AI and the Risk of Cognitive Atrophy in Human Cognition

Alexander Norrstam Vatamidis, Marcus Lindelöf, Valery Nkenguruke

This inquiry analyzes the ethical consequences of integrating artificial intelligence into human cognitive life. Grounding the analysis in frameworks such as the Free-Energy Principle and the Technological Approach to Minds Everywhere, cognition is defined as an active process of relevance realization. The inquiry traces the evolution from surveillance capitalism's attentional control to the generative AI era, highlighting ontological shifts toward the "cyborgic". We distinguish between restorative and augmentative applications, each with a different set of cognitive consequences. Ultimately, we propose a normative framework advocating for systems designed to return abundance to human practice rather than replacing it.

Introduction

The integration of artificial intelligence (AI) into human cognitive life presents a complex set of moral and theoretical questions. AI systems increasingly participate in cognitive processes once considered uniquely human, such as reasoning, memory, creativity, and decision-making, and by doing so they do not merely extend human cognition but reshape it. The guiding question in what follows is not whether AI will surpass human intelligence, but how its widespread use, especially in the form of large language models (LLMs), recommender algorithms, and emergent neuro-integrative technologies, reconfigures the structure and quality of human cognition. This is the ethical case to be analyzed: a socio-technical transformation in which tools that assist cognition also modify its conditions of possibility. The resulting ethical question is how we should relate to systems that extend cognition while simultaneously transforming it.

We propose that this transformation involves a dual movement: while AI has the potential to enhance cognitive capacities, it also risks a gradual reconfiguration that may, under certain conditions, diminish them. This process is common particularly when technological systems, when optimized for convenience, efficiency, or profit, can subtly alter attention, autonomy, and understanding. What begins as cognitive assistance can, in specific social and economic contexts, lead to new forms of dependency and passivity. The resulting dilemma is both philosophical and ethical: what are the ethical consequences of using systems that extend cognition while simultaneously transforming its underlying structure and quality, and how should we relate to them?

To address this question, our analysis proceeds in five parts. First, we provide a definition of cognition, drawing on contemporary frameworks such as the Free-Energy Principle (Friston, 2010) and the Technological Approach to Minds Everywhere (Levin, 2022), to establish the concept of relevance realization as foundational. Second, we describe the historical context, examining the pre-LLM stage of surveillance capitalism to show how algorithms engineered our salience landscapes. Third, we analyze the current state and trajectory of generative AI, detailing the structural risks of model collapse and the transition into new ontologies (the virtual, the artificial, and the cyborgic). Fourth, we utilize the distinction between restoration and augmentation to explore the immediate and prospective ethical challenges, including the mechanism of cognitive atrophy via the principle of cognitive efficiency, and systematically identify the interests of the relevant actors and stakeholders (corporations, governments, individuals, and the public). Finally, on the basis of this comprehensive analysis, we propose a normative framework for approaching these issues. While our framework draws selectively on elements from utilitarian, deontological, and virtue-ethical traditions, it does so contextually, arguing that ethical evaluation must remain sensitive to the diverse ways in which AI intersects with cognition, ranging from the collective regulation of surveillance technologies to the individual cultivation of intellectual virtue in everyday AI use.

Taken together, this inquiry aims to clarify both what is at stake in the technological mediation of cognition and how we might navigate this transformation responsibly. Throughout, we distinguish restoration uses (therapeutic scaffolding back to a functional baseline) from augmentation uses (efficiency beyond baseline). This distinction is the ethical hinge: scaffolding returns users to practice; substitution replaces it.

What is Cognition

Understanding the ethical and social implications of AI's impact on human cognition requires first clarifying what is meant by cognition itself. The term is widely used but often inconsistently defined across disciplines such as philosophy, psychology, neuroscience, and computer science. Without a clear conceptual foundation, discussions about the enhancement or degradation of cognition risk becoming ambiguous. What constitutes cognition has been a subject for debate for millennia, and to date still lacks consensus of what it is, and how it relates to other terminology such as mind,

sentience, consciousness, emotions, intelligence, and the like. Needless to say, the debate will not be settled here, so this section is a succinct overview of the main contemporary paradigms of cognition, how we construe cognition for this essay, and what it means for humans and AI. This framing allows us to address how cognitive systems, whether biological or artificial, can interact, align, and potentially interfere with one another.

Informed by advances in computer science, cognitive science emerged as a discipline joined at the hip with the advent of artificial intelligence at Dartmouth college in 1968. The insight was that cognition, or the mind, could be likened to a computer program, where cognition was construed as the distinct informational processing intermediary between perception and action. This view was serial in nature, going from perception, to cognition, to action. The models of cognition were heavily inspired by computer science, as Hutchins put it: "The computer was not made in the image of a person. The computer was made in the image of the formal manipulations of abstract symbols. And the last 30 years of cognitive science can be seen as attempts to remake the person in the image of the computer" (Hutchins, 1995). However, cognition also became construed in the light of cognitive human psychology, where cognitive capacities of humans became the norm for cognitive capacities, which was also not without its problems (Cisek & Kalaska, 2010). Another strand of theorizing about cognition has emphasized the embodiment of the agent, and have stressed the dialectical relationship between the world and the body, and the affordances and restrictions the body imposes on the capacities for the agent (Clark, 2014; Keijzer, 2015).

For this essay however, we consider cognition in light of the Technological Approach to Minds Everywhere (TAME) framework (Levin, 2022) and the Free-Energy Principle (FEP) framework (Friston, 2010). In short, cognition here is to be understood as the capacity of an agent to exploit the structured variation in its problem space such that it may act in accordance with or thwart the actions of the environment in service of self-preservation. This view is agnostic with regards to 1) the substrates of the cognitive agent (e.g., carbon vs. silicon); 2) the phylogeny (how the cognitive agent evolved); 3) temporal scale (e.g., milliseconds vs. aeons); and 4) spatial scales (micrometers vs. kilometers). There are multitudes of theoretical implications for this, but there are three implications of pertinence here for human cognition.

In this view, cognition can be understood as a fundamental property of adaptive systems and a necessary foundation for more complex phenomena such as sentience and emotion. While not all cognitive agents are sentient, all agents capable of subjective experience must, at minimum, be cognitive. Sentience and affect thus supervene on cognition; they emerge from but do not exhaust its operations. Moreover, cognition is not a discrete stage separate from perception and action, as posited in classical serial models, but is deeply embedded within both. Perception and action are active, inferential processes that continuously attempt to align top-down predictions with bottom-up sensory signals. Within this dynamic, the agent selectively adjusts its attentional landscape by realizing (in both sense of the word, thus both becoming aware and enacting) what is relevant to its goals and ongoing interaction with the environment, a process described as relevance realization (Andersen, Miller, & Vervaeke, 2022). Herein lies the notion of embodiment, central to contemporary cognitive science, from which arise the interrelated ideas that cognition is embedded, enacted, and extended (Thompson, 2007; Varela, Thompson, & Rosch, 1991). Together, these perspectives emphasize that cognition is not confined to the brain but emerges through the dialectical relationship between the body and the environment. Finally, this account rejects the notion of human cognition as the benchmark for all cognitive capacities. Instead, cognition is conceived as a minimal and scalable property that can manifest in varying degrees of complexity across different substrates and temporal scales.

With these implications, it means that we have an account of cognition as minimal and fundamental. In relation to human cognition, this means that emotions are not in contrast to cognition as some theorizing would have it; rather they are a form of cognition. Changes in cognition need not necessarily affect how we think, but how we see the world. These changes can also occur at different levels of cognitive complexity, ranging from basic cognitive functions (e.g., attentional allocation) to meta-cognitive functions (e.g., self-reflection).

The Pre-LLM stage: Surveillance of Cognition as a Consequence of Capitalism

Before we assess the current wave of AI, dominated by generative systems such as LLMs, we first need the pre-LLM baseline: the platform era in which surveillance and recommendation engines began engineering our salience landscapes. What follows describes the architecture of attentional control that generative systems now plug into.

One way in which AI has been commonplace in everyday life before the advent of LLMs has been in the form of recommendation algorithms among popular social media sites and search engines, such as Facebook and Google. These companies found a means of transforming behavioural data outside of the market domain to a novel source of revenue. This was achieved to a large extent through the surreptitious collection of behavioural data, which later was commodified to stockholders in the form of predictions of behaviour. The natural evolution for the use of these predictions were to use them to affect the behaviour of people at scale. To do so, various psychological methods can be employed, such as nudging, operant conditioning, and curated selection and presentation of content (Zuboff, 2019). Regardless of employed method however, the application of these are done in the service and interest of the stakeholders of the financial institu-

tions that exert this power, rather than the individuals being targeted or society at large.

The central problem here is well characterized by Shoshana Zuboff: "... the essence of the exploitation here is the rendering of our lives as behavioural data for the sake of others' improved control over us" (Zuboff, 2019, p. 94). An illustrative example of this control is a study showing that Facebook were able to affect voter behaviour in the U.S during the 2010 election (Bond et al., 2012). This in effect means that the autonomy of individuals is, or at least can be, undermined at scale.

This form of control is enabled by platforms with extensive AI technologies and data resources that is aggregated to a few handful of stakeholders on the market. And in the digital age, power need no longer conform to the spatial constraints of location or resource constraints of gold for instance. Instead, information can be accrued and concentrated to a exponentially higher degree than has been the case previously. This means that the arbiters who exert control have a noticeably higher reach and influence than traditional institutions.

The evolution of information technology has seen a noticeable shift in the active parts of information dissemination, shifting the active part from the user to the disseminator. A library adhering to liberal norms will generally exert little control over the access to information for its users. Libraries will inevitably have to make choices concerning selection of books to keep and where to place them (forcing to place some books in more or less favourable positions, yielding a form of nudging), but these issues are ubiquitous for any aggregator of information (selection and sorting) and cannot be circumvented, which is the case for other institutions such as traditional news outlets as well.

However, with the advent of radios and televisions, the medium through which information is disseminated became noticeably more active. In a traditional library, people would themselves have to enter the library, select the books they have to read, and subsequently actively engage with the material to take part of it. With radios and televisions, people could be more passive, in that the information was given to them rather than them having to find it. This lent credence to the winged words "the medium is the message" (McLuhan & Lapham, 1995) in that it fundamentally changed how information retrieval functions in society. Information is broadcasted and curated, at least nowadays, every hour of every day, where the curators gained increased power to influence what people could see and not see.

Now, with the advent of AI algorithms in social media and commercially accessible LLMs, this dynamic has further shifted the action of information retrieval from the individual to the curators.

The Current State of the Art and Trajectory

Against this backdrop, the current wave of AI dominated by generative systems like LLMs do not arrive on neutral ground. They slot into an already optimized attention economy and shift the locus of control from one-way curation to interactive co-production. The “current state” is thus an escalation: the same incentives and data flows, but with tools that talk back intelligently and personalize at dialog speed. The proliferation of generative models marks what might be called a Kairos moment, a juncture of both opportunity and peril. What was once a domain of experimental computation is now a central mediator of communication, knowledge production, and decision-making. Among its recent developments, large language models (LLMs), such as GPT-5, Claude, and Gemini, represent a significant step in the evolution of machine intelligence. These systems, built upon transformer architectures containing hundreds of billions or even trillions of parameters, exhibit capacities that appear strikingly general: they can reason, converse, generate code, and emulate creativity. Technologically, the breakthrough is the apparent solution to the 'silo problem'. We’ve moved from single-domain, specialized AI (e.g., the Go-player) to more general problem-solvers via hybrid machines that couple deep neural networks with language-like processing (the “language of thought” idea). This is a qualitative shift toward Artificial General Intelligence (AGI). Their performance has led some researchers to suggest that they display forms of “emergent” general intelligence (Bubeck et al., 2023).

Despite their impressive functionality, the mechanisms underlying such capabilities remain only partially understood. The dominant paradigm of AI development is scaling, increasing the size of models, data, and computational resources to improve performance. They are, although impressive, still technically a form of narrow AI, excelling at pattern recognition and data compression, not genuine, self-directed learning and autonomy outside their training parameters. While this has yielded rapid progress, it has also produced a widening gap between technological capability and scientific understanding. The success of LLMs is driven less by theoretical advances in the nature of intelligence and more by empirical optimization through massive datasets, reinforcement feedback and improvements in hardware. The result is an engineering triumph that has, paradoxically, deepened philosophical and epistemic uncertainty about what such systems are actually doing.

This decoupling of technological progress from scientific insight has ethical as well as epistemic consequences. AI research, once closely aligned with cognitive science, increasingly follows the imperatives of capital investment and market competition rather than the pursuit of understanding. Public discourse, as a result, tends to oscillate between three unproductive extremes: technological utopianism, radical skepticism, and apocalyptic fatalism. Between these poles lies the more pressing question, how these technologies, already embedded in our social and cognitive ecologies, are transforming the ways humans perceive, reason, and act.

The social implications of this transformation are profound. LLMs and similar systems increasingly

automate forms of labor once associated with human intellect, such as writing, summarizing, programming, and advising, thereby altering both the value and meaning of cognitive work. As such, the rise of AI compels us to rethink the very computational, propositional core of human identity as defined by modernity. As the "price of inference" decreases, the economic and existential worth of certain human capacities is called into question. Historically, such shifts have redefined identity and agency: the first "computers" were human clerks, whose disappearance marked a turning point in how intelligence was understood and valued. Today's automation of reasoning and expression suggests a similar threshold, one that compels renewed reflection on what constitutes human cognitive agency in a world of synthetic co-thinkers.

From a systems-theoretical perspective, indefinite exponential growth is unlikely for complex adaptive systems. Whether biological or technological, they encounter intrinsic limits as coordination and energetic demands rise, performance faces diminishing returns and reorganizations. They face general system collapse due to the exponential growth of inner complicatedness needed to manage real-world, inexhaustible complexity and genuine uncertainty/emergence (Shumailov et al., 2023). Scaling alone cannot bypass these constraints; instead, it tends to produce emergent trade-offs, feedback instabilities, and phases of reorganization. In AI specifically, training on model-generated data can degrade distributional diversity and performance

This phenomenon of AI model collapse further illustrates a structural vulnerability in current trajectories. As models are increasingly trained on data partially generated by other AIs, the informational diversity of their training corpora diminishes (Shumailov et al., 2023). This recursive loop amplifies statistical regularities while erasing rare or anomalous patterns, leading to homogenization and a gradual erosion of epistemic richness. Because human cognition now co-evolves with these systems, drawing from the same digital knowledge ecosystems, the effects of such informational narrowing may reverberate across both machine and human domains, producing a shared drift toward mediocrity and self-reference.

Seen through the lens of cognitive science, these developments signify not merely a quantitative extension of human tools but a qualitative shift in the structure of cognition. If, as previously discussed, cognition is the capacity to enact and realize relevance, then the increasing mediation of this process by artificial systems represents a transformation in the very conditions of sense-making. By proposing, filtering, and completing meaning, AI systems participate in the determination of what counts as relevant. This participation alters the boundary between use and understanding: tools that once extended human thought now reciprocally shape it. Artificial intelligence thus occupies an ambiguous role within the human cognitive ecology, both an extension and a co-author of cognition. It enables new forms of creativity and reasoning, yet it also risks attenuating the very capacities it augments. Understanding this tension is essential for any ethical evaluation of AI. The challenge lies not in forecasting whether these systems will surpass human intelligence but in discerning how their integration is already reshaping the distribution of attention, agency, and meaning across our shared cognitive landscape.

At this juncture, a deeper question arises: when does a difference in degree become a difference in kind? The Sorites paradox reminds us that continuity can conceal rupture where for example a heap of sand losing grains remains a heap until, imperceptibly, it is not. Likewise, while there is a historical continuum in the way technologies extend human capacities, for instance writing extended memory, electricity extended muscle, computation extended calculation, there are thresholds where quantitative scaling yields qualitative transformation. For AI, it can be argued that we've reached, or are fast approaching such a point. These systems no longer merely amplify symbolic reasoning; they instantiate autonomous, self-modifying processes that act back upon the human cognitive ecosystem. What was once an instrument of thought now participates in the very conditions of thinking. That moves us into new ontological terrain: the virtual (informational domains shaping material life), the artificial (constitutive rather than opposed to the "real"), and the cyborgic (human–machine agency fused in practice). The qualitative novelty is the feedback: tools that talk back and reshape our cognitive loops.

When artifacts become co-participants in cognition (extended cognition), not mere tools, we inhabit a different mode of being. The line between subject and object, use and understanding, begins to blur. These systems might do more than simply extend agency; they restructure it. When they predict, recommend, and converse, they alter the feedback loops through which humans form beliefs, coordinate action, and sustain meaning. That feedback, that is the tools that talk back, and do so intelligently, is the mark of the difference in kind. Hence AI systems are not just another rung on the ladder of tools like the wheel or the printing press; they belong to a qualitatively new category: reciprocal intelligences within the human cultural system.

Yet this new reciprocity conceals an asymmetry. Despite their fluency, LLMs do not care about the information nor truth in any technical or existential sense. As previously presented, genuine cognition involves the process of relevance realization defined as the dynamic realization (again in both sense of the word) of what matters for one's continued existence. Humans perform this because we are autopoietic (Maturana & Varela, 1980/1987) beings: self-organizing systems that must care for our own survival and coherence. We have to care for and take in this information or that information, and then that information literally becomes part of us. Machines, by contrast, are not autopoietic; they do not generate their own goals or sustain their own being. They can simulate caring, but they do not enact it. Their apparent understanding of meaning is derivative of human structures of relevance that we have encoded into data, language, and reinforcement signals.

This distinction reveals why their "knowledge" remains indifferent, for now. AI models operate on statistical relevance, not epistemic or existential relevance. They process correlations sans any intrinsic orientation toward truth, falsity, or care. What feels like understanding is a pantomime of cognition, one that can exploit our phenomenology of sense-making. Without prudence we risk mistaking their fluency for wisdom and in outsourcing judgment to systems that cannot care, we risk becoming instruments of our own instruments. Recognizing this difference is therefore not grounds for dismissal but for prudent discernment.

Cyborgs: Human–Machine Hybrids

If AI systems now co-author salience and transform our digital knowledge ecology, thus fully realizing the virtual and the artificial as new ontological grounds, the next question is what happens when that co-author moves closer to (or inside) the body, culminating in the cyborgic mode of being. The current wave of AI is not merely external; it is actively shaping the cyborg identity, where the boundary between human and machine agency blurs in identity and practice.

Before the emergence of LLMs, the vision of human-AI merger often revolved around physical enhancement. One might have imagined AI would merge with humans to improve certain attributes inherent to our species, such as enhanced perception (hearing, vision) or enhanced information processing (logical reasoning, overall processing speed). The method by which these enhancements were thought to be actualized is through external sci-fi-like gadgets or surgical interventions. Such is the idea behind Neuralink (Musk & Neuralink, 2019): to install a digital layer above the cortex with the objective of serving as a silicon extension of our carbon-based brain. This shift in embodiment is what defines the emergent cyborgic condition, demanding a new ethical analysis that moves beyond simple tool-use.

Perhaps it is the case that as an organism is gifted with more neurons, it becomes more sophisticated. It seems that one of the things that differentiates us from fruit flies is the number of neurons in our respective brains. It is important to note that a species' overall intelligence is not determined by the raw number of neurons alone, but more critically by the ratio of neurons to the organism's body size. A significant portion of any brain's resources is dedicated to basic somatic and motor control; a higher ratio therefore suggests more neurons are available for complex, higher-order processing. It is theorized that as this available neural mass increases, novel and more sophisticated emergent cognitive properties can occur. This implies a case of biological opportunism, given that there are more neurons available the system will eventually utilize them for what is deemed most essential. Such is the case in individuals who are blind from birth, these people's visual cortex does not lie dormant even if the eyes are not being brought into service, instead the available area is repurposed to enhance other senses, such as the processing of hearing sharpening it for spatial localization.

To properly analyze the moral challenges of the cyborgic transition, it is necessary to distinguish between two variations of human-machine integration. Direct Artificial Neural Enhancement (Hardware Integration) refers to the direct integration of non-biological components into the brain's neural architecture, with the objective of fundamentally augmenting the brain's intrinsic processing capabilities, such as memory, calculation, or perception. This is analogous to a hardware upgrade, where the core processing capacity of the biological system itself is increased. The ethical stakes here center on identity, agency, and access and raise questions of cognitive equality in a potential post-biological world. A surgical Brain-Computer Interface (BCI) like Neuralink offers an example of a difficult to define enhancement tool. While the implant is a form of hardware integration, its primary function today is to act as an interface. It facilitates the use of external software, such as LLMs and other apps, as cognitive assistance tools. The chip creates a pathway for these software add-ons but does not yet fundamentally alter the brain's biological processing architecture. Since its main effect is to enable faster, more seamless cognitive assistance and offloading, it is best classified on the software side of the current cognitive atrophy debate.

In contrast, cognitive assistance & offloading (software supplementation) involves the delegation of cognitive processes to an external tool or system (e.g., an LLM, a smart device). The agent's brain utilizes an external resource to achieve a cognitive outcome more efficiently. This is analogous to running a software application; the underlying biological hardware is not altered, but its function is temporarily supplemented by an effective external program. This is the primary vector for cognitive atrophy, as it modifies competence development. These are the 'Hutchins-like' cognitive artifacts onto which the brain offloads processes as a means for metabolic efficiency, which has been theorized as extended cognition in different guises (Chalmers & Clark, 1998; Hutchins, 1995). While these devices provide a clear utility, the cognition they contain is not integral to the agent. It is a pre-packaged, external resource, a form of reactive memory, rather than an expansion of the agent's own deliberative capacities, effectively being reminiscent of stigmergy (Heylighen, 2016). This is why the LLM's simulated caring, discussed in the previous section, is a key risk: it provides a fluent answer (lowering prediction error) without demanding the necessary internal relevance realization from the user. A distinction must be made, however. The skilled artist's pencil or the hockey player's stick, through sustained practice, become deeply integrated into the user's sensorimotor loops, functioning as genuine extensions of the body. Such tools extend the agent's competence; they do not supplant it.

On the Detrimental Effects of Cognitive Delegation

For Aristotle, the highest human good is eudaimonia (Aristotle, ca. 350 B.C.E./1999, bk. 1, ch. 7), a term

often translated as 'flourishing'. This good is not a passive state of mind, but is achieved through virtuous activity. This classical conception of well-being has a modern analogue in positive psychology frameworks such as Self-Determination Theory (SDT) (Ryan & Deci, 2000), which offers an empirical account of the psychological needs that are constitutive of a flourishing life.

Let us consider what makes our lives go well. SDT claims that a life is better to the extent that it involves the fulfillment of certain fundamental needs. These are: (1) Autonomy, the exercise of self-governance; (2) Competence, the possession and development of abilities; and (3) Relatedness, the formation of meaningful bonds with others.

We can now ask whether the widespread use of certain new technologies, such as LLMs, is likely to improve or worsen our lives based on these fundamental needs. We argue that the use of these systems will, if used imprudently, be against our own self-interest. This is because their function has the capacity to replace the very human activities that are required to fulfill all three of these needs. The result is that what may seem to make our lives easier in the short term makes them, in ways that matter more, worse.

The Diminishment of Autonomy

What it is to be autonomous is, in part, for our actions to be the result of our own intentions, actions, and values. It is to be the author of our own choices. The extensive use of LLMs has the capacity to undermine this. When a person has the intention of forming a plan or writing an argument, it is possible to use an LLM as an assistive tool, thus requiring minimal cognitive effort. The cognitive work that constitutes the formation of the intention is then not performed by that person, or not at least in any meaningful way. Instead, the person is presented with a finished product generated by an external process. The person's role is reduced from that of a creator to that of a curator, who merely endorses or rejects the system's output.

The grounds for the action are not the agent's own. The connection between the agent's deliberative faculties and the final product is severed. Over time, a life in which many of one's actions have this character is a life with substantially less autonomy. This diminishment can be expected to have negative psychological effects. The resulting loss of well-being makes such an outcome undesirable on several moral views. For a utilitarian, it is a net loss of welfare. For a virtue ethicist, it is a failure to exercise the capacities that are constitutive of the good life.

This is not to claim, however, that every use of such a tool is impermissible. The conditions under which its use might be justified are complex. An LLM can have as much positive impact on one's autonomy as it can have negative. A simple rule that either forbids all use or permits all use would be crude. The permissibility of the act plausibly depends on the context, the agent's intention, and the nature of the task. As with prior technologies, we can expect that norms governing the proper use of these tools will need to be developed and refined over time (though we probably ought to do so with a liberal framework).

Erosion of Competence

What we call competence is the possession of abilities that allow us to achieve our aims effectively. Such abilities are not innate; they are causally dependent on a process of effortful practice. To become a skilled writer or programmer, one must repeatedly engage in the difficult act of writing or programming. If you never do things you cannot do, you will never be more than you currently are.

LLMs are designed to remove this necessity by trying to predict our needs and output the exact outcome we desire. They provide a means to an end by supplying us with a finished text, a working block of code, that bypasses the skill-forming process.

The effect is not merely a subjective feeling of incompetence. By consistently avoiding the activities necessary for skill acquisition, a person is at risk of becoming less competent than they otherwise would have been. This creates a state of dependency, where the inability to perform a task without the system gives the person further reason to continue using it, leading to a positive feedback loop of an even greater deficit in their abilities and greater dependency on these systems.

A distinction must be made, however. It is plausible that bypassing certain processes could be beneficial. If a task is merely tedious and does not contribute to the development of valuable skills, delegating it to a system could free a person's cognitive resources for other activities that are more conducive to developing competence or other forms of well-being. This must be distinguished from the circumvention of "desirable difficulties" since those challenges contribute to the development of the abilities that are an essential component of a flourishing life (Bjork, 1994; Kapur, 2008). This further emphasises the need for a nuanced approach of when and how to use LLMs, rather than mindlessly dealing in absolutes on either end of the spectrum.

The Impoverishment of Interpersonal Relations

The value of our relationships with others, or relatedness, is the need to form meaningful bonds with other people. This requires a belief that we are interacting with another person's mind, meaning; their actual thoughts, feelings, and intentions, and that in turn becomes the ground upon which authentic communication rests on.

The use of LLMs presents a dual threat to this need. First, it degrades the quality of human interaction by replacing authentic expression with strategic performance. Consider a person using a system to craft a response in a difficult professional negotiation. Here, the goal is not to engage in a sincere exchange of reasons but to produce an output that is maximally persuasive.

This practice is objectionable on Kantian grounds, as it treats the other person merely as a means to an end. By circumventing authentic communication in favor of an optimized rhetorical strategy, by effectively becoming a 'bullshit artist' (Frankfurt, 2005), one fails to respect the other person as an end in themselves. The relationship is instrumentalized, and this corrupts the grounds for trust. If we suspect another's words are AI-generated artifacts, such as coming from integrated mail writing-assistance, we can no longer be confident that we are in a genuine relationship with that person.

A further issue arises when the system ceases to be a mere intermediary and becomes the substitute for human connection. An individual might turn to an LLM to discuss personal or psychological problems, seeking the connection that they would otherwise seek from other people. After all, LLMs are designed to be pro-social, are highly accessible, endlessly patient, and are sequestered from one's social life, thus reducing the risk of inadvertent gossip. The LLM may effectively simulate understanding and care, but this interaction is still asymmetrical. The need for relatedness and the ability to be connected and care for others, cannot, in principle, be fulfilled by a non-conscious artifact. Widespread use of such systems could therefore lead to a world in which our interactions are hollow and our genuine human relationships are increasingly displaced by simulated ones, leading to a sense of isolation among the population.

There is however antithetical empirical evidence for this claim. Spytska (2025) found that while human therapists were more effective at reducing anxiety, an AI chatbot also produced a significant reduction in symptoms for individuals in crisis situations where professional care was inaccessible. The utility of such systems, therefore, may not lie in their capacity to replicate human connection, but in their scalability and immediate availability. This suggests that the correct application of this technology may be a hybrid model, where AI systems supplement human care (Spytska, 2025). In such a model, the system could serve as an initial point of contact or a supplementary tool, rather than as a complete substitute for the therapeutic relationship. Such an approach is consistent with a virtue-ethical framework, in which the virtuous application of a tool is to find a mean between two extremes: the vice of rejecting a potentially beneficial instrument, and the vice of an excessive reliance that undermines the human capacities essential for a flourishing life. This mean is not absolute, but is relative to the particulars of the situation, requiring its application at the right time and to the appropriate degree.

By handing the keys to our mind to an LLM or otherwise, the over-reliance on the interaction; does not allow for us to go through that difficult transformation process, and by avoiding the transformational virtuous work, we also give up the tools that result in eudaimonia. From an Aristotelian perspective we avoid fulfilling certain virtues. From a utilitarian point of view we do not perform the actions that lead to the maximum amount of happiness. The core dilemma, therefore, is how to reap the immense benefits of AI without sacrificing the underlying effort, skills, and authenticity that constitute a meaningful, flourishing human life.

The Principle of Cognitive Efficiency and the Risk of Competence Erosion

The principle of cognitive efficiency provides the fundamental mechanism through which cognitive atrophy operates. As biological agents, we are heavily constrained by our metabolism; our actions and cognitive processing are always balanced against our energy expenditure, a consequence of which is that superfluous energy drains tend to be weeded out for efficiency. Frameworks such as the FEP view the brain as a prediction machine that seeks to minimize prediction errors, or "surprises," in the most metabolically efficient way possible (Friston, 2010). If an external cognitive tool or shortcut can resolve this error more efficiently—for instance, an LLM generating a summary or an answer—the brain, in its effort to save energy, will naturally take that path.

This creates a direct tension: the path to convenience offered by cognitive offloading is also the path away from the development of genuine competence. This is because genuine cognitive development, analogous to biological processes such as memory consolidation where muscles atrophy without activation, requires active, effortful processing. Relatedly, work on causal emergence and learning suggests organized, multi-scale structure emerges through active learning and model building (Hoel, 2017), aligning with the idea that effortful practice is not a dispensable cost but a driver of cognitive development. This lends credence to effort being a necessary component for cognitive development, since learning is an inherently active process of ascertaining, assimilating, and integrating new information into the inner knowledge of the self. Therefore, extensive outsourcing of various tasks to AI may well stall, if not outright halt, learning opportunities and skill acquisition.

Here is where the core ethical dilemma is enacted: For learning and competence development, the journey is quintessentially the destination. The teaching of mathematics, for example, is not done merely to ensure everyone knows specific geometrical relations by heart, but rather to foster general numerical literacy and a baseline competence in logical reasoning. Likewise, an LLM may be able to give an adequate summary of some corpus of text, but the person using the LLM will in the process deprive themselves of the opportunity to maintain or improve the attentional demands and text comprehension that reading requires. This discrepancy, between having more accessible interpretations of difficult material, may, on first glance, provide a more equitable epistemic landscape for people at large. However, it does so at the trade-off that fewer individuals may be able to penetrate and adequately understand the original material, thereby increasing our systemic dependence on these AI systems. This is the "no free lunch" constraint on skill acquisition: the short-circuit-

ing of effort, while metabolically efficient, short-circuits the very mechanism of cognitive growth, leading directly to the decline in quality and structural dependency that defines cognitive atrophy.

Restoration vs. Augmentation: An Ethical Branching Point

To arrive at a normative judgment about the use of these tools, a further distinction is required regarding the purpose for which they are used. Restoration (therapeutic) applications involve utilizing AI to bring an individual facing a cognitive deficit up to a functional baseline. The goal is to establish or regain the foundation for competence, meaning the tool acts as a necessary enabler to engage with the world as the individual would have been able to do without whatever deficit is plaguing the individual. This stands in contrast to augmentation (enhancement) use, which concerns the use of AI for "abundance," where the goal is to augment the abilities of a given individual beyond the functional baseline. The technology is therefore not compensating for a defect but is instead used to produce a better outcome such as a more efficient piece of code or a faster analysis. The long-term impact of these technologies depends on whether they act as a means to enable a user to engage in a difficult cognitive process, or as a means to the contrary, meaning to allow for a user them to escape the difficulties in a variety of cognitive process entirely.

At the software level of the cyborgic spectrum, the core ethical challenge is the widespread use of cognitive assistance & offloading for augmentation purposes. When LLMs are deployed at mass scale to produce speed and efficiency, they accelerate the systemic cognitive atrophy of public discourse and individual skills. A therapeutic application, when well-designed, is a cognitive prosthetic; an augmentation application, when improperly used, is a cognitive crutch that leads to atrophy. Meanwhile, the hardware level, direct artificial neural enhancement, presents a distinct, though equally vital, set of ethical challenges related to augmentation. This involves genuinely augmenting the brain's processing capacity—for instance, achieving supra-human memory or reasoning speed. If access to these enhancements is dictated by capital, society risks the emergence of a biologically and cognitively superior class, raising difficult questions of cognitive equality and social stratification among the nascent cyborgs.

Ethical Perspectives, Stakeholders, Harms/Benefits and Possible Resolutions

Ethical Perspectives: A Convergence

It is tempting to cast the major ethical traditions as rivals here, yet when we slow down and follow the problem where it leads, they begin to lean in the same direction. Take the utilitarian first. The core concern is the long-run welfare of humanity. A short-term calculus celebrates the efficiency, speed, and convenience offered by AI. However, true human welfare depends on capacities such as stable attention, rationality, trustworthy judgment, embodied skill, and deep relationships. If routine delegation to AI, in the pursuit of immediate efficiency, erodes these fundamental cognitive and relational stocks, then the system ultimately fails its utilitarian aim. The question is: Does an efficiency today lead to a "thinner attention" or "less resilient judgment" tomorrow? If so, the long-term utility is negative.

A deontologist, coming from a different angle, will emphasize duty and autonomy, insisting that individuals must never be treated merely as raw material for optimization (the Categorical Imperative, Kant, 1785/1997). Therefore, any system that profoundly shapes human perception and action (i.e., our salience landscape) must owe its users intelligibility, genuine consent, and principled limits on manipulation. Given the steep asymmetry of power between the user and the arbiters of these technologies (e.g., developers and corporations), a fundamental duty exists to ensure the systems do not covertly undermine or coerce free agency of their users.

The virtue-ethical perspective raises a complementary doubt in its emphasis on character and agency. This approach focuses on the agent rather than the outcome or the rule. It asks whether the ordinary use of these systems cultivates or withers the virtues (such as effort, patience, truthfulness, and honest interpersonal relations) upon which long-run human flourishing depends. Even if outcomes are good and procedures are followed, if the easiest path through our tools demands swimming against the current of intellectual effort and genuine engagement, the technology risks creating agents less capable of living a good life.

From within these different perspectives, a shared hinge appears. The moral weight of delegation does not turn mainly on the sophistication of the system but on what in us its use displaces or sustains. When assistance functions as scaffolding, lowering barriers while returning us to the work, the perspectives have fewer conflicts. When it functions as substitution, quietly removing the activity by which cognitive capacities are formed, they converge in their unease, albeit for different reasons.

Stakeholders, Harms, and Benefits

To understand the full scope of this challenge, we must map the stakeholders implicated by the movement from individual cognition to distributed cognition. While the individual user is the first stakeholder, whose moment-to-moment autonomy is immediately affected by the scaffolding-versus-substitution dynamic, the systemic impact is mediated through professional and institutional layers.

The next layer of stakeholders can be thought to be individuals occupying seats of authority in different professions. Here is just a snapshot: Educators and Professors are important stakeholders because they will need to adapt to the inevitable use of the systems and the new academic landscape LLMs bring about. In their

overseeing of the acquisition of skill and critical thinking for example, they can decide to integrate an LLM as a scaffolding tool (a mechanism that supports the student in reaching a level of competence they could not otherwise achieve) or as a substitution technology (a mechanism that bypasses the effortful practice necessary for the formation of durable competence). Without a (somewhat) high understanding of how the systems operate and the ramifications of its use we could see a degradation of the educational system. Clinicians and therapists are also stakeholders on the front lines of mental and relational health. As LLMs become integrated into communication and emotional processing (e.g., as conversational interfaces, synthetic companions, or diagnostic aids), clinicians must wrestle with the implications for their patients' capacity for genuine human connection, reality testing, and self-definition. The risk is that simulated care might hollow out authentic relation, substituting fluency for existential depth.

Beyond these mediating professions, the stakeholder map scales to those who control the environment: developers who define system defaults; regulators who define the acceptable bounds of influence; the public which inherits the resulting epistemic baseline; and, critically, future generations who will live with our current compromises as if they were natural constraints. The benefits of this new technology are genuine (access at the point of need, speed of iteration, creative amplification) but the systemic harms are equally real: erosion of autonomy, atrophy of skill and competence, decreasing in genuine interpersonal relations and, as explored with model collapse, the homogenization of collective knowledge via self-referential training.

And finally, the distributed cognitions of institutions or even society at large are stakeholders in their own right, representing the collective architecture of knowledge, meaning-making, and coordinated action. A university, a hospital network, a newsroom, a court, a standards body: each is more than a collection of individuals. Each is a cognitive architecture with memory (archives, records), inference procedures (peer review, clinical guidelines, legal reasoning), and action loops (publication, treatment, rulings, protocols). What these systems "notice," how they "update," and which outputs they authorize are shaped by tools, incentives, and norms just as surely as a person's attention and judgment are. When assistance at this level becomes substitution, automated summaries in place of reading committees, synthetic findings in place of experiments, policy generated from dashboards without deliberation, the result is not a faster institution but a shallower one. The stock of institutional competence is drawn down even as the flow of decisions increases. Society at large is the widest version of this claim. A public sphere has shared memory (libraries, media, law), shared procedures for settling claims (science, journalism, elections), and shared habits for disagreeing productively. These are the scaffolds by which a people learns in public. If AI-mediated systems compress plural signals into median patterns, train on their own exhaust, and hide the routes by which conclusions are reached, the culture can become synchronized and quick while losing the diversity and contestation on which correction depends. A society can feel more informed while becoming less capable of truth-tracking, because the practices that generate epistemic resilience—provenance, plurality, process—have been quietly replaced by convenience.

Resolutions

It is the risk of these cascading, systemic harms with regards to the intellectual, relational, and epistemic that makes prudence mandatory. The choice is not between using AI and shunning it; it is between arranging our collective cognitive systems so that abundance returns to practice or quietly replaces it. Returning abundance means designing tools and institutions that make provenance visible and portable, reward process evidence alongside outcomes, test for homogenization and self-reference, and cultivate multiple interoperable infrastructures rather than a single stack that sets defaults for everyone else. Replacement, by contrast, is what happens when speed and engagement become the master metrics and opacity is treated as a feature rather than a cost. Without sustained, interdisciplinary attention (scientific, philosophical, anthropological), the default path concentrates inference power in a few hands and substitutes algorithmic convenience for human agency.

What follows, then, is not a final judgment on the nature of machines, but a necessary reconfiguration of our duties to one another in their presence. If human flourishing is predicated on effortful autonomy, durable competence, and genuine relation, then the surrounding digital environment must be deliberately shaped so that abundance returns to practice rather than replacing it.

The technical architecture should default to scaffolding, not substitution. Systems should not merely deliver results; they should expose the route by which results are reached and encourage curiosity whenever applicable. Design for practice and provenance means outputs come with inspectable lineage (sources, transformations, constraints) and with metacognitive affordances (hints, exemplars, critique) that return the user to the work. When process is made visible and portable across contexts, AI stops being a black-box oracle and becomes a cognitive partner: it supports learning, enables verification, and lets the user integrate reasons not just consume conclusions of varying arbitrariness. This is how convenience becomes competence rather than erosion.

At the institutional layer, shallow acceleration must be resisted by rewarding process over outcome. Committees, courts, clinics, and newsrooms should audit not only what was decided but how it was known: what evidence was considered, how dissent was handled, how uncertainty was communicated, and whether automated subsystems narrowed signal diversity. Audits should explicitly test for homogenization and self-reference, warning signs that an institution is training on its own exhaust. By valuing process evidence in official review and incentives, institutions preserve

memory, sustain inference quality, and keep the "muscles" of collective judgment in use.

These architectural and institutional shifts imply a duty of care for agency. Designers have a duty to make the helpful route also the formative route; educators and clinicians to use systems as scaffolds that return people to practice; regulators to prohibit covert manipulation and require intelligibility proportionate to stakes; institutions to evidence their processes, not only their products; citizens to prefer tools and policies that keep human capacities in the loop. The test is simple and Socratic: What does ordinary use make of us? If a design makes people faster but shallower, it fails even if it "works."

Conclusion

The essay began with presenting the ethical problem of a socio-technical shift in which tools designed to help us think now participate in the conditions under which thinking occurs. The central question was not whether artificial systems surpass human intelligence, but how their widespread use reconfigures the structure and quality of human cognition, and how we ought to relate to such systems. The answer required a clear account of cognition as relevance realization within an embodied, predictive agent; a sober account of the pre-LLM platform regime that already managed attention at scale; and a careful reading of the generative turn as an escalation from one-way curation to interactive co-production of meaning. Along the way we identified both the promise and the danger: systems that widen access and speed iteration can also erode the very capacities by which a life becomes one's own, such as autonomy, competence, and genuine social relations.

References

Andersen, B. P., Miller, M., & Vervaeke, J. (2022). Predictive processing and relevance realization: Exploring convergent solutions to the frame problem. Phenomenology and the Cognitive Sciences. (Advance online publication).

Aristotle. (1999). Nicomachean Ethics (T. Irwin, Trans., 2nd ed.). Hackett.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), Metacognition: Knowing about knowing (pp. 185–205). MIT Press.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. Nature, 489(7415), 295–298.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., … Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712.

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

Cisek, P., & Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. Annual Review of Neuroscience, 33, 269–298.

Clark, A. (2014). Mindware: An introduction to the philosophy of cognitive science. Oxford University Press.

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering CBT to college students with symptoms of depression and anxiety using a conversational agent (Woebot): A randomized controlled trial. JMIR Mental Health, 4(2), e19.

Frankfurt, H. G. (2005). On bullshit. Princeton University Press.

Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138.

Heylighen, F. (2016). Stigmergy as a universal coordination mechanism I: Definition and components. Cognitive Systems Research, 38, 4–13.

Hoel, E. (2017). When the map is better than the territory. Proceedings of the National Academy of Sciences, 114(30), 7851–7856.

Hutchins, E. (1995). Cognition in the wild. MIT Press.

Kant, I. (1997). Groundwork of the metaphysics of morals (M. Gregor, Trans.). Cambridge University Press. (Original work published 1785)

Kapur, M. (2008). Productive failure. Cognition and Instruction, 26(3), 379–424.

Keijzer, F. (2015). Moving and sensing without input and output: Early nervous systems and the origins of the animal sensorimotor organization. Biology & Philosophy, 30(3), 311–331. https://doi.org/10.1007/s10539-015-9483-1

Levin, M. (2022). Technological approach to mind everywhere: An experimentally grounded framework for understanding diverse bodies and minds. Frontiers in Systems Neuroscience, 16, Article 768201.

McLuhan, M., & Lapham, L. H. (1995). Understanding media: The extensions of man. MIT Press. (1995 edition with Lapham introduction; originally published 1964.)

Maturana, H. R., & Varela, F. J. (1980). Autopoiesis and cognition: The realization of the living. D. Reidel. (Alternatively: Maturana & Varela, 1987, The tree of knowledge, Shambhala.)

Musk, E., & Neuralink. (2019). An integrated brain–machine interface platform with thousands of channels. Journal of Medical Internet Research, 21(10), e16194.

Ryan, R. M., & Deci, E. L. (2000). Self-Determination Theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, 55(1), 68–78.

Shumailov, I., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). The curse of recursion: Training on generated data makes models forget. arXiv:2305.17493.

Spytska, L. (2025). The use of artificial intelligence in psychotherapy: Development of intelligent therapeutic systems. BMC Psychology, 13(1), 175.

Thompson, E. (2007). Mind in life: Biology, phenomenology, and the sciences of mind. Harvard University Press.
Varela, F. J., Thompson, E., & Rosch, E. (1991). The embodied mind: Cognitive science and human experience. MIT Press.
Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.