Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Information Overload and Disutility in Generative AI Systems as Cognitive Extension: A Bounded Rationality Perspective

Rodrigo Laera

rodrigolaera@gmail.com

Consejo Nacional de Investigaciones Científicas y Técnicas

Research Article

Keywords: bounded rationality, generative AI, cognitive disutility, human–AI interaction, extended cognition

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11167481/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Information Overload and Disutility in Generative AI Systems as Cognitive Extension: A Bounded Rationality Perspective

Rodrigo Laera

Conicet/Sadaf

Beltrán 72 Buenos Aires, Argentina

rodrigolaera@gmail.com

Orcid: 0000-0002-5132-7631

Data Availability Statement
Not applicable.

Funding Statement
This research received no external funding.

Ethical Approval Statement
Not applicable.

Conflict of Interest Disclosure
The author confirms that there are no conflicts of interest.

Permission to Reproduce Material from Other Sources
Not applicable.

I confirm that this manuscript is original, has not been published before, and is not currently under consideration for publication elsewhere.

Abstract: Generative AI systems promise cognitive extensions by enhancing information access and use. However, from the bounded rationality perspective, their utility is fundamentally constrained by inherent human cognitive limitations. This article argues that when the volume of AI-generated information exceeds users' processing capacity governed by working memory constraints, attentional resources, and verification capabilities, it produces cognitive disutility rather than enhancement. We introduced a formal model to quantify this interaction, demonstrating that net cognitive benefits decline systematically as information overload increases beyond individual processing thresholds. The analysis reveals that successful cognitive extension depends not on maximizing informational output but on strategically aligning AI design with human cognitive ergonomics and individual differences in processing capacity. Ultimately, truly augmentative systems must minimize cognitive overhead to prevent overload, while preserving users' epistemic agency and decision-making autonomy.

Keywords: bounded rationality, generative AI, cognitive disutility, human–AI interaction, extended cognition.

1. Introduction

The proliferation of generative artificial intelligence systems -such as Large Language Models (LLMs) and related architectures- has changed the way people access, process,

and use available information, both in its factual and interpretative dimensions (Floridi and Chiriatti, 2020; Raiaan et al., 2024; Yenduri et al., 2024). These systems, which range from technical observations of empirical data to narratives, conceptual frameworks, or practical recommendations, not only diversify the modalities of access to knowledge, but also open up the possibility of reorganizing it in real time and adapting it to the needs of users, thus constituting a form of epistemic technology (Alvarado, 2023)[1]. The ability to filter, recombine, and contextualize the content they provide makes generative AI a mediator that not only passes on existing information but also actively reconfigures it according to statistical patterns and situational requirements (Haim and Hauser, 2024). In this sense, it can be argued that these large models can act as cognitive extensions that enhance our abilities to remember, analyze, and synthesize while reshaping the dynamics of knowledge production (Clark, 2025). That is, they not only mediate between user queries and the algorithmic selection of data, but also introduce new forms of epistemic interaction that challenge traditional notions of cognitive agency and knowledge dissemination (Hamilton and Benjamin, 2019; Rahwan et al., 2019).

However, the technological promise of augmented cognition based on generative AI faces a fundamental epistemic reality: the architecture of the human mind operates according to the principles of bounded rationality (Simon, 1955; 1957; March and Simon, 2005; Kahneman, 2011), which imposes structural constraints on any form of cognitive extension. While LLMs can generate seemingly unlimited answers to complex queries, human users face hidden cognitive costs: the need to verify accuracy and interpret extensive results, categorize information into relevant frames of reference, and make decisions under uncertainty regarding the quality of the answers (cf. Parasuraman and Riley, 1997). This paradox suggests that the net benefits of generative AI may be lower than expected and even negative in certain scenarios. For example, the limited capacity of human information processing (Miller, 1956) or working memory (Baddeley, 2007; Cowan, 2012) imposes restrictions that these AI systems neither recognize nor adequately consider.

This article proposes a theoretical framework and formal model to quantify the emergent disutility of human-AI interaction, considered a case of extended cognition. The central contention is that such disutility is not a mere technical side-effect, but inevitably

[1] That is, technology deployed in epistemic contexts to manipulate epistemic content through epistemic skills, where 'epistemic' refers to the broader categories of practices directly related to the acquisition, retention, use, and creation of knowledge. See also Alvarado (2023b).

arises from the human cognitive architecture when augmented by systems that lack an understanding of their own fundamental limitations[2]. While previous research has addressed specific aspects of human-AI interaction in terms of automation (e.g., Lee and See, 2004; Hoff and Bashir, 2015), this paper offers an integrated model that explains the global cognitive costs of such interaction from the perspective of bounded rationality. To this end, the article is organized in four steps: first, we provide the theoretical foundation of the problem in bounded rationality and the notion of extended cognition; second, we develop the specific cognitive implications of human-AI interaction; third, we formalize these phenomena; and finally, we address potential objections and draw implications for the design of systems focused on human agency.

2. Bounded Rationality and Extended Cognition

The development of cognitive sciences and artificial intelligence has produced a scenario in which the classical notions of rationality, cognitive limitation, and extension of the mind need to be reconsidered in an integrated way. In this context, the notion of bounded rationality should not be understood as a description of human inadequacy, but rather as a more accurate characterization of how real cognitive systems function under fundamental architectural constraints. That is, it is not just a practical limitation imposed by external computational costs, but a structural feature of systems that have evolved to function efficiently in complex environments with scarce resources.

From this perspective, people do not seek globally optimal solutions, but pursue solutions that are "good enough" for their immediate purposes given the constraints of time, information, and processing capacity (Simon, 1955). The effects of this cognitive architecture are both profound and complex. For example, human working memory, which is limited to approximately four elements at a time (Cowan, 2001), is not simply a technical bottleneck, but rather an evolutionary design feature that enables efficient processing in specific contexts. This limitation explains, among other things, the tendency to organize information in blocks or meaningful units, as can be observed in the

[2] See Gershman et al. (2015), who introduce the notion of *computational rationality*. According to this approach, intelligent agents make decisions oriented toward maximizing expected utility; however, given the inherent complexity of real-world problems, such decisions are typically realized only in approximate form. Computational rationality is thus proposed as a convergent framework for the study of intelligence across brains, minds, and machines. It aims to explain how cognitive systems—whether natural or artificial—arrive at effective solutions by taking into account not only the pursuit of utility, but also the computational costs associated with processing information in complex environments.

memorization of number sequences or language acquisition, where the grouping of elements facilitates the efficient storage and retrieval of information.

The use of heuristics in decision-making under uncertainty provides another compelling case. Tversky and Kahneman (1974) argued that people systematically resort to cognitive shortcuts- such as the availability or representativeness heuristic - that enable quick and generally adaptive responses, even if they do not always guarantee statistical accuracy. This reflects a cognitive architecture in which information processing is associated with opportunity costs, as resources allocated to one cognitive task are not available for others, creating a system of constant trade-offs that governs the phenomenological experience of human cognition (cf. Pashler, 1994). In this way, bounded rationality does not entail irrationality, but rather represents the adoption of strategies that maximize cognitive efficiency in contexts with incomplete or overloaded information (Gigerenzer and Goldstein, 1996).

In apparent contrast to this view of intrinsic limitations, the theory of extended cognition proposes a radically different ontology of the boundaries of the cognitive system. Clark and Chalmers (1998) have argued that, under certain functional conditions, external tools can constitute genuine parts of an extended cognitive system and not just auxiliary tools used by an agent[3]. Cognition is not necessarily confined to the boundaries of the skull, but can extend to the world if external resources fulfill the criteria of accessibility, reliability, and functional transparency. A notebook, a smartphone, or potentially an AI system can act as a genuine cognitive component if it is suitably integrated into an agent's thought and action processes.

Understood in these terms, the relationship between bounded rationality and extended cognition should not be reduced to a simple opposition. While the former describes the structural and evolutionary constraints under which human cognition operates, the latter redefines those very constraints by integrating external resources that effectively expand memory, processing, and analytical capacity beyond the biological domain. The tension between these perspectives opens fertile ground for inquiry: the limits of human cognition

[3] This gave rise to vigorous debate in which critics such as Adams and Aizawa (2001; 2008) challenged the thesis by introducing the "mark of the cognitive" argument, contending that external artefacts, however useful, lack the intrinsic processing features that define genuine cognition. A parallel line of criticism, developed by Rupert (2004; 2009), advanced the "integration objection," according to which external resources, though tightly coupled, do not constitute the kind of unified and enduring system necessary to qualify as cognition in its own right.

are neither absolute nor infinitely malleable, but are dynamically negotiated through the interaction between architectural constraints and technological affordances.

The fundamental tension between these two paradigms arises from the question of integration. If bounded rationality is understood as a fundamental architectural feature of human cognition, then successful cognitive extension cannot simply add external capabilities without considering how these new capacities interact with pre-existing limitations. By *extended cognition*, following Wilson and Clark (2008) and Clark (2019), we mean the idea that cognitive processes are not confined to the brain or nervous system, but are constituted through a distributed network of internal and external resources—tools, technologies, and symbolic environments—that together form a unified functional system.

The key lies not in the mere annexation of external devices or media, but in how they are fluidly and transparently coupled to the agent's perception–action–feedback cycle. When coupling is successful, internal limitations can be compensated for or reconfigured, resulting in a genuine increase in cognitive capacity. Conversely, when the coupling is fragile or discordant, the structural constraints of bounded rationality not only persist but may be amplified through new forms of dependency and error.

Effective extended cognition thus requires what we might call *architectural compatibility*: an alignment between the operational characteristics of external resources and the specific capabilities and limitations of the human cognitive system. For external resources to function as genuine components of an extended cognitive system, they must operate according to the principles of cognitive ergonomics that respect mental economy: they must be immediately accessible, their functioning must be transparent, and they must reduce rather than increase overall cognitive load.

Here, the central paradox that motivates the present analysis emerges: generative AI systems, precisely because of their power and seemingly unlimited capacity to generate information, can create irresolvable tensions with the bounded rationality architecture that characterizes human cognition. While historically successful extended cognition has involved tools that complement and amplify specific human capacities—calculators that extend arithmetic abilities, maps that extend spatial navigation, or notation systems that extend memory—LLMs present a qualitatively different case. These systems can generate responses to complex queries that exceed not only the user's specific knowledge but also their capacity to process, verify, and integrate the generated information.

Parasuraman and Riley (ibid.) anticipated this tension by explaining how automation can paradoxically lead to the degradation of human performance when the information provided by automated systems exceeds the operator's processing capacity, or when users develop dependencies that erode their autonomous skills. Similarly, Wickens (2002) argued that the introduction of additional information does not always improve performance and may, in fact, degrade it when it surpasses the cognitive resources available for effective processing. Both cases illustrate that the mere incorporation of external support does not guarantee a net increase in cognitive effectiveness; on the contrary, it may introduce new functional limitations and critical points of failure. Another particularly illustrative case of this dynamic is the phenomenon of complacency in automation, where operators tend to rely excessively on automatic systems and reduce their level of active monitoring, thereby increasing the risk of severe failures when automation malfunctions (Bainbridge, 1983; Sarter and Woods, 1995). Consequently, if automation is implemented without considering the limits of human cognitive capacity and patterns of interaction with technology, it may systematically impair operator performance, increasing the likelihood of critical errors.

In the specific context of generative AI, this tension manifests simultaneously across several domains. An AI system may generate a thousand-word response to a complex query; however, the human user, operating under bounded rationality, can process only a fraction of it effectively. The remainder becomes informational surplus—material that must be skimmed, ignored, or deferred—introducing costs of overload, trust calibration, and metacognitive effort. What appears as abundance from the system's perspective may, from the user's standpoint, translate into diminished utility and cognitive strain. That is, the cognitive cost of this processing can exceed the informational benefit, particularly when the user must simultaneously maintain uncertainty about the accuracy and relevance of the generated information. Thus, an asymmetry arises: while the AI system operates without apparent capacity constraints, the human user incurs marginal and limiting costs for each piece of information that must be processed.

This asymmetry suggests that the successful integration of generative AI as a cognitive extension requires a reconceptualization of how we approach optimization in hybrid cognitive systems. Instead of trying to maximize the technical capacity of the artificial component, optimization must also focus on minimizing the total disutility of the integrated system, explicitly taking into account the architectural limitations of the human component. This means that the design of cognitively effective AI systems should

be based not only on a deep understanding of their potential technical capabilities, but also on a precise understanding of the actual cognitive constraints that shape the human experience of interacting with these systems.

The theoretical synthesis that emerges from this dialectic is not a simple reconciliation between bounded rationality and extended cognition, but a more nuanced understanding of the conditions under which cognitive extension can be truly beneficial for systems operating under bounded rationality. Successful extended cognition does not transcend the limitations of bounded rationality; rather, it respects and optimizes them, creating configurations that expand effective capacities while minimizing the cognitive costs of integration. In the context of generative AI as extended cognition, this means that the design of truly human-centered systems must begin not with available technical capabilities, but with a precise understanding of the cognitive phenomenology of human–AI interaction under conditions of bounded rationality.

3. Cognitive Implications of Human–Generative AI Interaction

Bounded rationality prevents the exhaustive verification of AI outputs, leaving users with the dilemma of either trusting blindly—thus assuming the risk of error—or verifying extensively, thereby undermining efficiency. This cost can be particularly acute because LLMs may generate convincing errors that require subject-matter expertise to detect (cf., Kalai et al., 2025). Moreover, outputs are composed and selected by algorithms that operate according to opaque logics; that is, the processes of generation and selection function as black boxes for the end user, who lacks visibility into the criteria, biases, or internal mechanisms that produced the specific output (Pasquale, 2015)[4]. Algorithmic opacity intensifies the verification dilemma because, without access to the underlying logic, users must evaluate credibility based solely on the superficial properties of the generated text[5]. In this way, extension through LLMs becomes not only a source of cognitive overload but also a source of epistemic dependence, in which trust shifts from the content itself to the implicit authority of the system. The asymmetry between users and models thus deepens: the former lack the resources necessary to scrutinize internal processes, while the latter impose their results as accomplished facts.

[4] For a detailed discussion of the concept of opacity, cf., Burrell (2016).

[5] Moreover, Gerlich (2025) argues that the frequent use of AI tools is negatively correlated with critical thinking, a relationship mediated by cognitive offloading; that is, the strategy of externalizing or "offloading" a mental task onto an external support in order to reduce cognitive demand and effort. In essence, this practice frees up working memory to focus on other activities.

Rather than alleviating the limitations of human processing, an opaque system with costly verification exacerbates the trade-offs inherent to bounded rationality. The promise of expanded cognitive capacity collides with the reality of poor ergonomics, where the interface between the human mind and the AI system introduces frictions that consume the scarce attentional and working memory resources it was meant to liberate. Therefore, for generative AI to avoid functioning as a cognitive burden, it must incorporate aspects of practical reason that provide intrinsic evaluative scaffolds, a requirement that poses significant challenges for its conceptualization as extended cognition.

This would require designing outputs that natively integrate elements of inherent readability and contextual justification, enabling rapid assessment without demanding extensive analysis. The objective would be to transform the interaction from an ex-post verification process into one of fluid evaluative apprehension, with the reliability and relevance of content signaled through structural and design cues that align with natural human judgment mechanisms. Such an approach would not only mitigate cognitive load but also restore the user's epistemic agency, allowing cognitive extension under conditions of informed dependence rather than blind reliance on IA. Specifically, to address different types of cognitive load:

• Contextual justification helps reduce extraneous load by providing clear explanations that prevent unnecessary cognitive effort from being expended in deciphering information.

• Inherent readability improves intrinsic load management by organizing complex materials into a more digestible format.

• Design cues that align with natural judgment mechanisms enhance the germane load by supporting schema building and automation.

Consider, for example, how humans develop systematic biases in trust, tending toward both over-trust and under-trust depending on the context (Muir, 1994; Schaefer et al., 2016). These dynamics create persistent cognitive dissonance between the need to verify system outputs and the limited capacity to do so exhaustively, leading to reliance on trust heuristics that, while practical, may be suboptimal for judgment quality (Hoff and Bashir, 2015). Moreover, the time and resources invested in verification carry significant opportunity costs that are often underestimated in the design of human–AI interactions. Thus, the relationship with intelligent systems is shaped by tensions between efficiency, reliability, and cognitive load, which directly influence how users calibrate trust and make AI-assisted decisions. Additionally, institutional incentives can impact in these trust

dynamics by shaping who benefits when users over-trust or under-trust the system. For example, developers may benefit from over-trust as it reduces resistance to AI adoption, while under-trust may align with regulatory interests protecting user autonomy. These sociotechnical power dynamics highlight the importance of policy considerations that address system-level accountability beyond individual cognition.

As has been established, AI outputs are not consumed passively; they must be interpreted and integrated within the user's preexisting mental frameworks. For boundedly rational agents, this mediation can become counterproductive when the amount or complexity of information exceeds the available processing capacity. Cognitive Load Theory provides a useful framework for understanding these limits by distinguishing between intrinsic load (built-in difficulty), linked to the inherent complexity of the material; extraneous load (unnecessary difficulty), stemming from the way information is presented; and germane load (helpful difficulty), associated with resources dedicated to building and automating mental schemas (Paas et al., 2003; Sweller et al., 2019).

Within this framework, various mechanisms illustrate how AI outputs can saturate the available cognitive resources. For instance, working memory is quickly overloaded when lengthy responses exceed the simultaneous retention capacity. Additionally, there is the effort required to integrate new information with prior knowledge, a process that intensively consumes the resources available for learning and comprehension (Van Merriënboer and Sweller, 2005). Attention also becomes fragmented when users must simultaneously manage multiple informational elements or shift between levels of detail. Consequently, generative AI systems may generate what could be termed cognitive friction: additional resistance in the process of assimilating presented information that drains mental resources without necessarily enhancing comprehension. This friction is particularly problematic because it operates silently; users are not always aware of the extra effort required to decode, contextualize, and reconcile AI-provided information with their own mental schemas.

This challenge also arises because interaction with AI is inherently uncertain; users cannot know the quality of responses a priori. Under conditions of bounded rationality, they thus resort to simple heuristics, which, although they reduce immediate processing demands, may prove to be systematically suboptimal. The literature on trust in automated systems documents that individuals develop calibration strategies that do not always adapt to the specific dynamics of generative AI (Thorne, 2024). Common cases include

the length heuristic, which assumes that longer responses are of higher quality; the confidence heuristic, which takes a persuasive tone as a signal of truthfulness; and the effort heuristic, which leads users to dismiss responses requiring excessive cognitive effort. Although efficient from a cognitive standpoint, such heuristics can induce significant misjudgments when applied to AI outputs that do not conform to usual patterns of human communication. For instance, once the system generates a highly confident but incorrect response that aligns with the user's preexisting beliefs, reinforcing confirmation bias and making the error more likely to be accepted uncritically.

If these interactions are repeated over time, then feedback and dependency effects emerge, as the heuristics employed tend to reinforce themselves through experience, even when they lead to mistaken evaluations. What develops here is what we might call "cognitive scaffolding drift": a process whereby the extended mind incorporates not merely the capabilities of external tools, but also embeds both the systematic failures of these tools and our own recurring mistakes in managing them. Just as a skilled craftsperson's hand learns to compensate for the particular wobble in a familiar tool, our cognitive systems adapt to accommodate not only AI limitations but also our predictable patterns of suboptimal engagement, for instance, our tendency to excessively trust confident-sounding responses, our failure to cross-reference outputs, and our systematic neglect of uncertainty markers. The extended mind, in this view, becomes a hybrid system that institutionalizes both technological shortcomings and human error patterns. We do not simply incorporate AI as a neutral cognitive prosthetic; rather, we develop integrated mind-tool assemblages that perpetuate and normalize specific failure modes. The coupling becomes so seamless that we lose track of where technological limitations end and our own cognitive blind spots begin. The ultimate result of this scaffolding drift can be distributed cognitive dysfunction, a stabilized state where the extended cognitive system operates smoothly and efficiently while systematically reproducing compound errors that neither the human nor the AI component would generate independently.

However, not all individuals experience these dynamics in the same way. Research on individual differences has revealed significant heterogeneity in cognitive capacities and processing strategies, creating differentiated vulnerabilities, especially when it comes to complex problem solving (Fleischhauer et al., 2010). Domain experts can verify outputs more efficiently in their areas of expertise, while those with greater working memory capacity are better able to withstand high informational loads. Experience with AI systems also plays a central role, as users who are accustomed to interacting with these

models tend to develop better-calibrated heuristics. Tolerance for uncertainty is also an important factor, as some people are better able to deal with the ambiguity of AI's results. Taken together, these dynamics suggest that interactions with AI are neither cognitively neutral nor homogeneous in their effects on users. Rather, it is permeated by tensions between efficiency, trust, and overload, which define both the potential and the limits of its integration into everyday practice.

To quantify the cognitive implications discussed, the following section introduces a formal model that captures the dynamic interplay between human cognitive limits and AI-generated information flows. This framework formalizes how information volume, processing capacity, and temporal recovery interact to determine whether AI functions as an effective cognitive extension or becomes a source of overload and disutility.

4. Bounded Rationality and Cognitive Extension through Generative AI: A Brief Formal Model

The proposed model seeks to formalize the interaction between human cognitive capacity and the information supplied by a generative artificial intelligence system, framing this relationship within the principles of bounded rationality. Let $C(t)$ denote the cognitive capacity available at time t, understood as the quantity of effective cognitive resources, such as attention or working memory, while the information provided by AI is denoted by I, considered constant for simplicity. Thus, the information effectively processed by the user is defined as $P(t) = min\big(I, C(t)\big)$, which implies that processing capacity can never exceed the cognitive resources available at a given moment. For instance, if a physician queries an AI system about a complex diagnostic case and receives a comprehensive 2,000-word analysis covering multiple differential diagnoses, treatment protocols, and research citations, but their available cognitive resources during a busy shift only permit thorough processing of 800 words, the model predicts that only this limited subset will be effectively integrated into their clinical decision-making.

The maximum capacity that the user can reach, determined by the architectural constraints of human cognition, is denoted as C_{max}. The dynamics of cognitive capacity are governed by two parameters: α, which represents the recovery rate of capacity as a function of factors such as rest or experience, and β, which expresses the rate of cognitive resource consumption per unit of information processed. Additionally, the model introduces variables associated with the costs and benefits of interaction: b, the benefit

per unit of information processed; c, the cost per unit of information received, understood as the minimum attention effort required; and d, the additional cost associated with overload when the information exceeds processing capacity. A practical example would be a lawyer using an AI system to generate legal drafts: each well-processed clause provides an immediate benefit (b), but even if not, all clauses are reviewed, the lawyer must bear a minimum reading cost (c), and if the generated contract contains more information than can be reviewed, cognitive overload is incurred (d). Consequently, the evolution of cognitive capacity can be described by the following differential equation:

$$\frac{dC}{dt} = \alpha(C_{max} - C(t)) - \beta P(t)$$

The first term, $\alpha(C_{max} - C(t))$, represents the natural recovery of capacity toward its maximum value. The second term, $\beta P(t)$, corresponds to cognitive depletion resulting from the processing of effectively assimilated information; for instance, when someone must review extensive AI-generated reports and their attentional resources become depleted. The instantaneous net utility of the interaction is expressed as: $U(t) = bP(t) - cI - dmax\big(0, I - C(t)\big)$. Here, $\beta P(t)$, reflects the benefits obtained from the processed information, c I the unavoidable costs of reception and attention, and $dmax\big(0, I - C(t)\big)$ the overload cost when the flow of information exceeds the available capacity. In a corporate environment, this is equivalent to an analyst who gains value only from the data they manage to effectively integrate, while incurring fixed reception costs and additional costs if the AI-generated report exceeds their processing capacity.

If the analysis is conducted in steady state, two relevant cases arise. In the first case, where $I \leq C(t)$, the information never surpasses cognitive capacity and is fully processed, so $P(t) = I$. The equilibrium condition reduces to: $0 = \alpha(C_{max} - C_{ss}) - \beta I$, from which we obtain $C_{ss} = C_{max} - \frac{\beta}{\alpha} I$. It is important to note that this result is consistent only if the constraint holds $I \leq C'_{ss}$, which is equivalent to $I \leq \frac{\alpha}{\alpha+\beta} C_{max}$. An example of this situation occurs when a user requests a brief executive summary of a few paragraphs from the AI: the amount of information is low enough to be processed without overload, and the net utility is positive whenever $b > c$. In this scenario, the stationary utility simplifies to $U_{ss} = (b - c)I$.

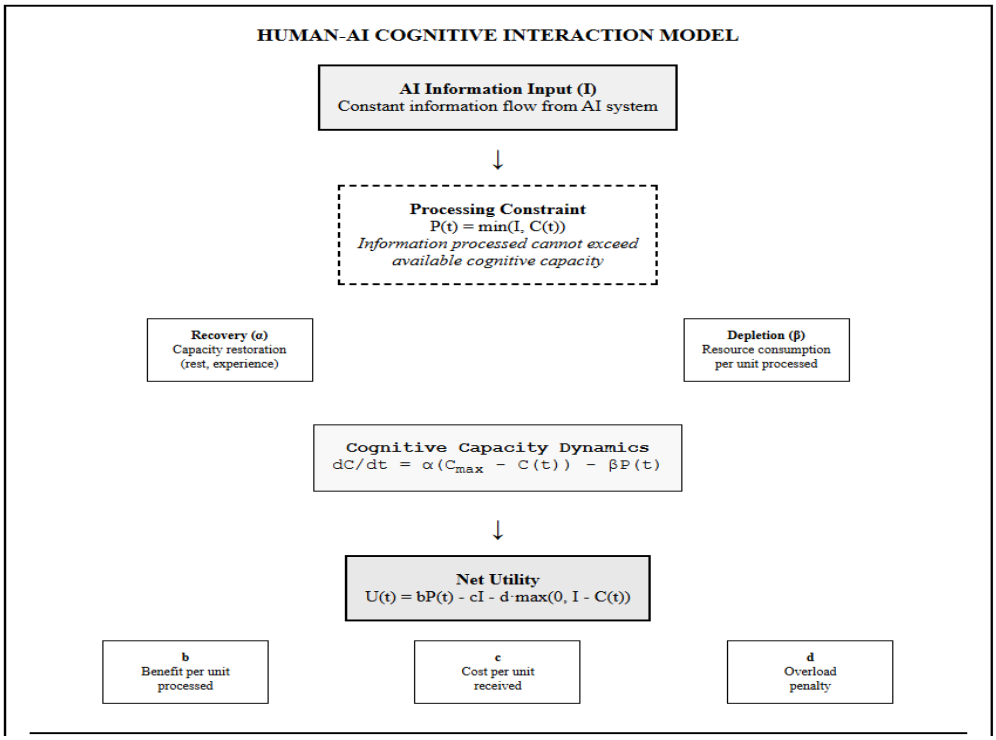


Figure: 1. Human-AI Cognitive interaction model.

Source: Author's own elaboration

In the second case, where $I > C(t)$, the information provided exceeds cognitive capacity. Here, $P(t) = C(t)$, and the equilibrium equation becomes: $0 = \alpha(C_{max} - C_{ss}) - \beta C_{ss}$, which leads to $C_{ss} = \frac{\alpha}{\alpha+\beta} C_{max}$. The stationary net utility can be expressed as:

$$U_{ss} = (b + d)\frac{\alpha}{\alpha + \beta} C_{max} - (c + d)I$$

This resembles the case of a user requesting an excessively long report from AI, such that their processing capacity stabilizes at a level below what is required, and cognitive overload introduces costs that may render net utility negative, even when part of the information is valuable.

In interpretive terms, the model captures the inherent tension between the contribution of generative AI, which amplifies access to information, and human cognitive limitations, which restrict processing capacity. Disutility arises when the amount of information exceeds the capacity for assimilation, generating additional overload costs. The dynamics of $C(t)$ show how cognitive capacity degrades with information processing, while also exhibiting recovery mechanisms that reflect the architecture of bounded rationality. In sum, the model emphasizes that cognitive extension through AI is beneficial only insofar as the flow of information remains compatible with the user's cognitive capacities.

Otherwise, overload erodes net utility and transforms the promise of cognitive augmentation into a source of disutility.

Having established a formal model that reveals the potential cognitive disutility in human-AI interactions, we now turn to address critical objections that highlight the complexity and potential oversights of our analysis. This will allow us to refine the theoretical framework, clarify the scope of its claims, and situate its relevance within broader debates about cognitive extension and human-centered AI design.

5. Objections and Responses

The discussion about the interaction between humans and artificial intelligence leads to a number of objections to the claim that generative systems can induce cognitive disutility. Critics emphasize alternative interpretations of overload, the role of design and the general benefits of AI. To clarify these debates, this section presents three objections and responses that situate the argument within the limits of bounded rationality while acknowledging opposing perspectives.

First objection: *Disutility is a temporary design problem, not an inevitable consequence of human cognitive architecture*. One could argue that the problems of overload and disutility identified here are not inherent to the interaction between humans and AI, but rather flaws in current interfaces and the way information is presented. With advances in prompt engineering, algorithmic personalization, and user-centric interface design, it is possible to optimize LLM outputs so that they are more concise, relevant, and easier to process. For example, a system could learn to adapt the length and complexity of its answers to the user's cognitive profile, or automatically summarize important information. In this sense, disutility would not be an inevitable outcome, but rather a technical challenge that technological development will eventually solve so that AI can ergonomically integrate with the human mind.

Response: Although it is true that improvements in design can mitigate some cognitive costs, the objection underestimates the depth and inevitability of the constraints imposed by bounded rationality. The problem lies not only in how the information is presented, but also in the fundamental asymmetry between the unlimited generative capacity of AI and the finite, structurally constrained processing capacity of human beings. The model presented in this paper formalizes this asymmetry. The variable $C(t)$, which represents the available cognitive capacity, has an upper bound C_{max} posed by biological architecture (working memory, sustained attention, etc.). No matter how much I (the

information provided by AI) is optimized, if $I > C_{max}$, the user will never be able to process it in full. Design optimization can reduce the cost per unit of information received, or even the overload cost, but it cannot eliminate them. Such cost arises precisely when $I > P(t)$, that is, when there is an informational surplus that the user must ignore, discard, or process superficially, which generates frustration, anxiety, and an expenditure of resources in metacognitive decision-making (What should I read? What should I skip?). Moreover, this objection ignores the problem of algorithmic opacity. Even with a perfect interface, the user is still faced with the trust dilemma: how is he supposed to know whether the concise summary presented by the AI corresponds to the original content or whether it omits crucial information? Paradoxically, checking the accuracy of a summary requires access to the full text, reintroducing the information overload that should be avoided. As explained here, this leads to an "epistemic dependence" in which the user's agency is undermined. Thus, disutility is not just a "bug" in the system, but a property that arises from the coupling of an unlimited generative system with a limited processing system- a tension that cannot be fully resolved by refining the interfaces, only managed.

Second objection: *The model ignores the value of latent information and the human capacity for learning and adaptation*. Another plausible objection is that the proposed model is excessively static and pessimistic. It assumes that cognitive capacity C_{max} is a fixed and immutable limit, and that information not processed at time t is entirely useless. However, humans do not process information only in real time; they also archive it, revisit it later, and learn from it. A lengthy output from an LLM, even if not fully assimilated in the first reading, may serve as a future reference source; a kind of second brain that extends long-term cognitive capacity. Moreover, continuous interaction with AI can train users, effectively increasing their C_{max} through learning and familiarization with the system. In this sense, the initial cognitive investment, though costly, yields long-term returns that the instantaneous utility model fails to capture.

Response: This objection raises valid points about the dynamic nature of human learning, but it does not invalidate the core argument; rather, it qualifies it within a broader temporal horizon. The model presented here focuses on instantaneous net utility and steady-state conditions precisely to isolate and quantify immediate cognitive costs—where friction is most acute and where disutility may become negative. Regarding "latent information," its potential value is conditioned by factors that bounded rationality directly constrains: the cost of retrieval and uncertainty about quality. For archived text to be

useful in the future, the user must remember its existence, know where to find it, and trust that its content is accurate and relevant. Each step entails cognitive expense. The very act of archiving and organizing AI-generated information (labeling, indexing, storing) consumes $C(t)$ resources that could be devoted to other activities. In cognitive load theory, this constitutes extraneous load: effort that does not directly contribute to learning or task resolution, but to managing the learning environment.

Regarding adaptation and increases in C_{max}, experience can indeed improve efficiency. Expert users develop better heuristics and require less effort to process similar outputs. However, this does not eliminate the problem—it redistributes it. The model can incorporate this through the parameter α (recovery rate), which could increase with experience. Yet even with high α, the limit C_{max} remains a physical barrier. One may become faster and more efficient, but cannot simultaneously process ten pages if working memory manages only four information chunks. Adaptation enables better handling of overload; it does not transcend it. Therefore, while long-term value exists, it remains subject to bounded rationality principles: the user must strategically decide which information merits investment in future storage and retrieval—itself a cognitively costly metacognitive task.

Third objection. *Focusing on individual disutility ignores the systemic and collective benefits of generative AI*. The objection can be raised that the analysis is excessively individualistic. While a specific user may experience disutility when interacting with an LLM, the system as a whole generates immense value by democratizing access to knowledge, enabling large-scale collaboration, and accelerating collective innovation. For instance, a researcher might use an LLM to generate 20 hypotheses, discard 18 after a preliminary evaluation, and explore the remaining ones in depth. Although processing the 18 discarded hypotheses incurs a cost, the benefit of identifying the two valuable hypotheses more than justifies the expenditure. At the macro level, individual overload is the necessary price for systemic efficiency and creativity.

Response. This objection is strong and reflects a common utilitarian perspective in technology evaluation. However, it confuses systemic potential with the utility experienced by the individual cognitive agent, which is the focus of this paper. It does not deny that generative AI can yield net positive benefits at the social or economic level. What it maintains is that, from the user's perspective as a cognitive agent with bounded rationality, the interaction can be disutilitarian. The researcher's example is illustrative. The formal model can be applied perfectly to this case. Suppose the benefit of a valuable

hypothesis is $b = 100$ (arbitrary units), but the cost of superficially evaluating a hypothesis is $c_r = 5$. If the researcher evaluates 20 hypotheses, their total cost is 100. If they find 2 valuable ones, the benefit is 200, yielding a net utility of $+100$. This seems positive. But this assumes the researcher has sufficient cognitive capacity available to evaluate the 20 hypotheses without incurring overload costs. If the task of evaluating 20 hypotheses drive, them into cognitive fatigue ($C(t)$ rapidly decays), the overload cost may be very high: errors of judgment, loss of motivation, and reduced quality in the evaluation of later hypotheses. In this scenario, the researcher could discard a valuable hypothesis (false negative) or, worse, pursue an erroneous one (false positive), which would carry an even greater cost. By warning about "feedback and dependency effects," the argument also holds that cognitive extension can become a source of long-term disutility. If the researcher becomes accustomed to generating 20 hypotheses with AI, their ability to generate hypotheses independently may atrophy, making them dependent on the system. This could erode their cognitive agency and critical thinking capacity, even if this lack of autonomy is a systemic cost not immediately reflected in the short-term utility of finding good ideas. Systemic benefits are real, but they do not cancel out the disutility experienced at the individual level. A system may be socially efficient while being cognitively costly (or even harmful) for its users. This paper proposes that we must design systems that are truly human-centered not only to maximize system output, but also to minimize the agent's disutility, while respecting the limits of bounded rationality. True innovation is not about generating more information, but about generating the right information at the right time and in the right form that the cognitively impaired user can actually process.

These objections highlight the complexity of assessing the cognitive costs of human-AI interaction. They point to the promise of improved design, the dynamic nature of learning and the undeniable systemic value of generative technologies. Yet, as the responses show, none of these considerations overturns the fundamental asymmetry between unlimited generative capacity and bounded human cognition. Rather than dismissing the disutility as a temporary problem, the analysis suggests that it is an inherent tension that must be carefully managed if AI systems are to remain truly human-centered.

6. Conclusion

This analysis argues that generative AI as a cognitive extension is characterized by the architecture of bounded rationality. These systems promise to improve memory, analysis and synthesis, but their cognitive utility is not guaranteed. The main problem lies in asymmetry, as generative AI can produce vast amounts of information, while humans are limited by memory, attention and processing speed. Our framework and model show that information flows beyond the capacity of the user are a disadvantage. This makes a tool for augmentation a source of overload, dependency and strategic error.

The promise of cognitive extension depends on the AI's compatibility with the human mind, not its raw power. Successful extension works within the limits of bounded rationality; it does not override them. This requires a shift in AI design. Instead of maximizing output, we need to minimize the cognitive cost of interaction. Ideal systems, therefore, act as ergonomic partners, providing outputs that are structured for rapid evaluation and seamless integration.

Ultimately, the challenge of generative AI is not only technical, but also human. We need to rethink what it means to think and know with intelligent systems. True cognitive enhancement is measured by understanding, not by the amount of information. Human-centered AI demands respect for our cognitive limitations. This ensures that the extended mind grows not only larger, but also wiser. The future of human-AI collaboration depends on the development of systems that complement, rather than overwhelm, our bounded rationality.

Bibliography

Adams, F., & Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology, 14*(1), 43–64. doi:10.1080/09515080120033571

Adams, F., & Aizawa, K. (2008). *The bounds of cognition.* Oxford: Wiley-Blackwell.

Alvarado, R. (2023). AI as an Epistemic Technology. *Science and Engineering Ethics, 29*(32). doi:10.1007/s11948-023-00451-3

Alvarado, R. (2023b). What kind of trust does AI deserve, if any? *AI and Ethics, 3*(4), 1169-1183.

Baddeley, A. (2007). *Working memory, thought, and action.* Oxford: Oxford University Press.

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19*(6), 775-779.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1). doi:10.1177/2053951715622512

Chalmers, D. (2019). Extended cognition and extended consciousness. In M. Colombo, E. Irvine, & M. Stapleton (Eds.), *Andy Clark and his critics* (pp. 9-20). Oxford: Oxford University Press.

Clark, A. (2025). Extending Minds with Generative AI. *Nature Communications, 16*(4627). doi:10.1038/s41467-025-59906-9

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7-19.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences, 24*(1), 87-114.

Cowan, N. (2012). *Working memory capacity.* New York: Psychology Press.

Doshi, A. R., & Hauser, O. (2024). Generative artificial intelligence enhances creativity but reduces the diversity of novel content. *Science Advances, 10*(28). doi:10.1126/sciadv.adn5290

Fleischhauer, M., Enge, S., Brocke, B., Ullrich, J., & Strobel, A. (2010). Same or different? Clarifying the relationship of need for cognition to personality and intelligence. *Personality and Social Psychology Bulletin, 36*(1), 82-96. doi:10.1177/0146167209351886

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and machines, 30*(4), 681-694. doi:10.1007/s11023-020-09548-1

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1). doi:10.3390/soc15010006

Gershman, S. J. (2020). Origin of perseveration in the trade-off between reward and complexity. *Cognition, 204*, 104394. doi:10.1016/j.cognition.2020.104394

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science, 349*(6245), 273-278.

Gigerenzer, G., & Goldstein, D. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*(4), 650-669.

Hamilton, K. A., & Benjamin, A. S. (2019). The human-machine extended organism: New roles and responsibilities of human cognition in a digital ecology. *Journal of Applied Research in Memory and Cognition, 8*(1), 40–45. doi:10.1016/j.jarmac.2019.01.001

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors, 57*(3), 407-434. doi:10.1177/0018720814547570

Hoff, K., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors, 57*(3), 407-434. doi:10.1177/0018720814547570

Kahneman, D. (2011). *Thinking, fast and slow.* New York: Farrar, Straus and Giroux.

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors, 46*(1), 50-80.

March, J. G., & Simon, H. A. (1958). *Organizations.* New York: John Wiley & Sons.

Miller, G. (1956). Information and memory. *Scientific American, 195*(2), 42-47.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics, 37*(11), 1905-1922.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational psychologist, 38*(1), 1-4.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230-253. doi:10.1518/001872097778543886

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin, 116*(2), 220-244.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information.* Cambridge, MA: Harvard University Press.

Rahwan, I., Cebrian, M., Obradovich, N., & et al. (2019). Machine behaviour. *Nature, 568*, 477–486. doi:10.1038/s41586-019-1138-y

Raiaan, M. A., Mukta, M. S., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M., & Azam, S. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access, 12*, 26839-26874. doi:10.1109/ACCESS.2024.3365742

Rupert, R. D. (2004). Challenges to the hypothesis of extended cognition. *Journal of Philosophy, 101*(8), 389–428.

Rupert, R. D. (2009). *Cognitive systems and the extended mind.* Oxford: Oxford University Press.

Sarter, N. B., & Woods, D. D. (1995). How in the World Did We Ever Get into That Mode? Mode Error and Awareness in Supervisory Control. *Human Factors, 37*(1), 5-19. doi:10.1518/001872095779049516

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors, 58*(3), 377-400. doi:10.1177/0018720816634228

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics, 69*(1), 99-118.

Simon, H. A. (1957). *Models of Man.* New York: Wiley.

Sweller, J., Van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational psychology review, 31*(2), 261-292.

Thorne, S. (2024). Understanding the interplay between trust, reliability, and human factors in the age of generative ai. *International Journal of Simulation--Systems, Science & Technology, 25*(1). doi:10.5013/IJSSST.a.25.01.10

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science, 185*(4157), 1124-1131.

Van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational psychology review, 17*(2), 147-177. doi:10.1007/s10648-005-3951-0

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science, 3*(2), 159-177.

Wilson, R. A., & Clark, A. (2008). How to situate cognition: Letting nature take its course. In P. Robbins, & M. Aydede (Eds.), *The Cambridge Handbook of Situated Cognition* (pp. 55-77). Cambridge: Cambridge University Press.

Yenduri, G., Ramalingam, M., Selvi, G. C., Supriya, Y., Srivastava, G., Maddikunta, P. K., & ... Gadekallu, T. R. (2024). Gpt (generative pre-trained transformer)—A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE, 12*, 54608-54649. doi:10.1109/ACCESS.2024.3389497