2012; Murugesan & Cherukuri, 2023). Altogether, we want to stress processes that produce—generate—structured material outcomes: text, drawings, sound, voice, shapes, etc. Midtended or midtention is a neologism that wants to capture a space situated between traditional conceptions of *intention* or *intended* creation, that is, generated from within, and *extended*, processes that bring material exo-biological processes into the creative process.

In the next section, we introduce so-called “generative AI technologies” and their internal workings. Then, we argue that existing theories of cognition that have incorporated external or environmental components into cognitive processing fall short of adequately capturing the new forms of cognition and agency that generative AI makes possible. Section 3 introduces the concept of generative midtention with the examples of drawing and writing. We then articulate the relationship between the concepts of intention and extension and characterise the singularity of midtended cognition. We provide an explicit definition of midtended cognition and distinguish two fundamental dimensions along which generative midtended cognition can be demarcated. Section 4 introduces some future scenarios that are relevant to deeper senses of generative midtended cognition, we evaluate some potential benefits and risks of authenticity, generative power asymmetry, and creative atrophy and alienation. Finally, Sect. 5 recapitulates on the main ideas of the paper and offers some concluding remarks.

2 Extending extended cognition: generative AI’s new challenges

2.1 The era of generative AI

Generative Artificial Intelligence has reached maturity and widespread adoption after transformer architectures first managed to deliver highly proficient text generation (Vaswani et al., 2017). Today it is possible for anyone to access AI services that can generate text, image, audio and video, and whose quality is often indistinguishable from that created by humans. To characterise these systems, it is important to understand in some detail their internal workings (see Barandiaran & Almendros, 2024 for a more detailed philosophical analysis).

Generative systems take as input a string or matrix of data (a text document, audio file, image, or combinations of them) and break them down into tokens or basic processing units. The tokenized input is then transposed to an intermediate structured “representation” (*embedding* in the case of language or *latent space* in the case of image diffusion models) that takes the shape of a matrix in a high dimensional space. Attention mechanisms make it possible for points of this high dimensional vector to be related to each other in multiply structured manners (grammatical relationships between subjects and verbs in a sentence, instruction steps or a receipt, continuity of shape in a picture). Neural networks take over now. They process the matrix by transforming it further through massive non-linear distributed computations that are

said to express “knowledge” or to constitute a model of the training data (see below). Attentional and neural processing blocks (96 in the case of GPT3, see Brown et al., 2020) are chained in this processing line. The resulting matrix is one that assigns a weighted probability to each possible token. An algorithm then picks up one among the highest probability output tokens. The whole process is mediated by a processing architecture composed of billions of parameters.

Importantly, the generation of a complete output (paragraph, full text, complete image, etc.) is not something created in “one shot” but recursively, token by token, pixel by pixel, by feeding the system with its own output (where the human agent can intervene, by correcting, suggesting, or adding new tokens to the input). Whereas LLMs’ *autoregressive* dynamics rely on reintroducing the output token into the new input (adding the new token to the previous input), generative diffusion models work on a kind of internalised noisy sketch that gets iteratively completed by a similar process of modifying pixels while processing previous image and text (or even sound) input (like image instruction).

In its raw initial state, a generative AI system is almost a blank slate, waiting to be shaped by the training data to determine embedding, attentional and neural parameters. The basic training procedure is to provide the system with an input, to let the system process it, and then to compare the result with the next token (word) of the training data. If the training data includes Clark and Chalmer’s “The extended mind” paper, the training process will include delivering as an input pieces of the text, like the beginning of the paper “Where does the mind stop and the rest of the world [...]”, and letting the system “guess” the next word. During training, the result will range from absolutely random at the beginning, (e.g. “Paris”) to an approximate bad guess in earlier stages, (e.g. “rest”), to a reasonable match (e.g. “start”). The output is then compared to the next word on the original text: “begin”. The “distance” between the generated and the correct output is used to train the system backwards, adjusting all the parameters to reduce the error. Therefore, “Paris” will generate a big error, but “start” a small one.¹

In this sense, the system is trained to predict the most likely follow-up sequence of a given input (often called prompt). Sextillion operations of error backpropagation and parameter fitting have been shown to deliver very efficient systems able to reasonably approximate training data. But, most importantly, able to generate reasonable completions to previously unseen data. To the input “Deleuze is very inspiring, but his writing is too” (never seen in the training phase), ChatGPT responds with “dense and abstract for some readers.”² Furthermore, foundation models (as they are called after the first intensive training phase) can be fine-tuned for specific tasks or to adapt their generic abilities to specific content, styles, or different types of customizations. It is possible to tune ChatGPT (and other LLMs) with personal conversation records, emails, messaging conversations and papers to fit my personality and complete the above sentence with “abstruse”. In addition, the input to a generative system can

¹ The real procedure is more complicated (e.g. all output’s error is 1 except for the “correct” term) but the net effect of the error propagation will be smaller for terms that are closer in the embedding space (e.g. “start” and “begin”).

² <https://chatgpt.com/share/8a56c0cb-95fc-48ed-aa01-b8d2ec6adac4>.

include contextual information (like the set of papers you want to reference in a paper you are about to start writing, or all the emails you exchanged with a particular person, or a full genocide-victim database).³

Being the fastest growing digital service in history (Hu & Hu, 2023), ChatGPT and similar generative technologies producing text (Gemini, LLaMA, Mixtral, Claude, etc.), images (DALL·E, Midjourney, Stable Diffusion), music (Udio, AIVA, WavTool) or video (Sora, Pika Labs, or Runway) are part of our cognitive resources. Moreover, generative AI services, increasingly involve the disposal of joint human–machine creative digital spaces (Canvas on ChatGPT, NotebookLM on Gemini, Artefacts on Claude or NotebookLLaMA) and complementarily, digital productive environments (from Gdocs and Gmail to the Microsoft or Adobe suites) incorporate generative AI support for a variety of creative tasks (from email writing to image editing). What are our theoretical resources to accommodate these tools and environments to our best understanding of how human cognition emerges out of recurrent agent–environment interactions?

2.2 Extended cognition theories and the challenge of generative environments

Many philosophical accounts, particularly during the twentieth century, have emphasised the role of the technical environment in human cognition and experience. Notable examples include Haraway’s cyborgs and situated knowledge (1991), Stiegler’s Heideggerian account of technique (1998) and Hutchins’ distributed cognition approach (1995). However, the extended mind hypothesis (Clark & Chalmers, 1998) marks the first significant attempt, within philosophy of mind to strongly argue on the triviality of retaining the mind, within the boundaries of the skull. This hypothesis, rooted in a functionalist philosophy of mind, coined the parity principle as the main support for its argument: if a process considered cognitive when performed by the brain can also be instantiated including elements of the environment, then those elements are cognitive to the same extent as the brain.

Clark and Chalmers illustrate this with the example of Inga, a healthy individual, and Otto, an Alzheimer’s patient. Both individuals successfully perform the same cognitive process (remembering the address of their favourite museum), but in each case it is instantiated differently—Inga using only her brain and Otto using his brain and a notebook. Otto has an extended mind. To clarify this claim, Clark & Chalmers propose some criteria that environmental resources must meet to qualify as cognitive extensions: trustworthiness, reliability and accessibility. Known as the “trust & glue” criteria, they seem to follow the intuition that the less controversial cases of extension are prostheses or bodily extensions; fixed and static environmental resources that the agent always carries with herself.⁴

³Generative AI systems like Gemini can handle up to 2 million token contexts and be accurately sensitive to their content when generating the output.

⁴This connects to earlier phenomenological intuitions explored by Merleau-Ponty or Heidegger, such as the blind man’s cane as a transparent extension of his embodied cognitive abilities (Merleau-Ponty, 1945).

After a first wave focused on defending the possibility of extended minds, a second wave of theorists moved away from strict functionalism and the stringent parity principle (Sutton, 2010). Instead, they adopted complementarity-based accounts, where external resources integrated in cognitive processes need not mirror brain functions exactly, fostering different kinds of cognitive extension based on varying dimensions of integration. Heersmink (2015) offers a taxonomy of these dimensions as presented in the literature, retaining some of the early intuitions of durability, reliability and trust, reformulating others such as phenomenological and informational transparency, and including newly proposed dimensions such as informational flow, individualization and transformation.⁵ This allows a more nuanced and fruitful description of extended cognition.

But it is still noteworthy for our purposes that the most used examples of extended cognitive processes in early theorising were memory processes, where it is particularly relevant that the environmental resources remain fixed and stable, thus making a certain passivity of environmental elements the ideally characterised integrators of extended cognitive processes (in open contrast with generative AI). In this sense, Material Engagement Theory (MET) (Malafouris, 2016) as a follow-up of extended cognition theories that takes a more enactive and dynamical approach to cognition, is clearer in the non-passive role that it assigns to the environment. MET not only claims that cognition is extended beyond the skull, but that it is dynamically co-shaped by material interactions. Thinking is not previous to the encounter with matter. Thinking is *thinging*, doing things with matter. Malafouris (2019) uses the example of pottery (note the difference in the choice of example, opting here for an uncontroversially creative and active process). Contrary to the idea of the potter imposing a predefined form onto passive matter (as if having a 3D model that is then applied to the clay), the form starts to emerge in the process of the potter manually interacting with the materiality of the clay. The malleability of the clay in response to hand movements is what constitutes the creative process. Whereas in traditional theorising of mind the generative cognitive locus was “in the head” of the human agent, here the generative locus of cognition is situated in the meeting with the environment, driven both by material constraints and human creative agency.

Extended cognition theories, as part of the 4E paradigm in cognitive science, have proved to be rich frameworks to study contemporary digital technology (Smart et al., 2017). AI-based technologies, however, pose particular challenges, and ethical debates have questioned the suitability of adopting extended cognition frameworks for these contemporary technologies (Cassinadri, 2022; Farina & Lavazza, 2022). The challenge stems from the intuition that extended cognitive processes should retain a certain locus of control or authorship within the human agent, so we can still consider it *her* cognitive process. We could make use of trust & glue criteria and dimensions of integration to explore these issues of control, but they turn out to be insufficient in

⁵ Sterelny (2010) also focuses on the ideas of individualization and entrenchment, as opposed to interchangeability. These concepts highlight the intuition that an important aspect of integration is the tailoring of the artifact to our specific cognitive needs so that our cognitive processes become more efficient. These dimensions have remained relatively undertheorized but stress the relevance of adapting the environment to the agent’s specific cognitive functions—and, we could say, style—, which we will argue becomes an important area to explore in the case of generative AI.

many cases, since they operate under the assumption that the environment does not have any sort of generative power. Under this supposition, it makes sense to think, for example, that the more transparently we use an artifact, the more integrated it is in a cognitive process: more transparency would mark that the artifact is more directly *responding* to the users' intentional agency (and only to it), since the human agent is the only locus of generative, purposeful interventions in the interaction.

But the intuition that more transparency equals more integration (in this sense of clear authorship of behaviour) starts to break down when we consider AI-powered technologies (Clowes, 2020; Wheeler, 2019). One of us (Pérez-Verdugo, 2022) analyzed this problem with the example of a smart racket that adjusts your movements to optimise your play. Here, we can have a very strong phenomenological feel of transparency, since we can enter an uninterrupted flow of action. Nevertheless, the artifact is not directly responding to our intentional agency, but is actively anticipating it and *intervening* on the joint action at a scale the agent is not fully aware of. The artifact constitutes its own locus of generative, purpose-*structured* interventions in the cognitive process that involves both agent and artifact. The generative locus of cognition is now no longer only on the side of the agent in its meeting with materiality. There is now a second locus in the environment that can generatively alter the materiality of the encounter.

The important novelty of Generative AI is that, on the pottery analogy of material embodied cognition, clay is starting to shape itself by anticipating our movements. And it does so based on the huge amounts of data coming from: (a) the history of pottery and all contemporary catalogues, (b) the current shape being moulded, (c) our prior declared intentions, (d) the particular work context and, (e) our personal-style expressed on all our previous production. And we can engage with it in ways that range from the AI *inter-actively* grasping and directing our hands, to simply letting us rest at a distance to see the clay take shape. Thinking (cognition) might not be anymore a process understandable as “mere” *thinging*. Our computational civilization has created *thinging things*, that is, things (programs) that recursively (autoregressively) transform digitality (as a type of materiality) and generate other things (structured digital objects, including other programs). We live, thus, in an era where cognition can be extended in a very specific form, as a very special form of thinking: *thinging with thinging things*.

3 Generative midtention

3.1 Intention and authorship in creative action: drawing and writing

Imagine you are a professional designer. You received an email with a client's request. You externalise your thoughts: “This work requires an 80's style drawing with a contrasting combination of sharp and round shapes”. Creating this work implies many recursive processes: from looking at the blank piece of paper or screen and projecting imagined figures, lines, shapes and compositions (maybe even using hands or other objects to get a grasp of proportions), to sketching in pencil the imagined future drawing, to inking it (making it more definitive) and to finally colouring it. So, as *you* start

drawing, *you* get to imagine the completion of your traces on the white background as a grey coloured (pencil) sketch. Sometimes *you* augment the focus of your anticipated sketch and keep imagining. Some other times, *you* directly intervene on it: *you* either keep drawing following the grey lines that *you* drafted and continue imagining them as *you* draw them. But you also often choose to diverge from the draft, and *you* imagine how the picture re-organizes accordingly. Sometimes you surprise yourself as *you* have sketched, on *your* first try, exactly what you wanted. Some other times, you take your distance, *you* sketch again and keep correcting until *you* get it right.

We have all had similar extended creative experiences before. But with generative AI, the role of imagination and of the material pencil sketch (or even that of the “inker”), can be played by a machine. By a generative AI working in real-time *within* our creative process. Read the passage above again and substitute the “you” in italics by a generative AI. Although still relatively fictional, this possibility is increasingly foreseeable. It is partly a reality for computer programmers making intensive use of code completion. And it is partly available as autocompletion and other generative AI services embedded within office environments or audiovisual creative and editing tools. We explained above how these systems can be trained, tuned and contextually fitted to a particular user, her past work and surrounding information. The system could have read the client’s request, taken the designer’s expressed intentions, and projected the style and features of her past designs to this joint generative process; while actively adapting to the designer’s intentions in real-time.

In a similar vein, Barandiaran and Almendros (2024) provide the following explanation of writing as a generative process:

The process of writing (in paper or on the screen) is one that it is often experienced as the very act of writing driving itself the intentions of the writer: the interaction process of writing pulls agency out of the head. It is the recurrent hand-keyboard-PC-screen-vision-brain-body-hand loop that produces text. Yet, we don’t only write. We also supervise and edit recurrently. Thus, at least two loops are involved here, one is more pulled by the direct writer-text dynamics, the other by the more detached editorial supervision that either continuously or intermittently follows the former. At times, one finds the non consciously written text as right and owned, as proper, and it is left untouched. Other times... “That is not what I meant exactly” ensues, “it needs a rewrite”. Both loops are person anchored. The environment (pen and paper or keyboard and screen) served as a support structure, a well integrated, creative scaffold, providing the material basis of extended memory, recomposition, and tinkering. But the writer was the extended agent, the organized center of the scaffolded subject. This might start to change. (p. 29).

A variation of Hemingway’s motto “write drunk, edit sober” is becoming widespread in many areas of human text generation: “let the LLM write, edit yourself” and very often the inverse “write drunk, let the LLM edit”. Both loops used to be person anchored, but now LLMs can intervene on any or both of them and create, alter or shape (parts of) the final creative product.

These are all examples of generative activities. We shall use the term “generative” as a more generic or neutral concept (regarding mindful attribution) for what, under normal circumstances and performed by a human, would be considered a *creative* production process. Unlike other forms of cognition (like decision-making, recalling, navigating, imagining, judging, understanding⁶) generative cognition is characterised by the generation of a sequence of transformations on the environment to compose a final product. Under most circumstances, generative processes are creative, purposeful processes. Intentions take shape along the process and the final result is authored. You identify yourself with the result. When honest, a sense of authorship ensues, like the sense of agency that accompanies intentional and voluntary action. But, with generative AI, as we have seen, the environment becomes populated and shaped by digital automata capable of producing (as Barandiaran and Almendros put it) purpose-structured (although not fully purposeful) generative processes.

What happens when the generative process is hybrid? When machine-generated units (tokens in LLM technical vocabulary) are systematically combined with human-generated ones? When production and edition, sketching and inking, are recurrently intertwined between machines and humans? What happens when machine contributions are tuned to context and author-style and knowledge? When the machine inputs augment, condition and expand authorship?

3.2 Intention, extension and midtention

For classical approaches to human action, intentions are the result of the right combination of beliefs and desires causing the action (Davidson, 1980). This kind of event-causal account of intentionality leaves little room for explaining generative midtended action. ϕ becomes an intentional action if X (the agent) holds the belief that doing ϕ will satisfy desire D. Both ϕ and D are specified prior to the action and the desire D and the belief that “ ϕ will bring D” must be contained and causally efficacious in the mind of X in order to make ϕ an intentional action.

Contrary to the standard view, enactive and embodied contemporary approaches to cognition (Calvo & Gomila, 2008; Clark, 1998; Gallagher, 2023; Varela et al., 1991) stress the emergent nature of intentionality through sensory-motor interactions between agents and their environments. There are various families of such approaches. On the active inference or predictive coding side (Clark, 2013), intentionality is understood as emergent from recurrent predictive loops, where action is directed at reducing uncertainty and updating the agent’s future predictions of its environment. As Barandiaran and Almendros stress, from this point of view, the novelty of generative technologies is that: “we might be encountering, for the first time, that the environment is delivering to the brain-body the very predictions that the brain-body is about to make about the effect of its own activity on the environment (...) In a textualized manner, this form of autocomplete is equivalent to injecting predictive efferent signals into the body movement.” (Barandiaran & Almendros, 2024, p. 30).

⁶Although it is possible to do all these things generatively, they are not defined by their generativity.

Enactive, autonomy-centred sensorimotor approaches, on the other hand, stress the importance of the agent-centred adaptive (i.e. norm-sensitive, purposeful) modulation of the sensorimotor coupling with the environment (Barandiaran, 2008, 2017; Barandiaran et al., 2009; Di Paolo et al., 2017). This approach assumes that the sensory-motor coupling is modulated from within or centred on the agent's side. How to make sense of the increasing active automaticity of the environment, producing normatively-structured changes in the environment while coupled to the activity of the agent? As the word-suggesting example in the opening of the paper illustrates, generative cognition might be better understood through the conceptual resources of social cognition (Pérez & Gomila, 2021) than those of extended cognition. In the enactive approach, this would imply conceptualising generative cognition as a case of participatory sense-making (Cuffari et al., 2015; De Jaegher & Di Paolo, 2007). However, generative technologies do not qualify as autonomous agents (Barandiaran & Almendros, 2024; Gubelmann, 2024; Harvey, 2024). Thus, it is problematic to apply social cognition categories to interactions with and between generative AI systems, if we make justice to their lack of genuine autonomy, agency and understanding.

Technology, in turn, has also started to be analysed within enactive theorising. In a recent paper (Pérez-Verdugo & Barandiaran, 2023), we define technical behaviour as the active transformation and organisation of elements of the body and environments “with the intended effect of constraining or regulating couplings with (or between) other aspects of the environment” (p. 83), and technology is seen in this framework as the “sedimented effects of this technical behaviour at different scales within a systemic context” (p. 84). This does indeed underline the strong relation between agency, productive practices (see also Di Paolo et al., 2023) and technical interventions, and the enactive idea of coupling stresses the co-constitutive and constraining role of the environment in cognitive processes. However, this account still retains the idea that the coupling with the (technological) environment is only *generatively* regulated by the agent -which is the only one capable of technical behaviour in the coupling. Enactivism, thus, has not yet accommodated scenarios in which the environment is *generatively* norm-following or purpose-structured and tuned to the autonomy of the human agents and the sociotechnical context in which they operate.

Following Barandiaran and Almendros (2024), we differentiate between the *purposeful* (or intentional) behaviour characteristic of human agents, and the merely *purpose-structured* and *purpose-bounded* behaviour that generative AI produces. The gigantic amounts of data that are used to train generative AI are the result of human purposeful activity and the expression of human purpose. Massive iterative parameter tuning techniques result in AI systems becoming purpose-structured themselves *and* able to generate purpose-structured output. Moreover, the training process with reinforcement learning (often with human feedback) channels the resulting behaviour within the bounds of human expectations, desires, goals and values, constraining its generative capacities within purpose-bounded outputs.

To further clarify this point, we can make use of the distinction between intrinsic and derived intentionality (Searle, 1983). Generative AIs are not intrinsically intentional systems, but they possess derived-intentionality, like other human creations (from simple utterances to complex technologies). However, unlike previous signs and technologies, this time the derived intentionality is also *generative of derived*

intentional products. Imagine a set of road signs in a city; they are certainly not intrinsically intentional entities, not purposeful systems. The STOP signs do not understand English, nor do they “know” that they are there to stop cars, or even to signify anything. However, they are all purpose-structured and purposefully-arranged, they operate as systems with derived (human designed and repaired) intentions to avoid car crashes, to reduce car speed and pollution, to direct the traffic towards the widest avenues, and to reduce it in school areas. Now imagine that, taking a set of initial prescriptions and based on immense road traffic data, an AI is capable of generating a similar kind of purpose-structured traffic-sign distribution. We have here a second-order derived-intentionality system (the generative AI), capable of generating third order derived-intentional or purpose-structured products (texts, images, etc.).

The framework we are developing in this paper allows for the integration of both social-cognitive aspects and technological constitutive aspects into a unified theory (for both enactivism and predictive coding approaches). A proper characterization of the cognitive processes that occur in interaction with generative AI is required for this task. As a first step, we provide the following definition of generative midtention:

Given a cognitive agent X, a generative system Y (artificial or otherwise) and cognitive product Z, midtention takes place when generative interventions produced by Y become constitutive of the intentional generation of Z by X, whereby X keeps some sense of agency or authorship over Z.

Figure 1 illustrates the definition and the various ways in which grounded, embodied and extended cognitive processes interrelate with each other and with generative processes on enacting midtended cognition. Agent X, interacts with digital product Z, which is partly generated by Y, an artificial generative technology.

Different agent-environment loops generate different aspects of cognition: *Grounded cognition* implies neurodynamic loops involving multimodal pre-motor and sensory processing areas, *embodied cognition* implies body proprioceptive feedback and neuromuscular synergies, *enactive cognition* involves the coupling with environmental elements, and *generative AI* (not a kind of cognition per se) produces digital structures through autoregressive loops. *Generative midtended cognition* brings all components together on the production of somewhat intentional, purposeful, works.

Creative intentional loops through the agent (A loop) are nested with the autoregressive loops of the generative system (B loop). Mere concurrence of both loops separately, however, is not sufficient for midtended cognition; both loops need to interact, in the sense that generative interventions by the B loop become nested with the A loop without the latter losing its intentional character. The agent needs to *accept* or *adopt* B loop’s interventions as part of her intentional creative process. Then, the whole system becomes a case of generative *midtended* cognition.

Why and how are AI-generated tokens or interventions adopted or midtended by the agent? At a first approximation, taken in the context of the examples of word suggestion or text completion, the adoption can be attributed to the predictive power of the generation. The generated word(s) are those that the agent was about to use or would have used given other circumstances (a clearer mind, more time to recall, etc.). But these circumstances can be amplified further. Generative interventions might be adopted, or midtended, if the agent finds the generatively suggested information as

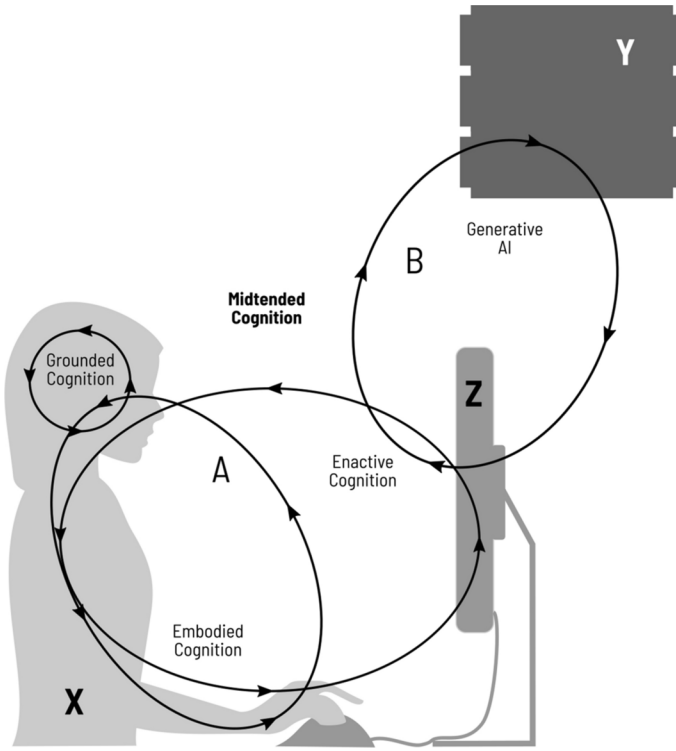


Fig. 1 Different agent-environment loops generate different aspects of cognition. Midtended cognition involves the intersection between more standard agent anchored A-type loops and AI produced B-type loops. See text body for details. [Copyleft 2024 © Barandiaran and Pérez-Verdugo, with CC-by-sa licence, the person’s silhouette and the server image are derivative works of original SVG files of Vectorportal.com, with permission]

a good approximation of what she could have found herself by doing some research or, and here we enter a complicated ground, had she read and understood some additional theory or author (e.g. the work of Gilles Deleuze). In fact, generative tokens can be midtended or made your “own”, by cognitive proximity, comfortable familiarity and self-identification, or simply by trust, laziness or fear that one’s own standards are lower than those produced by the AI system.⁷

3.3 Dimensions of generative midtended cognition

Inspired by Heersmink’s (2015) dimensions of integration in extended cognition, we propose that cognitive processes with intersecting loops of generative power can also be conceptualised through dimensions of (active) integration. Note that whereas Heersmink’s (or any traditional extended cognition literature) account refers to the degree of integration of external vehicles in a single locus of generative cognition

⁷And here is where hallucinations, biases and reasoning errors are still preventing a generalised rolling down into the generative road.

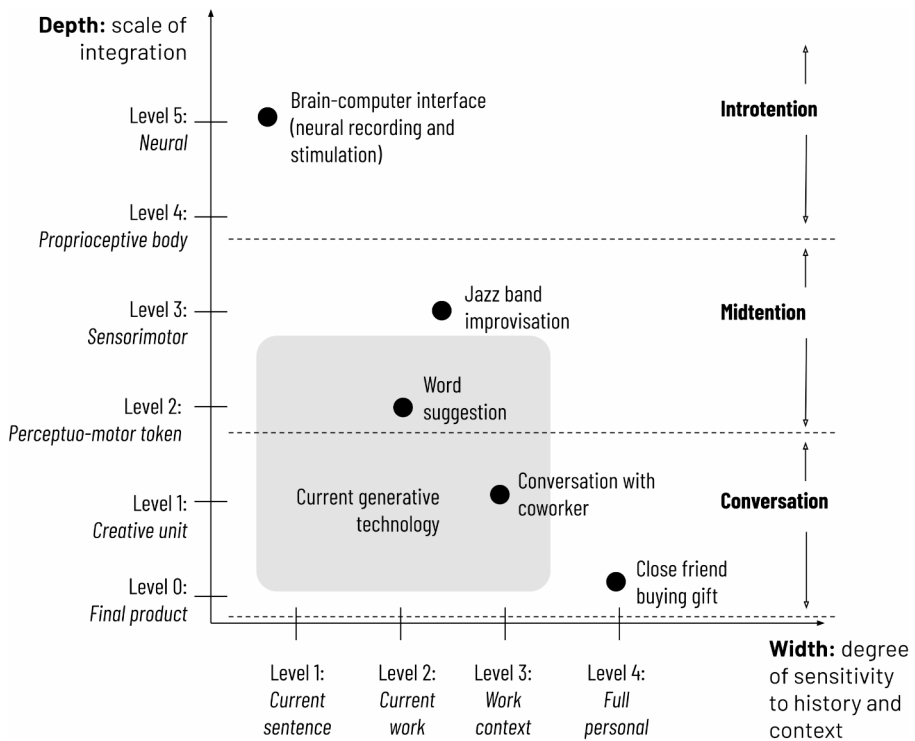


Fig. 2 Two main dimensions of generative hybrid space: depth and width. The first refers to the degree of reciprocal causal integration. The second refers to the length of the context in which Generative AI is trained in relation to the agent’s history. [Copyleft 2024 © Barandiaran & Pérez-Verdugo, with CC by-sa licence, the person’s silhouette and the server image are derivative works of original SVG files of Vectorportal.com, with permission]

that is anchored in the agent, we are proposing a different sort of (active⁸ or generative) integration; that of at least two different loci of generative power—each of which potentially realised by agent or environmental resources or both—jointly bringing about a cognitive process. We hereby propose a conceptualization along two dimensions (see Fig. 2)—although it could potentially be extended to include more (see footnote 6).

The horizontal dimension, labelled *Width*, involves the degree of generative sensitivity to the increasingly wider data samples on which the Generative automaton can be trained, tuned or can take as context. We assume that the raw foundational systems have been trained on a massive dataset that includes a huge sample of human-gener-

⁸We tentatively term the kind of integration we are targeting “active” integration, instead of the more passive kind of integration theorised in extended cognition so far. We are aware, however, that this might be misleading as, in a different context, extended cognition was precisely characterised as a sort of “active externalism” against the “passive externalism” of externalist semantics (Hurley, 2010). We have chosen to maintain it like that for now, given that the “active” etiquette there is used in a sufficiently different comparison and terrain.

ated digital content (text, video, audio, image, etc.). Beyond that, it can be increasingly tuned to adapt to the agent's personal history and context.⁹

Academic writing can serve as an example to understand what we mean by this horizontal axis. Level 0 would correspond to the system's sensitivity covering only the last word typed. Level 1 would include the whole previous sentence, and level 2 the whole text that is being written (e.g. the whole paper including a first draft of the abstract as sent to a previous conference, and some notes, quotes and bulleted ideas to be fully developed). Level 3 would involve contextual access beyond the current work, including, e.g., the documents you might have selected to re-read and cite on the paper, the editorial guidelines of the journal, your teaching notes on the subject, etc. Level 4 includes everything you have ever written and read (or even from multi-modal sources: heard, seen, etc.).

The second dimension, labelled *Depth*, refers to the scale or granularity of integration of reciprocal interaction of generative loops. At level 0 we have a final product (a text, a song, picture, etc.) that might result from a generative or creative process: e.g. a text, a song, a picture, etc. Level 1 involves conceptual or creative units: e.g. a sentence or paragraph or a musical phrase or rhythm. Level 2 involves perception–action tokens: e.g. a word, a note. Level 3 involves sensorimotor dynamics, generally sub-personal or unconscious: e.g. a letter within a text, which we are rarely aware of, neither in reading nor in writing. If we continue further up within this dimension, we can get to level 4, the proprioceptive body, where we have to turn to bodily prostheses as an example. Finally, level 5 would include integration at the neural scale—potentially achieved by Brain–Computer Interaction technologies.

This allows us to distinguish three levels of integration along the depth dimension, one of which targets midtended cognition as we have defined it. In lower levels along this dimension, where integration occurs at the level of the final product or creative units, we have *conversation*, the kind of integration we usually think of when we think of social interactions. A conversation with colleagues or a co-authored paper where each author writes a section and then revises other people's sections would fall within this level. Many of our current interactions with Generative AI technologies also fall here: e.g., when we give a prompt to an image generator, and seconds later, without further intervention, it provides an image close to what we requested, or when we enter into a conversation with ChatGPT. And, as we can see in the figure, the width dimension nuances the kind of integration in these conversational cases: there is a clear difference between a conversation with a co-worker and a conversation with a life-long friend, or between your landlord choosing a new wall paint for your flat after you asked for renovations and your mum gifting you some (requested) decoration for your new home.

On the other end, the deepest level, *introtention*, seems to (luckily) be still quite out of reach, at least for the public. Although progress has been made in Brain-Computer Interface technologies, and marketing and hype surround it (see Neuralink

⁹We have chosen to mix both in a single dimension for simplicity's sake. However, we could be more precise and differentiate between two dimensions, keeping the label "Width" for the dimension covering the system's sensitivity to the particular context in which the generative process is situated, and adding a second dimension labelled "Length" that covers the system's sensibility to the specific agent's previous history of creative processes and outputs, displaying a temporal character.

as an example), we have yet to see integration at this deeper scale to be adopted in somewhat standard contexts. Our analysis, however, helps forecast its potential risks and offers grounds for its future theorising.

But we are more interested in the level in between both; the kind of integration that occurs at scales deeper than complete creative units (an email, an image, a song), up until the sensorimotor level. The possibility of this occurring with generative technologies is, we argue, what remains within the blind spot of previous theorising of extended or social cognition. As with the conversational level, this does not imply that midtention only occurs with generative AI. As the participatory sense-making framework has shown, integration with other agents also occurs at this deeper scale, and many human intentional processes are, in a sense, socially midtended. A well-documented example of this is the integration between band members in jazz improvisation (which also becomes clearly modulated along the width dimension) (Martínez et al., 2017). The central claim we make is that productive processes where generative AI is working *within* someone's creative process, would also fall in a similar category. They are cases of *generative midtended cognition*.

4 Should we stand on the shoulders of generative giants? Future risks and benefits

Before assessing the future of Generative AI, we need to take into account that it is a technology that has grown very fast, but that might be showing signs of fatigue and plateauing. The bending of the exponential curve is evident for many. It is perfectly possible that, as a species of technological entity, LLMs and the like might have reached important limits on their generative capacities. If that is so, there are different scenarios that can ensue: 1. this is the peak of AI fantasy, and we are deemed to conform with what we have, or 2. other infrastructural or algorithmic innovation might lead to increased generative capacities. If the first scenario is right, generative midtention might not expand very far.

The state of the current technology does not afford a comfortable hybrid generativity. The current problems with hallucinations, reasoning limitations and deviant chained errors make midtended cognition a relevant but narrowed phenomenon. In fact, we are aware that in our analysis we have conceptualised the “width” dimension as assuming a high precision and accuracy. This is definitely an idealisation; the predictive capacity of different technologies might vary, due to technical shortcomings, but also to being inaccurate *by design* (see later discussions about branded content).

However, if the technology keeps evolving, we might increasingly shift to a cognitive digital culture that relies on generativity as much as previous civilizations were partially born and sustained by the cognitive transformations that writing brought about (Ong & Hartley, 2012). Midtended cognition (and other forms of generative interactions, closer to those found in functionally equivalent social cognitive tasks) will spread quickly and become an integral part of our producing and being in the world.

Technology can evolve in the two dimensions of generativity we outlined above, towards more granularity on the recursive technological interventions, and on the

increasingly wider contextual tuning or adaptation. Regarding the first dimension, if cognition emerges out of recursive loops and, so far, generative technologies only had access to relatively low-resolution external tokens (words, pixels, etc.) the potential of training, prompting and contextualising the functioning of generative AI with neuromuscular, proprioceptive, and brain data might be enormous. The risks of intervening and that scale would be even bigger.

Perhaps the most promising evolution of the current technology involves the hybridization of social cognition with generative AI along the horizontal dimension (including others in the interactive and generative context). LLMs can facilitate collective intelligence, boosting midtended agency in the direction of genuine participatory cognition. But this also poses a delicate risk. Social interaction is to social autonomy what neuronal interaction is to individual autonomy. Technologies directed to actively, automatically and massively intervene in the social interaction fabric, come with high risks. Existing biases and limitations might well export and be amplified in the social realm, but the implications go deeper.

Generative AI, when considered from a sociotechnical perspective (see also Hipólito et al., 2023), moves beyond the cases of individual creativity we have explored so far; and can become an integral component of social and political cognition. It is starting to play an increasing role in mediating and intermediating social interactions. First, by occupying positions traditionally held by human mediators in social coordination—e.g. as a family therapist mediating in a couple's conflict, a secretary organizing and scheduling meetings, or a lawyer mediating between a client and a public institution. Generative AI can also intermediate between human or machine mediators: e.g. automatically providing meeting schedules for a secretary, bringing up relevant laws for the lawyer or identifying significant events or *lapses* to the therapist. More simply, when AI breeds its generative capacity with you to reply to an email, it is already intervening in social interaction. This does not only entail a contribution to the information packages we exchange between humans (emails, lawsuits, confessions or meeting schedules). By selectively ignoring some aspects of a conversation or social interaction, by amplifying others, by invigorating politeness or demanding concreteness, by accelerating response times or reinforcing standardized phrases, generative midtention intervenes on the participatory cognitive interactions that generate and sustain social norms and regulate social coordination.

Shared intentionality is key to understanding how humans align their goals and actions to engage in cooperative behaviour (Tomasello et al., 2005). Generative AI can support these mechanisms by facilitating joint goals and promoting shared understanding; e.g. merging individual and shared context on midtended generative processes. Moreover, if explicit intentions serve as guiding frameworks that direct behaviour and enable effective cooperation (Bratman, 1987), midtended cognition is called to play an increasing role in planning, proposing and adjusting intentions and actions, which allow for smoother alignment of individual contributions within collective efforts. This generative support for social adaptability might enhance the fluidity and coherence of social coordination.¹⁰

¹⁰But it might also weaken the stability-inducing effect that non-AI-reshaped prior intentions had toward future plans, by increasing resources for practical reasoning and intentional reconsideration and leaving

Digital infrastructures are an essential part of contemporary political processes, and they are especially useful for accelerating and scaling participatory democracy (Barandiaran et al., 2024; Berg & Hofmann, 2021). AI and Artificial Life systems can play an important role in empowering these infrastructures (avoiding the dangers of digital authoritarianism and centralism) towards models of democratic ecology augmented by (open and participatory) artificial systems of care and facilitation (Barandiaran, 2019). Generative AI can act as a democratic catalyst, occupying roles traditionally reserved for human intermediaries—breaking information bubbles, identifying latent conflicts or facilitating consensus. In fact, generative AI is already, albeit in an experimental way, becoming an infrastructural component of social organization, participating in roles that are integral to democratic deliberative, propositional and decision-making activities actively synthesizing collective inputs and (re) writing citizen proposals, that are further modified and selected by participants, to later bridge the gap between expert input and participatory engagement, and finally delivering public policies (Bjarnason et al., 2024).

The risks increase if we consider that, unlike social conquests like public education and public libraries, the development of generative AI rests mostly on the private sector and demands increasingly higher resources that remain out of reach for anyone other than a few big tech corporate actors. Moreover, we are still at an early stage of generative AI service delivery. As the recent history of social media services has shown, the early stages of corporate success depend on increasing the user-base and their trust, only to later deploy monetization strategies that are mostly directed at commoditizing human behaviour (Zuboff, 2019). The commodification of human intentionality through its generative midtended extension might be one of the deepest risks we face in the near future of AI's sociotechnical expansion. If the *attention economy* is already showing its devastating effect on human autonomy and well-being (Bhargava & Velasquez, 2020; González de la Torre et al., 2024; Williams, 2018), it is hard to imagine the potential negative impact of an *intention economy* deployed through generative human cognitive steering. The attention economy has so far mostly developed by grabbing our attention and influencing future behaviour by shaping or creating desires (e.g. TV advertisement) or by conditioning or facilitating attention-action possibilities (online clicking to buy a product). If human intentionality however is increasingly meshed with midtended generative processes as we have defined it, the way is open for commodifying the generative process.

Going back to the opening example of word suggestion, we can find a parallel example that illustrates the intention economy to come. When the waiter comes, and you are about to complete your order, you hesitate with the wine. “We are having fish, and it is a hot day, how about a fresh white wine, like...” when suddenly you hear “Txakoli Generative”. It is not a suggestion of your partner or the waiters. It is the voice that comes out of the speaker of an AI embedding glasses (that include cameras that capture everything you see and a small speaker to your ears). The IA,

increasing details of planning open to artificial generative coordination. In the same vein as social habits and rituals suffered the destabilizing effect of instantaneous social communication (e.g. submitting social meetings every Wednesday to a constant negotiation in WhatsApp) we might be at the doorsteps of a volatile and continuously automatically mediated and re-negotiated planning under generative AI supported shared intentionality.

however, knows both your tastes and the current context, i.e. the meal you are about to order, as you discussed it with your meal colleagues, and images of the long wine cart that overwhelmed you. Whispering the right choice, at the right time, in the right context, is a way to commodify midtention. This opens the way for an economy of intention. Like the word suggested by a friend and integrated into your discourse about Deleuze, products, places or services might also be recurrently injected into your verbal or interactive intention-generating spaces. Is Txakoli Generative the wine you “really wanted”? Well, you did not fully know, or maybe you did, you tried it in the past and, in fact, it passed unnoticed to you on the wine chart. Had you known it was there, you might have chosen it yourself. But the same might have been the case for the other two wines on the list. Yet, Txakoli Generative paid a premium plan for its promotion on the intention economy market for the AI-embedded glasses you are wearing. More examples can be added to illustrate this tendency. Following the manifestation of generative midtention in writing an academic paper, it is not difficult to think of the generative assistant suggesting references from authors of universities that have paid a marketing package on their subscription.

Beyond this, even if we dodged those risks, one further line of worry that has already been brought up in the literature (Hernández-Orallo & Vold, 2019) is that of a sort of *cognitive atrophy*. Motor vehicles induce movement atrophies that we need to compensate with gyms and sport habits. Calculators severely reduced our capacity to carry out fast and efficient mathematical calculations. What kinds of creative atrophies might result from generative midtended cognition in humans? We also need to take into account that not only humans are exposed to creative damage. Generative AI can also collapse when recursively trained and tuned to cannibalised content (Shumailov et al., 2024). Could a generalized adoption of generative AI result in a global creative atrophy, incapable of generating sufficiently rich and original content to feed itself forward?

Finally, issues of authenticity (Zawadzki & Adameczyk, 2021), which are already at stake within current digital contexts (see Pérez-Verdugo & Barandiaran, 2023, for an exploration within an enactive framework), might also get amplified with generative midtention. If personal identity is the result of our creative and narrative capacity to write and tell ourselves in the world (Ricoeur, 1995), and generative interventions occur at scales that usually fall outside our awareness by technologies that we are unable to regulate (since they are mostly opaque for us), our personal autonomy and identity will be at stake. While the user can retain a sense of authorship over the final product, as she is the only true agent, midtended cognition implies that this does not guarantee that it is a completely authentic product of her agency in the traditional sense. To assess the extent of authenticity, the width dimension could be called in to shed light on some nuances: even when integration is deep, we cannot ascribe the same degree of authenticity to a wider midtended process where generative AI is tuned to the level of full personal context of the user, than to shallower cases where the tuning is more local. This is also the intuition that the individualization dimension of integration in extended cognition was tapping upon.

With generative AI, however, this can be misleading. While in the dimensions of individualization of extended cognition the tuning is assumed to be made and regulated by the human agent, here the tuning relies on the generative AI itself -or rather

the companies that develop them. This means that it can be more difficult to elucidate the true scope of the tuning, beyond our definition of width: a particular LLM could be tuned to my full personal context *and* to the context of the company that developed it, that has a particular set of values and goals different from mine—something I cannot be aware of (unless the system is fully open-source, explainable and properly audited by trustworthy agents). I might find myself using a seemingly hyper-integrated AI technology (think of a future personal generative AI that accompanies you during your life to be continually trained in the whole history of your creative outputs and processes, and that is integrated at the deep level of perceptuo-motor token generation) that nevertheless is also inadvertently trained on branded content for an investor of the developer company. The authenticity of my outputs will surely be at stake. Future developments of the work we propose here, such as finer analytical work on the width dimension or the incorporation of further dimensions, are necessary to better account for current and future cases of generative AI use.

Will we stand on the shoulders of generative AI? If we understand, as Barandiaran and Almendros suggest, that LLMs operate as a library-that-talks (or a media-library that generates “new” media), generative midtended cognition implies an automatic standing on the writings of others. And this is already becoming an important productive and creative boost for those that have the opportunity to explore it. In this sense, the most optimistic conceptions of the future of generative AI consider its democratising potential, given its promise of levelling anybody up to an interdisciplinary PhD or expert-level intelligence. But we might also soon witness the emergence of a “generative class” whose main goal is to co-pilot generative machines within an assembly network of digital intelligence processes. Deep forms of alienation might ensue also at the productive side of generative cognition (not only at the “consumerism” and marketing-targeted side), analogous to, yet psychically deeper than, those that industrial capitalism brought with it.

However, this can also be turned around. An unexpected (certainly not industry-promoted) form of benefit from generative AI could be that of “authentic creativity through negative dialectics with generative AI”. It is possible to let yourself go by AI generative waves, but, on the contrary, it might be equally possible to surf or ride the waves tricking against the flow by a sort of negative dialectics: to take impulse on the generative wave just to twist direction and find creativity always *against* what the AI has generated (as suggested by Rushkoff, 2024). If instead of a conservative repetition of what was done in the past (by oneself and others), the ultimate and deepest sign of authenticity is to be genuinely open to change and becoming, then generative AI might boost the means to become more openly authentic by dialectically building against generically and even personally-tuned generative suggestions.

5 Conclusion: active integration beyond extension

The framework of generative midtended cognition that we have proposed aims to provide a theoretical basis to capture the hybrid processes where AI-generated suggestions become integral to the intentional creation of cognitive products by human agents. Understanding the ways in which artificial generative technologies

can become integrated in our cognitive processes, without admitting to them being full-fledged cases of social interaction or of extended cognition as theorised so far, becomes a crucial step to analysing our relationship with AI. By articulating the entanglement of different nested loops of generative power that contribute to a shared cognitive creative outcome, we have conceptualised two dimensions of active integration, width and depth. In doing so we have been able to characterise the specific case of generative cognitive processes occurring at a scale that was not displayed in our relations with previous forms of technologies: that of *midtended* cognition.

The novelty of our approach lies in the fact that previously existing theories of extended (or enactive) cognition did not foresee that the environment would be populated by the generative technologies we now have at our disposal. The environment was not thought of as being *generative* in the relevant sense (a purpose-structured kind of generativity similar to human creative, purposeful, practices) nor tuned to the specific context and agent in its generativity. This is not only the case for extended cognition theorising; postphenomenology, particularly as developed by Verbeek, has advanced a strong paradigm on understanding hybrid (human-technology) kinds of intentionality. Verbeek's (2008) account of cyborg and composite intentionality aims at something similar to our account of integration, since he recognises the intentional character of technology. However, the sort of intentionality instantiated by technology, in his account, is always directed towards the world, and not back to the agent. This account falls short of capturing current generative technology, where the arrow of (derived) "intentionality" (or purpose-structured generativity, in our account) departing from technology would also point towards the human agent.

Analysing the challenges and implications of these technologies through the lenses of these previous approaches (as developed so far) would tend to attribute all purpose-structured or normatively shaped contributions to the human agent, with the AI tool functioning only as a vehicle or a constraint in its realization, or would fail to capture the uniquely generative character of the products of the interaction. This misses the importance of the contributions of the AI technology in the generative process. As such, our approach allows us to further analyse the ethical risks that come with digital technologies and that were already starting to be raised within extended cognition literature. Coming back to the issue of transparency, for instance, our dimension of depth allows us to account for cases where the scale of interventions occurring below the creative unit grounds the phenomenological feel of transparency, while still retaining the fact that the cognitive process is a result of two different sources of generative power. The creative units that emerge from these hybrid processes (for instance, a sentence created with an autocomplete function) are a result of midtended cognition, not merely extended cognition.

On the other hand, treating these processes as true forms of social cognition would obscure the fact that generative AI technologies, while capable of producing generative outcomes, are not autonomous agents capable of participating in a true form of social interaction, with its constitutive tensions (see Di Paolo et al., 2018). As such, our capacity to negotiate our interactions with these kinds of automata is not similar to what occurs with other human agents. While this does not devoid midtended cognitive processes that include AI technologies of their creative or generative character,

it does have important implications for how we might be able to (collectively and individually) regulate them.

We have aimed, in the previous section, to hint at the specific risks, questions and possibilities that generative AI can pose if we understand them as bringing about midtended cognition processes. The granularity of the intervention in the joint generative process (captured by our depth dimension), together with the immense capacity for context-sensitivity (width dimension) and the status of the genAI automata as (currently market-driven) artificial products, has brought about specific concerns that are unprecedented in the history of technology. The political strategies needed to navigate these issues are different from those used to overcome issues emerging in purely social interactions between autonomous agents (but also from passive, non-generative technologies, as those amenable to extended-mind style theorizing). And we have shown that we do not need to consider LLMs as possessing human qualities to account for the fact that generative AI can intervene on cognitive processes in a generative manner not only at the conversational level, but also at deeper levels in the shaping of the process.

We need new concepts to talk about our relation with and use of generative AI. Cognitive science, and extended cognition in particular as the vantage tradition in including the environment in analyses and descriptions of cognitive processes, has a strong (political, philosophical and scientific) responsibility in offering analytical tools to assess not only the mode of existence of generative AI (Barandiaran & Almendros, 2024), but also our modes of interactions with it; its risks and its benefits.

Acknowledgements XEB and MPV acknowledge IAS-Research group funding IT1668-22 from Basque Government, Grants PID2019-104576GB-I00 for project Outonomy, and PID2023-147251NB-I00 for project Outagencies funded by MCIU/AEI/<https://doi.org/10.13039/501100011033>. MPV is developing her PhD thanks to the Grant PRE2020-096494 of the Spanish National Research Agency (AEI) and cofunded by the European Social Fund.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Akhtar, Z. B. (2024). Unveiling the evolution of generative AI (GAI): A comprehensive and investigative analysis toward LLM models (2021–2024) and beyond. *Journal of Electrical Systems and Information Technology*, 11(1), 22. <https://doi.org/10.1186/s43067-024-00145-1>
- Barandiaran, X. E. (2008). *Mental life: A naturalized approach to the autonomy of cognitive agents*. [PhD Thesis, University of the Basque Country (UPV-EHU)]. <https://xabier.barandiaran.net/phdthesis/>
- Barandiaran, X. E. (2019). Artificial democratic life. Re-engineering the autonomy of the social, a research program. *The 2019 Conference on Artificial Life*, 31, 11–12. https://doi.org/10.1162/isal_a_00131
- Barandiaran, X. E., & Almendros, L. S. (2024). *Transforming agency: On the mode of existence of large language models* (arXiv:2407.10735). arXiv. <http://arxiv.org/abs/2407.10735>
- Barandiaran, X. E. (2017). Autonomy and enactivism: Towards a theory of sensorimotor autonomous agency. *Topoi*, 36(3), 409–430. <https://doi.org/10.1007/s11245-016-9365-4>
- Barandiaran, X. E., Calleja-López, A., Monterde, A., & Romero, C. (2024). *Decidim, a technopolitical network for participatory democracy. Philosophy, practice and autonomy of a collective platform in the age of digital intelligence*. Springer. <https://link.springer.com/book/10.1007/978-3-031-50784-7>
- Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, 17(5), 367–386. <https://doi.org/10.1177/1059712309343819>
- Berg, S., & Hofmann, J. (2021). Digital democracy. *Internet Policy Review*. <https://doi.org/10.14763/2021.4.1612>
- Bhargava, V. R., & Velasquez, M. (2020). Ethics of the attention economy: The problem of social media addiction. *Business Ethics Quarterly*, 31(3), 321–359. <https://doi.org/10.1017/beq.2020.32>
- Bjarnason, R., Gambrell, D., & Lanthier-Welch, J. (2024). *Using Artificial Intelligence to accelerate collective intelligence: Policy synth and smarter crowdsourcing* (arXiv:2407.13960; Version 1). arXiv. <https://doi.org/10.48550/arXiv.2407.13960>
- Bratman, M. (1987). *Intention, plans, and practical reason*. Harvard University Press.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. arXiv:2005.14165 [Cs]. <http://arxiv.org/abs/2005.14165>
- Calvo, P., & Gomila, T. (2008). *Handbook of cognitive science: An embodied approach*. Elsevier Science.
- Cassinadi, G. (2022). Moral reasons not to posit extended cognitive systems: A reply to Farina and Lavazza. *Philosophy & Technology*, 35(3), 64. <https://doi.org/10.1007/s13347-022-00560-0>
- Clark, A. (1998). *Being there: Putting brain, body, and world together again*. MIT Press.
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(03), 181–204. <https://doi.org/10.1017/S0140525X12000477>
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7.
- Clowes, R. (2020). The Internet extended person: Exoself or doppelganger? *LÍMITE Revista Interdisciplinaria de Filosofía y Psicología*, 15(22).
- Cuffari, E. C., Di Paolo, E., & De Jaegher, H. (2015). From participatory sense-making to language: There and back again. *Phenomenology and the Cognitive Sciences*, 14(4), 1089–1125. <https://doi.org/10.1007/s11097-014-9404-9>
- Davidson, D. (1980). *Essays on actions and events (underlining)*. Oxford University Press.
- De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences*, 6(4), 485–507. <https://doi.org/10.1007/s11097-007-9076-9>
- Di Paolo, E. A., Buhmann, T., & Barandiaran, X. E. (2017). Sensorimotor Life: An enactive proposal. *Oxford University Press*. <https://doi.org/10.1093/acprof:oso/9780198786849.001.0001>
- Di Paolo, E. A. D., Cuffari, E. C., & Jaegher, H. D. (2018). *Linguistic bodies: The continuity between life and language*. MIT Press.
- Di Paolo, E. A., Lawler, D., & Vaccari, A. P. (2023). Toward an enactive conception of productive practices: Beyond material agency. *Philosophy & Technology*, 36(2), 31. <https://doi.org/10.1007/s13347-023-00632-9>

- Farina, M., & Lavazza, A. (2022). Incorporation, transparency and cognitive extension: Why the distinction between embedded and extended might be more important to ethics than to metaphysics. *Philosophy & Technology*, 35(1), 10. <https://doi.org/10.1007/s13347-022-00508-4>
- Gallagher, S. (2023). *Embodied and enactive approaches to cognition* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/9781009209793>
- González de la Torre, P., Pérez-Verdugo, M., & Barandiaran, X. E. (2024). *Attention is all they need: Cognitive science and the (techno)political economy of attention in humans and machines* (arXiv:2405.06478). arXiv (submitted to AI & Society). <http://arxiv.org/abs/2405.06478>
- Gubelmann, R. (2024). Large language models, agency, and why speech acts are beyond them (for now)—A Kantian-Cum-Pragmatist case. *Philosophy & Technology*, 37(1), 32. <https://doi.org/10.1007/s13347-024-00696-1>
- Haraway, D. (1991). *Simians, cyborgs, and women: The reinvention of nature*. Routledge.
- Harvey, I. (2024). Motivations for Artificial Intelligence, for Deep Learning, for ALife: Mortality and existential risk. *Artificial Life*, 30(1), 48–64. https://doi.org/10.1162/artl_a_00427
- Heersmink, R. (2015). Dimensions of integration in embedded and extended cognitive systems. *Phenomenology and the Cognitive Sciences*, 14(3), 577–598. <https://doi.org/10.1007/s11097-014-9355-1>
- Hernández-Orallo, J., & Vold, K. (2019). AI extenders: The ethical and societal implications of humans cognitively extended by AI. In *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society*, 507–513. <https://doi.org/10.1145/3306618.3314238>
- Hipólito, I., Winkle, K., & Lie, M. (2023). Enactive artificial intelligence: Subverting gender norms in human-robot interaction. *Frontiers in Neurorobotics*. <https://doi.org/10.3389/fnbot.2023.1149303>
- Hu, K., & Hu, K. (2023). ChatGPT sets record for fastest-growing user base—Analyst note. *Reuters*. <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>
- Hurley, S. (2010). The varieties of externalism. In R. Menary (Ed.), *The extended mind*. The MIT Press.
- Hutchins, E. (1995). *Cognition in the wild*. MIT Press Cambridge.
- Jebara, T. (2012). *Machine learning: Discriminative and generative*. Springer Science & Business Media.
- Malafouris, L. (2016). *How things shape the mind: A theory of material engagement*. MIT Press.
- Malafouris, L. (2019). Mind and material engagement. *Phenomenology and the Cognitive Sciences*, 18(1), 1–17. <https://doi.org/10.1007/s11097-018-9606-7>
- Martínez, I. C., Damesón, J., Pérez, J., Pereira Ghiena, A., Tanco, M., & Alimenti Biel, D. (2017). Participatory sense making in Jazz performance: Agents' expressive alignment. In Van Dyck, E. (Ed), *Proceedings of the 25th Anniversary Conference of the European Society for the Cognitive Sciences of Music, 31 July–4 August 2017, Ghent, Belgium*. https://sedici.unlp.edu.ar/bitstream/handle/10915/68089/Documento_completo.pdf-PDFA.pdf?sequence=1
- Merleau-Ponty, M. (1945). *Phénoménologie de la perception*. Gallimard.
- Murugesan, S., & Cherukuri, A. K. (2023). The rise of Generative Artificial Intelligence and its impact on education: The promises and perils. *Computer*, 56(5), 116–121. <https://doi.org/10.1109/MC.2023.3253292>
- Ong, W. J., & Hartley, J. (2012). *Orality and literacy: The technologizing of the word*. Routledge.
- Pérez, D. I., & Gomila, A. (2021). *Social cognition and the second person in human interaction* (1st ed.). Routledge.
- Pérez-Verdugo, M. (2022). Situating transparency: An extended cognition approach. *Teorema: Revista Internacional de Filosofía*, 41(3), 7–24. <https://www.jstor.org/stable/27185808>
- Pérez-Verdugo, M., & Barandiaran, X. (2023). Personal autonomy and (digital) technology: An enactive sensorimotor framework. *Philosophy & Technology*. <https://doi.org/10.1007/s13347-023-00683-y>
- Ricoeur, P. (1995). *Oneself as another* (K. Blamey, Trans.). Chicago University Press.
- Rushkoff, D. (Director). (2024). *Artificial Creativity| Team Human* [Video recording]. <https://shows.acast.com/teamhuman/episodes/artificial-creativity>
- Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge University Press.
- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631(8022), 755–759. <https://doi.org/10.1038/s41586-024-07566-y>
- Smart, P., Heersmink, R., & Clowes, R. W. (2017). The cognitive ecology of the internet. In S. J. Cowley & F. Vallée-Tourangeau (Eds.), *Cognition beyond the brain: Computation, interactivity and human artifice* (pp. 251–282). Springer International Publishing.
- Stelzly, K. (2010). Minds: Extended or scaffolded? *Phenomenology and the Cognitive Sciences*, 9(4), 465–481. <https://doi.org/10.1007/s11097-010-9174-y>

- Stiegler, B. (1998). *Technics and time, 1: The fault of epimetheus*. Stanford University Press.
- Sutton, J. (2010). Exograms and interdisciplinarity: History, the extended mind, and the civilizing process. In R. Menary (Ed.), *the extended mind* (pp. 189–225). The MIT Press. <https://doi.org/10.7551/mitpress/9780262014038.003.0009>
- Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5), 675–691.
- Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. MIT Press.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- Verbeek, P.-P. (2008). Cyborg intentionality: Rethinking the phenomenology of human–technology relations. *Phenomenology and the Cognitive Sciences*, 7(3), 387–395. <https://doi.org/10.1007/s11097-008-9099-x>
- Wheeler, M. (2019). The reappearing tool: Transparency, smart technology, and the extended mind. *AI & Society*, 34(4), 857–866. <https://doi.org/10.1007/s00146-018-0824-x>
- Williams, J. (2018). *Stand out of our light: Freedom and resistance in the attention economy*. Cambridge University Press. <https://doi.org/10.1017/9781108453004>
- Zawadzki, P., & Adamczyk, A. K. (2021). Personality and authenticity in light of the memory-modifying potential of optogenetics. *AJOB Neuroscience*, 12(1), 3–21. <https://doi.org/10.1080/21507740.2020.1866097>
- Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Xabier E. Barandiaran¹  · Marta Pérez-Verdugo¹ 

✉ Marta Pérez-Verdugo
marta.perezv@ehu.eu; perezverdugo.marta@gmail.com

Xabier E. Barandiaran
xabier.barandiaran@ehu.eu; xabier.academic@barandiaran.net

¹ IAS-Research Centre for Life, Mind and, Society (Department of Philosophy), University of the Basque Country (UPV/EHU), Tolosa Hiribidea 70, Gipuzkoa 20018 Donostia, Spain