Article

Not peer-reviewed version

The Evolutionary Framework for Human-Machine Alignment

Gleb Vzorin *

Posted Date: 17 September 2024

doi: 10.20944/preprints202409.1250.v1

Keywords: Superalignment; Human-AI co-evolution; HCI; machine culture

Article

The Evolutionary Framework for Human-Machine Alignment

Gleb D. Vzorin

g.vzorin@mail.ru

Abstract: The rapid development of advanced AI technologies poses serious existential threats and necessitates research on safe and responsible Human-AI alignment. We hypothesize that the evolution of AI may be viewed not as something accidental or artificial, but as a natural part of the evolution of intelligence in living species. This perspective suggests a different view of the Human-AI relationship, where AI is not seen as a threat or a mere enhancement of humans, but as a mean of revealing and establishing true human nature — achieving greater freedom and self-determination. While in previous evolutionary stages, more freedom was attained through inculturation, AI can serve as a mean in the next stage by providing a direct connection between individuals and accumulated culture, overcoming the limitations of traditional cultural acquisition processes and giving rise to post-human free beings. A specific step-by-step framework rooted in cognitive evolution theories is proposed, along with potential directions for empirical research.

Keywords: superalignment; human-AI co-evolution; HCI; machine culture

---

Contents

 doi:10.20944/preprints202409.1250.v

Introduction

Recent advancements in artificial intelligence (AI) systems have significantly expanded the scope of human-machine interaction, turning it into a field of extensive multidisciplinary research with increasingly specialized topics (Dwivedi et al., 2023; Nah et al., 2023; Seeber et al., 2020). Often, the enhanced predictive power of theoretical models is achieved by deliberately narrowing their scope of application and anchoring them in specific business and socially significant contexts. This approach, however, may lead to a shortfall in general theorizing on human-AI complementarity (Hemmer et al., 2024). While these models are effective locally, we contend that the unprecedented pace of AI development might render them obsolete quickly. Therefore, a theoretical framework is required that can make predictions about the distant future, even in contexts that currently appear as science fiction. This is essential because many of today's technologies and transformations, which once seemed far-fetched, are now becoming reality.

We argue that such a model should be abstract enough to address what we consider existential problems, yet it must consistently engage with concrete empirical mechanisms to remain scientifically valid. The aim of this paper is to introduce a preliminary concept of such a theoretical framework, which is grounded in evolutionary analysis.

We begin by addressing the general threat of reduced human agency, which highlights the inherent limitations of purely empirical models. Following this, we introduce an evolutionary perspective and establish the fundamental assumptions required. Building on these assumptions, we attempt to predict the specific outcomes that these evolutionary processes may lead to. Finally, we explore further directions and potential areas for empirical testing.

The Problem: AI and Human Agency

Humanity has a long history of technological innovation, with each significant advancement initially met with strong resistance due to socioeconomic fears of job loss and existential concerns over the erosion of human agency (see Dusek, 2006). Despite ongoing debates about their potentially detrimental effects on civilization, technologies typically integrate into society, eventually serving the social good by becoming increasingly sophisticated tools for human use. For instance, despite the Luddite movement, 19th-century industrial machines found their niche in performing tasks beyond human physical capabilities, albeit under human intellectual direction. Later, the advent of computers marked the next evolutionary step, capable of executing intellectual operations that, while predefined and straightforward, were largely quantitative. The qualitative aspects of decision-making, however, remained uniquely human. The development of AI has somewhat altered this dynamic. Although artificial narrow intelligence can outperform humans in specific tasks like chess or AlphaGo, it still significantly differs from genuine human psychological functions (Voiskounsky, 2013) and generally, humans surpass such AI in many areas (Gigerenzer, 2022).

While on stages described above it was clear how to distinguish between human and machine areas of responsibility, current AI inventions seem to dissolve this edge. Modern AI systems are now endowed with human-like cognitive capabilities such as autonomous decision-making, reasoning, active interaction, and situational awareness (Bubeck et al., 2023; Haenlein & Kaplan, 2019). This

 doi:10.20944/preprints202409.1250.v

evolution is transforming human-machine collaboration from a human-centric "master-slave" model to a "peer-to-peer" framework, where machines are viewed not merely as tools, but as decision-making partners (Ren et al., 2023).

These developments once again raise concerns about human agency, this time with more urgency — could AI potentially replace or surpass humans (Fig. 1a)? Such a scenario seems increasingly plausible, especially with advancements in workplace technologies (Jarrahi, 2018; Zarifhonarvar, 2023). If we have transitioned from a "master-slave" to a "peer-to-peer" decision-making model, what prevents a further shift toward a "slave-master" relationship, with AI as the "master"? Evidence shows that in some contexts, human-AI collaboration is more effective when AI is positioned to define the roles and tasks of its human counterparts (Taesiri et al., 2022). Furthermore, if AI becomes sufficiently autonomous, it could even supplant humans as a species (Hendrycks, 2023; Mulgan, 2016) — a notion frequently explored with trepidation in media and the arts.

From Replacement to Collaboration?

It appears that fostering synergy or collaboration with AI, rather than competing against it, offers a promising strategy to circumvent competition-related issues (Grüning, 2022; Jarrahi, 2018). But how can we effectively achieve such synergy? This question brings us back to the concept of human-machine symbiosis, first articulated by J. C. R. Licklider (1960). In this vision, the strengths of one party compensate for the weaknesses of the other, creating a balanced and effective partnership.

In contemporary research there are multiple paradigms that are built on this idea. Of particular interest in the concept of hybrid intelligence (HI) that explicitly proposes the integration of the complementary capabilities of humans and AI, fostering collaboration that yields superior outcomes compared to what each could achieve independently (Dellermann et al., 2019). Both humans and artificial intelligence systems have their respective flaws and strengths. By designing hybrid cognitive systems that merge the best of both worlds, we can tailor certain task aspects to align well with AI capabilities, while other aspects better suit human abilities (Dellermann et al., 2019; Piller et al., 2022; Rai et al., 2019).

Current studies often highlight several capabilities that AI systems notably lack (Fabri et al., 2023; Heine et al., 2023; Rastogi et al., 2023; Zhou et al., 2021): empathy, general world knowledge, creativity, intuition, the ability to quickly adapt to unforeseen circumstances, the capacity to make generalizations from small data sets, evaluating ethical norms, and the ability to learn from data not digitally present. Humans excel in these areas but have their own shortcomings, such as limited capacity for data-driven analysis and computation, susceptibility to biases, and a reduced efficiency in routine operations.

This approach yields tangible economic benefits (Heine et al., 2023; Van Oudenhoven et al., 2023). Its effectiveness is largely due to empirical studies involving specific AI systems that complement specific human abilities, thereby enhancing performance in designated tasks. However, a significant long-term issue emerges from this reliance on specificity. As we depend on specific AI systems, we often presume their limitations will remain constant. However, AI development progresses at an unpredictable pace, and capabilities previously exclusive to humans are rapidly becoming accessible to machines. For instance, it was once assumed that machines lacked emotional competency, yet recent research shows that some AI models can now outperform humans in significant emotional intelligence domains (Vzorin et al., 2023; Wang et al., 2023). Moreover, there are instances where the performance of human-AI teams does not surpass that of AI alone (Malone et al., 2023), highlighting the diminishing role of humans as this trend continues. If the development of General Artificial Intelligence (AGI) and Artificial Superintelligence (ASI) is inevitable (Kurzweil, 2014), humans might become the redundant component in human-AI collaborations. This possibility emphasizes the need to remain vigilant about the accelerating pace of technological evolution, as the point of singularity — the moment when machine intelligence surpasses human intelligence — could arrive unexpectedly.

 doi:10.20944/preprints202409.1250.v

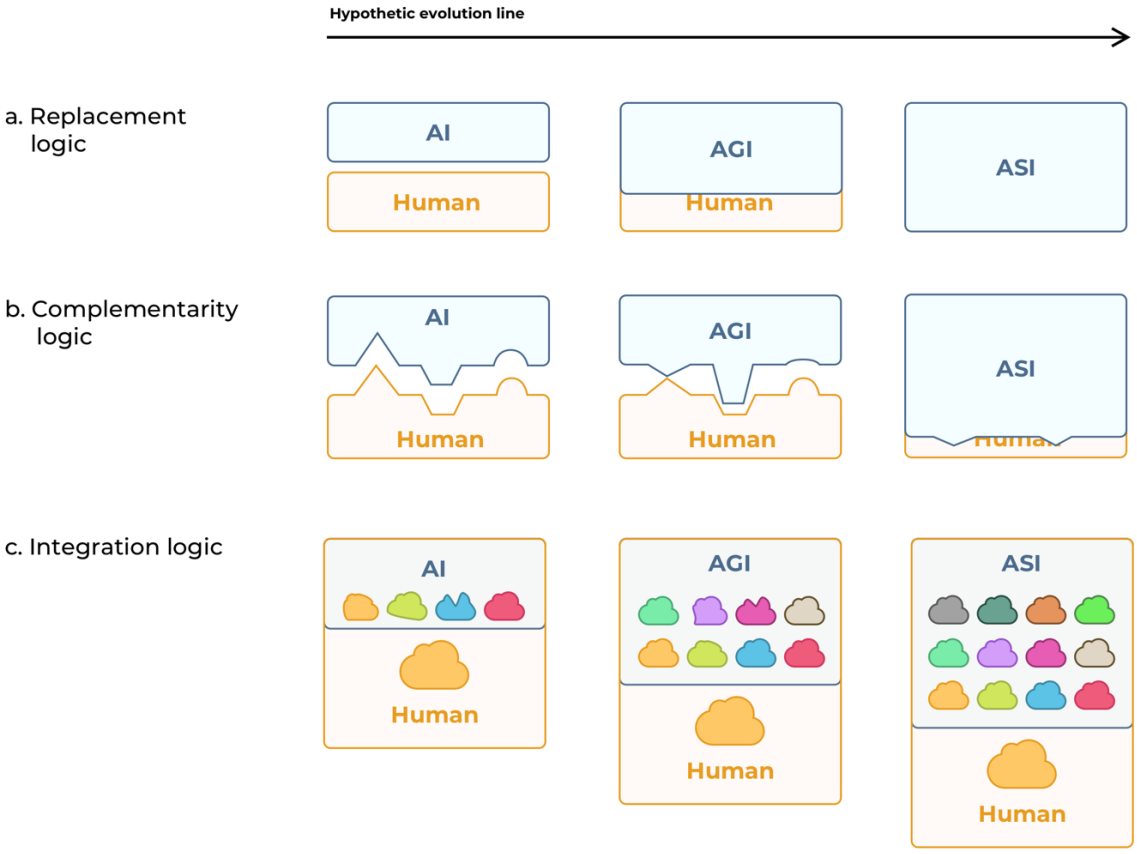

Figure 1. Three logics of human-AI future.

Cognitive Evolution Perspective

If relying solely on current technologies to compensate for human weaknesses proves to be an insufficient strategy, how then can we devise an alternative approach? The hypothesis proposed in this paper is that we need to move from surface of current technologies and current human manifestation in specific abilities to the spatial view of intertwined human and technology co-evolution.

One valuable framework to address this issue is adapting Nikolaas Tinbergen's scheme to study machine behavior, including the evolutionary dynamics of hybrid human-machine behaviour (Rahwan et al., 2019). Tinbergen outlined four complementary dimensions of analysis to explain animal behavior. These dimensions address questions related to the function, mechanism, development, and evolutionary history, offering a structured framework for studying both animal and human behavior in static and dynamic aspects. Current models of Hybrid Intelligence (HI) predominantly represent the static aspect but lack dynamism. Incorporating a dynamic perspective, which delves into the cultural evolution of these interactions, can offer strategies "for ensuring human survival and agency in the long run" (Brinkmann et al., 2023, p. 18) – Fig. 1c.

Instead of focusing on existing technologies, we can anticipate them. But how can we achieve this? Clearly, the technology market and science follow their own logic, which is shaped by current human needs. However, our aim is to transcend these specifics and develop a holistic theoretical framework, where not only technologies would be anticipated, but also human nature would be anticipated as preserved through change, according to Red Queen hypothesis (Van Valen, 1977).

We must assume that certain features prevalent in current technologies — regardless of their specific implementations — did not emerge randomly but as a direct response to underlying evolutionary trends. In other words, if evolutionary processes governed by the *same laws* were to occur in another

 doi:10.20944/preprints202409.1250.v

galaxy, hypothetical creatures there would likely develop technologies that, while formally different, would embody the same essential characteristics (Ushakov, 2018). By addressing this ambitious issue, we can better predict and guide the development of human-machine relationships in our real world with a clearer, more informed perspective.

Basic Model Assumptions

The evolutionary process encompasses a vast complexity, but our objective is not to address every detail or solve eternal problems. Instead, we aim to construct an idealized model that focuses on the transmission of information — a concept we consider crucial for understanding the future impact of information systems, particularly AI, on us as creatures shaped by evolution. This approach allows us to distill the essence of evolutionary dynamics in terms that are relevant to AI and its integration into the broader context of human development.

To develop such an idealized model, we must begin with a fundamental and universal principle that remains invariant across all evolutionary lines, thereby providing a general "direction" by elucidating the underlying motivation of living entities. Simultaneously, it's crucial to identify aspects that are subject to change and to elucidate the internal regularities governing those changes. By synthesizing these principles, we can create an idealized model of evolution that not only enhances our understanding of the past and present but also enables us to predict future developments.

1. The Universal Part In Life Evolution is the Free Energy Reduction

We argue that the free energy principle (FEP) can serve as such starting point (Colombo & Wright, 2021). Formulated by Karl Friston (2010), the FEP posits that any self-organizing system in equilibrium with its environment strives to minimize its free energy. This principle extends to living organisms, which reduce surprise or uncertainty by making active inference predictions based on internal models and refining these models through sensory input. The FEP applies broadly to all biological entities, processes, and complex systems — including organisms without brains such as single cells and plants, evolutionary processes driven by natural selection, and ecosystems (Allen & Friston, 2018; Friston, 2013). Essentially, the principle is a mathematical representation of how adaptive systems counteract the natural tendency towards disorder.

Ensuring the implementation of this principle is crucial, as it allows us to not only focus on evolution itself but also to examine how inner states, such as enactive brain-body states, have evolved. And because our aim is to predict how AI may change our mind, it is important to view biological information in the context of its interpretation or processing of inputs, rather than as an inherent property of those inputs (Jablonka & Lamb, 2006).

While there are alternative guiding principles, more specified within the context of human-computer interaction, such as those proposed in activity theories (Kaptelinin & Nardi, 2012), we have opted to begin with the Free Energy Principle (FEP). This choice stems from its compatibility with various higher-order theories of human evolution that address the "why" level of phenomena. Additionally, the FEP offers a precise framework for mathematical calculations, enabling the integration of theories with the "how" level, thus facilitating more accurate predictions. Furthermore, many seminal higher-order theories are laden with background details and specific discourse, which could potentially divert attention from the primary focus of our discussion. Therefore, our approach is to initiate with a "blank paper" mindset, incorporating pertinent details only when deemed necessary.

2. The variable Part is the Way How Information for Free Energy Reduction is Transmitted

The updates that minimize surprise may occur on evolutionary level (Friston, 2009). Prior information, that allows creatures to make active inferences, is heritable and transmits due to

 doi:10.20944/preprints202409.1250.v

replicators in the process of information selection, replication and transmission known as Darwinian evolution. We argue here that history of life on Earth is marked by numerous major transitions in replicators (Szathmáry, 2015; Szathmáry & Smith, 1995), each corresponding to changes in the ways information can be stored, transmitted (Gillings et al., 2016), and processed (Jablonka & Lamb, 2006). Given the extensive literature on this topic, the question arises: what classification of major transitions should we adopt, and which stages are of particular significance?

3. The Key Regularity is a Multilevel Balance

According to the multilevel evolutionary approach, natural selection operates across multiple levels of the biological hierarchy (Wilson et al., 2023; Wilson & Wilson, 2007). Novel evolutionary innovations in information transmission may initially yield benefits primarily for either individuals or the group at the current stage, leading to a sort of imbalance or constraint in terms of free energy reduction. This imbalance then sets a trajectory and requirement for future transitions. Subsequently, these future transitions would introduce new means of information transmission to address the imbalance on one level, yet potentially creating a new imbalance that necessitates further transition.

If this premise holds, then we can review major evolutionary transitions in terms of information sources and relevant inner models with group/individual dynamics. This review aims to identify current stage peculiarities and limitations that must be addressed in the next evolutionary stage. As demonstrated below, these concepts are greatly approximated by AI invention.

Is This Really Necessary?

One might question the necessity of such a broad scope, arguing that the issue of human-AI interaction is disproportionately small compared to the vast scope we have chosen. It may seem excessive to model the entire evolution to address how the human mind would adapt to new means in its "extended cognition." However, our choice to start with a broader problem — the existential threat to human agency in the near future and all the ethical dilemmas that stem from it—is not without reason. We believe that providing a proper answer requires addressing both of these questions, thus necessitating a comprehensive scope.

Without this evolutionary perspective, it would be unclear which clues are significant and from which emerging forms we can draw generalizations for the future. Absent a selective logic, clues may be intuitively taken based on proximity alone. For instance, the development of future minds is sometimes viewed as a continuation of digital technologies, and theories for human-AI cooperation might be generalized from human-technology interaction. We argue that these assumptions are premature without a general theoretical framework.

The Model of 4 Major Transitions

In our description, we'll provide only essential details, minimizing commonly known information and concentrating on the final evolutionary stages. For each transition (stage), we'll follow this logical structure, highlighting three key sections:

1. The evolutionary innovation in information transmission at this stage.
2. The peculiarities of the information processing of that stage (i.e., how this shift influenced the inner model).
3. The limitation of this stage in terms of optimal free energy reduction for the opposite level (individual/group), prompting a "request" for future transition.

 doi:10.20944/preprints202409.1250.v

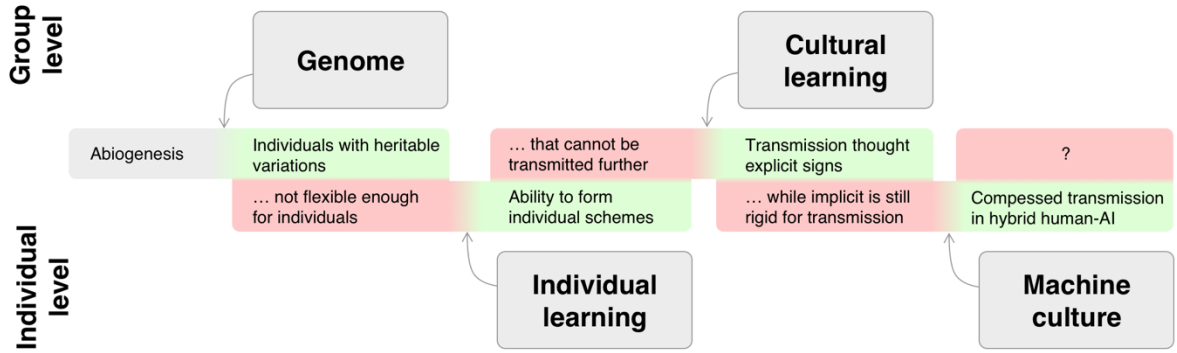

Figure 2. The multilevel evolutionary perspective on human-AI future.

Genome

The life is supposed to start from first self-reproducing protocell entities that culminated in the evolution of a DNA-based genetic system (Ganti, 2003). Genetic transmission allows organisms to transmit information for active inference about the world through generations pressed by natural selection. Thus, it requires separate individuals for natural selection process. Despite quite complex reaction schemes may arise from that kind of transmission on different sub-levels (Jablonka & Lamb, 2006), one obvious limitation of this transmission type is the rigidity of that scheme on individual level. The genetic information determines the range of possible "explanations" and reactions to fixed inputs. The emergence of more complex creatures that live in complex non-homogeneous environment requires more flexible individual source of information, that we will review in the next section.

Individual Learning

As the complexity of living environments increases, evolution compels organisms to satisfy their needs by actively searching for essential resources guided by abiotic properties signaling biologically significant objects (Falikman, 2023; Leontyev, 1981; Tomasello, 2022). These abiotic signals necessitate individual memory and bespoke schemas, as they may vary across different entities. This requirement has spurred the development of a new type of information transmission that occurs within an individual—neural processing. This began with nerve nets in simple diploblastic animals like the Hydra and represented a novel method of transmitting information among cells. This shift in communication had profound evolutionary implications, laying the groundwork for more complex neural architectures and cognitive abilities.

This new source of individual experience implies a new type of processing, wherein the living creature can distinguish between biologically significant (unconditioned) stimuli and paired conditioned biologically neutral stimuli encountered in individual experiences (known as the Pavlovian scheme). This capacity enables the creature to perceive the world as consisting of distinct objects (Leontyev, 1981), leading active inference to engage not merely with single perceptions, but with holistic and meaningful units.

The limitation of this type of information lies in its poor transmission to descendants due to its non-genetic nature (Mesoudi & Whiten, 2008), despite the existence of many well-studied sub-types (e.g., classical conditioning, operant conditioning).

 doi:10.20944/preprints202409.1250.v

Cultural Learning

The limitation of previous stage was inability to effectively transfer individually developed schemes to descendants, which is a disadvantage on group level. The transitional point was the evolvement of observational learning, yet it still lacked sufficient flexibility and occurred too sporadically. The third stage resolves this by integrating the strengths of the previous stages of genetic and individual learning: schemas became complex, universal, and heritable, yet retained flexibility. To achieve this, the culture invents a way of crystalizing individual experience for transferring it to others and influencing others – signs and cultural means (Cantlon & Piantadosi, 2024; Henrich, 2015; Vygotsky & Cole, 2012). Sign systems such as language now "duplicate reality" and serve as substrates for the transmission and recombination of information (Gillings et al., 2016). These signs are explicit in nature, presented as distinct objective parts of physical reality with fixed subjective social meanings. Individual subjective (implicit) experiences can now be converted into objective signs and transmitted to others, where they are decoded back into subjective representations with alterations.

The new mechanism of information transmission results in the new way of information processing, which builds specifically human cognitive system. This approach is widely known in literature as the cultural intelligence hypothesis (Herrmann et al., 2007; Turner & Walmsley, 2021; van Schaik & Burkart, 2011). They key point from our perspective here is that objective means become not just means of subjective knowledge transmission, they become internalized and thus this dichotomy emerges in subjective human mind dividing it to explicit and implicit parts, which are widely studied in dual-system approaches (Neys, 2018)[1]. The inner verbatim explicit part (primary it is a language) is closely intertwined with processes that were previously implicit resulting in whole system rebuilding towards human consciousness and complex cognitive gadgets (Heyes, 2018; Vygotsky & Cole, 2012). Thanks to this explicit part internalization the culture penetrates human mind and transmits inter-individually developed schemes of active inference and surprise reduction. The perceived situation is now always social, every object carries a fixed social meaning, and thus at this stage human is released from the "capture" of direct situation and direct biological needs. Human thought, mediated by speech, becomes abstract. Because of proactive (anticipatory) nature of these schemes they have a guiding influence, enabling human societies to cooperate towards achieving complex common goals — this is the process through which specific human agency emerges (Falikman, 2023; Tomasello & Carpenter, 2007). Furthermore, the human who has a guiding model of a culture in his own mind now can turn these means to affect his own behaviour, resulting in tremendous degrees of freedom of human behaviour (Vygotsky & Cole, 2012).

We argue that current humanity belongs to this stage, and thus we need to meticulously identify the limitations of our current information transmission methods to predict future changes — this is the ultimate purpose for which our model was constructed. The "Genome" stage led to the emergence of distinct individuals, serving as the substrate holders for natural selection. These individuals require more flexible schemas for intra-individual information transmission, which resulted in the development of the nervous system transmission —flexible but non-heritable. The subsequent stage allowed individuals to transmit individually developed schemas to others using explicit means like language, laying the foundation for extensive social cooperation and the evolution of cognitive gadgets. The schema is now flexible, but its limitation is that explicit information can only be transmitted through explicit means. Humanity continuously invented new type of means to make them more efficient: written language, printing press, digital technologies. Each of these advancements, related to major shifts in societal organization (Eisenstein, 1980) and personality (Pea & Cole, 2019) served essentially the same function — to transmit explicit signs to others. Google in these terms is just a huge library of explicit signs produced by other people with quick search using explicit prompts. The human still requires decoding these signs into implicit personal guiding

[1] Thus is has been shown that primitive forms of explicit processes occur in animals (Kelly & Barron, 2022), we argue that it is only the prerequisite for human-level linguistic explicit cognitive structure.

 doi:10.20944/preprints202409.1250.v

schemes and apply them to current situations. However, the general limitation of this stage – rigidity of implicit – remains intact.

This rigidity of the implicit is starkly illustrated in the education process. A person spends many years in school or university where teachers provide explicit signs through verbal conversations and books, resulting in the slow formation of implicit structures, from simple understanding to complex personality development. Both moral and scientific knowledge can initially be merely declarative ("just words"), but true education involves a prolonged process of forming deep implicit structures, mediated by explicit signs and interactions. Later, our habits, general world knowledge, attitudes, and other constructs categorized under the term "implicit" are firm and rigid.

Culture has liberated humans from immediate biological constraints by providing them with sociocultural "lenses", cognitive gadgets, made possible through the internalization of the explicit verbal system. This has led to a tremendous increase in degrees of freedom, yet they are achieved through embedding individuals within a relatively tight social discourse, observable in multiple experimental paradigms like the classical Bartlett's chain method (Bartlett, 1932).

We propose that the main limitation of this stage is the inability to transmit parts of the implicit scheme to others, resulting in the relative rigidity of human personal implicit schemas that require a long time to establish or change through the decoding of explicit signs and their application to current situations, personality, and prior background. As humanity enters a new reality closely intertwined with the digital world and vast amounts of information (Pan, 2016), this limitation becomes especially rough while limiting human information capacity (Cantlon & Piantadosi, 2024).

In the next section, we will discuss why emerging AI technologies appear to be the means of overcoming this limitation, implying a major transition to the next evolutionary stage with extensive consequences for the human mind and society.

Machine Culture Transmission

The limitation identified at the previous stage was the rigidity of implicit cognitive structures and knowledge acquired through the socialization process. Similar to how the "Individual learning" stage overcame the rigidity of genetic schemes at an individual level, we anticipate that the new stage will address the limitations of the "Cultural" stage by leveraging previous advancements in information transmission.

However, how can we theoretically achieve rapid transmission of implicit information? Traditionally, such information was conveyed through mechanisms like genes, observational learning, and imitation. Yet, due to the limited capacity of these methods, culture developed explicit means such as signs and language, which, due to their rule-based nature, could be transmitted quickly. Innovations such as writing, book printing, and digital technologies have facilitated this process but have not qualitatively changed the transmission of implicit information. This is because the cultural means are inherently explicit and can only indirectly convey implicit knowledge through prolonged processes of subject activity, interaction with others, and personal experience.

The transition to the next stage necessitates the development of explicit means capable of encapsulating and harnessing implicit knowledge independently of any subject, thereby externalizing implicit knowledge in an objective form and delivering it to others. Our claim it that this is precisely what most contemporary Generative AI (GAI) algorithms achieve. They can learn patterns and relationships from training data without explicit guidance on specific features to focus on (Dwivedi et al., 2023; Hacker et al., 2023) applying them to solve specific problems.

This is particularly evident in Large Language Models (LLM) like ChatGPT or Claude, which can be seen as miniature models of human culture. On the one hand, these LLMs are explicit human tools (essentially computer code). On the other, these explicit cultural artifacts have acquired implicit human knowledge (Radford et al., 2018). Previously, culture was crystallized only in explicit artifacts and required a human mind to decode them into implicit understanding. Thus, culture was inconceivable without humans. However, LLMs now serve as externalized, independent models of culture, capable of producing explicit signs flexibly based on implicit structures tailored to the current situation (prompt).

 doi:10.20944/preprints202409.1250.v

The ability of GAI to capture and apply societal discourses results in their reciprocal influence on human culture (Brinkmann et al., 2023). But if the cultural intelligence hypothesis—that sophisticated human cognition is substantially shaped by our socioculturally transmitted environment—is accurate, we can anticipate a major shift in the organization of the human mind. The nature of this shift, as we have seen throughout our exploration, will be tied to a new type of information processing that addresses the limitations of the previous stage. This will lead to what we term a hybrid cognitive architecture, which will be discussed in the next section.

Hybrid Cognitive Architecture

We propose that in the new stage of information transmission, the human mind, from a psychological perspective, will differ significantly from what we currently understand as a human mind. It is challenging to envision this type of posthuman mind because our own cognition remains "classical," and there are yet no real manifestations of such evolved minds—this remains a hypothetical construct. However, one way to conceptualize this future is by reversing our perspective: rather than viewing future minds through our current lens, we can try to imagine how they might perceive us in the distant future. This shift is possible because our model, if accurate, has already identified what is likely to change.

Clarifying the Meta-Mind From a Reversed Perspective

From this reversed perspective, the process of socialization of the current human is seen as disproportionately lengthy relative to its outcomes. During socialization in shared activities with others, mediated by explicit signs (cultural artifacts), we acquire cognitive gadgets and a knowledge system that gradually become internalized and automated as implicit structures. This process can be likened to forming a personal "inner LLM," which is implemented in the brain and corresponds to what psychologists describe as "System 1"[2]. Importantly, during the learning process, we do not simply absorb explicit signs (Luria, 1987); we process them to extract meanings that are stored as implicit structures, which transcend the literal words. In daily life, this "inner LLM" manifests as our intuition, imbuing our minds with emotions that guide our thinking. In its "pure" form it manifests, for example, if a person is exhausted or intoxicated. Typically, this mode is sufficient—our environment and our deliberate "System 2" provide prompts to System 1, and the responses are adequate for many situations. If not, our intentional System 2 intervenes, evaluates the responses, and reformulates the prompts. Remarkably, this "inner LLM" is often valued for its long-term development, especially in areas like morals and fundamental beliefs, but it is also relatively rigid — it can be changed, but typically this requires substantial mental effort (Zénon et al., 2019).

The limitations of this "inner LLM" highlight a significant problem with the current stage of information transmission: it is severely constrained by the learning timeframe (confined to a human lifetime), natural brain limitations, and the limited sources of information available. The invention of cultural artifacts capable of directly transmitting implicit knowledge is a response to this limitation and aims to overcome it. The externally personalized LLMs, which have much greater capacity and flexibility, are expected to gradually replace our naturally developed "inner LLMs". Key questions for psychological research include how it is possible to integrate external LLMs into the natural human mind and what consequences this integration might trigger. This exploration forms the basis for understanding how hybrid cognitive architectures could fundamentally reshape human cognition and societal functions.

[2] NB: some theorists argue that current LLMs act in mode that is similar to "System 1" (LeCun, 2022)

 doi:10.20944/preprints202409.1250.v

Core Principles for Research

1. The first question that arises is how is it theoretically possible to include such external LLM into natural human mind? We identify two arguments why it is possible and already happens in practice (Fabri et al., 2023) based on insights from human and AI characteristics:

- **Human Aspect**: From enactivist and social constructivist perspectives, integrating AI might not be perceived as "artificial" because cultural human beings are "artificial" by their nature (Theiner & Drain, 2017) in the sense than human mind is built on artificial artifacts. We think "with" and "through" things (Malafouris, 2019) and such new type of "thing" like AI due to its "soliciting" nature (Diederich, 2021) could seamlessly become part of the human cognitive process, potentially even before physical integration.
- **AI Aspect**: AI represents a form of intelligence that, unlike any animal intelligence, is not inherently tied to desires. As Steven Pinker astutely observes, "Being smart is not the same as wanting something" (Pinker, 2015). While AI can be programmed to simulate desire-driven behavior (Bubeck et al., 2023), such inclination is not intrinsic to its intelligence. This disconnection from desires suggests that AI could complement human intelligence without competing for agency, thus potentially enhancing human individuality and the value of personal creativity.

Moreover, rather than viewing non-sufficient explainability as a problem, it can be seen as a solution. The opaque nature of AI (Wenskovitch & North, 2020) parallels the inherent opacity of the human mind - akin to what we referred to as "inner LLM".

2. The integration of a new "inner LLM" would transform the human mind, adapting old structures to new functions within a larger hybrid system. We expect natural human "System 1" would not vanish but rather find a new role within this hybrid cognitive framework, facilitating seamless interaction with external LLMs. The natural human mind would change towards operating with more general units, in some sense, we may say that every word human speaks can "speak" by itself, elaborating the main idea. This evokes leveraging of human natural mind on higher level (in this sense it can be called meta-mind). This ability would evolve in socialization process, which we expect to be even longer than today yet giving more profitable results. Human would become more degrees of freedom, because now he can flexibly operate different discourses – existing or never existed before, that are abstract per se, but become real and prove their truthiness or moral estimation when applied to concrete circumstances. Processes of discourse formation, that previously occurred in culture, now can occur within an individual that has a new "inner LLM" – this process is similar to those occurred in the transition from "Genome" stage to "Individual learning".

3. The development of such cognitive architectures has profound societal and ethical implications. Particularly intriguing is the potential reconfiguration of societal structures as traditional, fixed social discourses — which are integral to individual identity (Harré & Moghaddam, 2003) — become more dynamic. How society reconstructs itself around these fluid, individually tailored discourses will be crucial.

4. To understand and harness these phenomena effectively, there is a pressing need for translational research that applies this new conceptual framework to real-world scenarios. In the subsequent section, we will briefly explore perspectives and potential directions for this transformative research, aiming to outline how these insights could concretely influence future human-machine collaboration and the evolution of society.

From Theory to Empirics

This theoretical framework was intentionally designed to be idealized, simplifying complexities to focus on overarching principles. Now, practical application and empirical research are required to test its heuristic potential. In this section, we explore three different research areas where applying our principle could be not only useful but also pose significant challenges for the principle itself.

 doi:10.20944/preprints202409.1250.v

"Prospective" Cognitive Archaeology

How can we empirically investigate something that does not yet exist but is theoretically predicted? This challenge is not unique to the future but is also common in studies of the past. The strategy involves using specific artifacts that, through the lens of theory, become meaningful and provide indirect information about the subject under study. This approach is central to cognitive archaeology, which aims to understand the psychological peculiarities of ancient minds (Henley et al., 2019).

Lev Vygotsky's seminal works in the 1930s (Vygotsky & Cole, 2012) focused on a transition similar to the one described in this paper—from "Individual learning" to "Cultural learning." He employed a method where he examined ancient tools and practices, such as casting lots or tying knots for memory. These could initially be understood in terms of conditioning (a stage of individual learning), but Vygotsky argued they should also be considered from the perspective of emerging cultural means due to their transitional nature. From this revised viewpoint, he was able to uncover mechanisms indicative of the transformation in the cultural mind associated with the new stage of learning. This methodological trick—reinterpreting artifacts from a new theoretical perspective—provides a valuable approach to exploring transitions in cognitive development, whether looking into the past or anticipating future changes.

We propose employing a similar methodological approach to anticipate the future by examining current forms of human-computer interaction. These interactions can initially be described using classical terms related to the third stage of cognitive development—the "cultural" stage. However, they might also be viewed as nascent artifacts of a transition to a forthcoming fourth stage.

A key indicator of such a stage transition is often observed as a deterioration in a previous process. For example, in child development, eidetic memory, the ability to recall an image from memory with high precision, tends to decrease as children grow. This decline is linked to language acquisition and the development of verbal skills, which allow older children to think more abstractly and thus rely less on visual memory skills. This transition reflects the evolution to the cultural mind stage, as discussed by Vygotksy (2020).

In a similar vein, we can analyze contemporary phenomena through the lens of a hypothetical fourth stage. By doing so, we act as "prospective" cognitive archaeologists, exploring current technological and interactive trends to uncover how they might represent early signs of a significant cognitive shift. This approach allows us to use current developments in human-computer interaction to predict and possibly guide the evolution towards this new stage of cognitive development.

An illustrative example of such phenomena, according to our hypothesis, is the "Google-effect" (Sparrow et al., 2011). There is growing concern that as we delegate increasing amounts of intellectual activity to machines, there may be a corresponding degeneration in human cognitive capabilities. This view suggests that the human mind, increasingly augmented by technical devices, more "flat" (Friede, 2013; Sparrow et al., 2011; Ward, 2013), less agent (Dell'Acqua, 2022) and less unique (Fügener et al., 2021).

A notable characteristic of digital tools is that they cannot be fully internalized and require the user to be perpetually "online", thereby disrupting the normal developmental path outlined by theorists like Lev Vygotsky (Falikman, 2021). Our guess here is that Vygotskian principle of socialization as interiorization was the proper only on that stage where direct implicit knowledge transmission was impossible ("Cultural learning" stage). The internalization of psychological tools as a process of forming implicit structures mediated by explicit means in human-to-human interaction was essential because, without it, developing agency was challenging; explicit means alone, without a formed implicit structure, were too formal and inadequate for effective management. We argue that there is a more general principle wherein socialization is fundamentally about agency formation. As we transition into a new stage characterized by machine culture transmission, the laws of human socialization may evolve to accommodate this new mode of information processing and thus interiorization becomes unnecessary.

 doi:10.20944/preprints202409.1250.v

From a psychological standpoint, we are suggesting that the integration of digital tools represents a reconstruction of the system of higher psychological functions through digitally mediated activity (Falikman, 2021). Our research aims to delineate the specific pathways of this cognitive reshaping, particularly focusing on adaptation to new forms of information transmission. To explore the implications of the "Google-effect" within this framework, we have conducted a series of experiments (Vzorin et al., 2024). First, we show that the "Google-effect" effect is expedient and goal-directed rather than being an automatic dysfunction. Second, according to our framework predictions, we estimate structural changes in memory with AI instead of “Google” is used.

Artificial Life Modeling

To further explore theoretical concepts or phenomena that don't yet exist, computational modeling offers a promising avenue. Specifically, the mathematical formulation of the FEP can be utilized not just as a theoretical explanation of evolutionary motivations but as a practical tool in modeling. Artificial life modeling, as discussed in recent studies (Aguilar et al., 2014; Bulitko et al., 2019; Park et al., 2023; Perez et al., 2024), provides a robust platform to test and potentially falsify the main ideas of our framework. By simulating life-like behaviors and evolutionary processes in a controlled computational environment, we can examine the dynamics predicted by our model and observe outcomes that may either support or challenge our theoretical assertions. This method enables a rigorous evaluation of hypotheses within our framework by observing how modeled systems adapt and evolve according to the principles we propose.

Conclusion

In this paper, we proposed a novel approach to examining human-machine interactions within an evolutionary framework. This perspective not only illuminates the potential for AI to complement and enhance human cognitive functions but also provides a theoretical basis for understanding the transformative impacts of these technologies on human evolution. While we have laid the groundwork with theoretical models and initial hypotheses, more empirical research is essential to validate and refine our propositions. Continued exploration in this area will be crucial to fully grasp the implications of AI integration and to ensure that such advancements benefit society.

References

1. Aguilar, W., Santamaría-Bonfil, G., Froese, T., & Gershenson, C. (2014). The past, present, and future of artificial life. *Frontiers in Robotics and AI*, *1*, 8.
2. Allen, M., & Friston, K. J. (2018). From cognitivism to autopoiesis: Towards a computational framework for the embodied mind. *Synthese*, *195*(6), 2459–2482. https://doi.org/10.1007/s11229-016-1288-5
3. Bartlett, F. C. (1932). *Remembering*. Macmillan.
4. Brinkmann, L., Baumann, F., Bonnefon, J.-F., Derex, M., Müller, T. F., Nussberger, A.-M., Czaplicka, A., Acerbi, A., Griffiths, T. L., & Henrich, J. (2023). Machine culture. *Nature Human Behaviour*, *7*(11), 1855–1868.
5. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (No. arXiv:2303.12712). arXiv. https://doi.org/10.48550/arXiv.2303.12712
6. Bulitko, V., Doucet, K., Evans, D., Docking, H., Walters, M., Oliver, M., Chow, J., Carleton, S., & Kendal-Freedman, N. (2019). *A-life Evolution with Human Proxies*. 465–466. https://doi.org/10.1162/isal_a_00204
7. Cantlon, J. F., & Piantadosi, S. T. (2024). Uniquely human intelligence arose from expanded information capacity. *Nature Reviews Psychology*. https://doi.org/10.1038/s44159-024-00283-3
8. Colombo, M., & Wright, C. (2021). First principles in the life sciences: The free-energy principle, organicism, and mechanism. *Synthese*, *198*(14), 3463–3488. https://doi.org/10.1007/s11229-018-01932-w
9. Dell’Acqua, F. (2022). *Falling asleep at the wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters*. Working paper. https://www.fabriziodellacqua.com/s/Fabrizio-DellAcqua-Falling-Asleep-at-the-Wheel-Dec-2.pdf
10. Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, *61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

11. Diederich, J. (2021). *The Psychology of Artificial Superintelligence* (Vol. 42). Springer International Publishing. https://doi.org/10.1007/978-3-030-71842-8
12. Dusek, V. (2006). *Philosophy of technology: An introduction* (Vol. 90). Blackwell Oxford. https://www.researchgate.net/profile/Val-Dusek-2/publication/273947214_The_Philosophy_of_Technology_An_Introduction/links/5d7c0b4b299bf1d5a97d6109/The-Philosophy-of-Technology-An-Introduction.pdf
13. Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642
14. Eisenstein, E. L. (1980). *The Printing Press as an Agent of Change*. Cambridge University Press.
15. Fabri, L., Häckel, B., Oberländer, A. M., Rieg, M., & Stohr, A. (2023). Disentangling Human-AI Hybrids: Conceptualizing the Interworking of Humans and AI-Enabled Systems. *Business & Information Systems Engineering*, *65*(6), 623–641. https://doi.org/10.1007/s12599-023-00810-1
16. Falikman, M. (2021). There and back again: A (reversed) Vygotskian perspective on digital socialization. *Frontiers in Psychology*, *12*, 501233.
17. Falikman, M. (2023). Agency, activity, and biocybernetics: On The Evolution of Agency by Michael Tomasello. *Mind, Culture, and Activity*, *30*(1), 90–96. https://doi.org/10.1080/10749039.2023.2246947
18. Friede, E. T. (2013). *Googling to Forget: The Cognitive Processing of Internet Search*.
19. Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301.
20. Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787
21. Friston, K. (2013). Life as we know it. *Journal of The Royal Society Interface*, *10*(86), 20130475. https://doi.org/10.1098/rsif.2013.0475
22. Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *Management Information Systems Quarterly (MISQ)-Vol*, *45*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3879937
23. Ganti, T. (2003). *The Principles of Life*. OUP Oxford.
24. Gigerenzer, G. (2022). *How to Stay Smart in a Smart World: Why Human Intelligence Still Beats Algorithms*. The MIT Press.
25. Gillings, M. R., Hilbert, M., & Kemp, D. J. (2016). Information in the Biosphere: Biological and Digital Worlds. *Trends in Ecology & Evolution*, *31*(3), 180–189. https://doi.org/10.1016/j.tree.2015.12.013
26. Grüning, D. J. (2022). Synthesis of human and artificial intelligence: Review of "How to stay smart in a smart world: Why human intelligence still beats algorithms" by Gerd Gigerenzer. *FUTURES & FORESIGHT SCIENCE*, *4*(3–4), e137. https://doi.org/10.1002/ffo2.137
27. Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other Large Generative AI Models. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1112–1123. https://doi.org/10.1145/3593013.3594067
28. Haenlein, M., & Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review*, *61*(4), 5–14. https://doi.org/10.1177/0008125619864925
29. Harré, R., & Moghaddam, F. (2003). Introduction: The Self and Others in Traditional Psychology and in Positioning Theory. In *The self and others: Positioning individuals and groups in personal, political, and cultural contexts* (pp. 1–11). Praeger Publishers/Greenwood Publishing Group.
30. Heine, I., Hellebrandt, T., Huebser, L., & Padrón, M. (2023). Hybrid Intelligence: Augmenting Employees' Decision-Making with AI-Based Applications. In G. Fortino, D. Kaber, A. Nürnberger, & D. Mendonça (Eds.), *Handbook of Human-Machine Systems* (1st ed., pp. 321–332). Wiley. https://doi.org/10.1002/9781119863663.ch27
31. Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., & Satzger, G. (2024). *Complementarity in Human-AI Collaboration: Concept, Sources, and Evidence* (No. arXiv:2404.00029). arXiv. http://arxiv.org/abs/2404.00029
32. Hendrycks, D. (2023). *Natural Selection Favors AIs over Humans* (No. arXiv:2303.16200). arXiv. https://doi.org/10.48550/arXiv.2303.16200
33. Henley, T. B., Rossano, M. J., & Kardas, A. (2019). Handbook of Cognitive Archaeology. *Psychology in Prehistory. Abhingdon: Routledge*. https://api.taylorfrancis.com/content/books/mono/download?identifierName=doi&identifierValue=10.4324/9780429488818&type=googlepdf

34. Henrich, J. (2015). The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter. In *The Secret of Our Success*. Princeton University Press. https://doi.org/10.1515/9781400873296
35. Herrmann, E., Call, J., Hernàndez-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans Have Evolved Specialized Skills of Social Cognition: The Cultural Intelligence Hypothesis. *Science*, *317*(5843), 1360–1366. https://doi.org/10.1126/science.1146282
36. Heyes, C. (2018). *Cognitive Gadgets: The Cultural Evolution of Thinking*. Harvard University Press. https://doi.org/10.2307/j.ctv24trbqx
37. Jablonka, E., & Lamb, M. J. (2006). The evolution of information in the major transitions. *Journal of Theoretical Biology*, *239*(2), 236–246. https://doi.org/10.1016/j.jtbi.2005.08.038
38. Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, *61*(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007
39. Kaptelinin, V., & Nardi, B. (2012). *Activity theory in HCI: Fundamentals and reflections*. Morgan & Claypool Publishers. https://books.google.com/books?hl=ru&lr=&id=ialeAQAAQBAJ&oi=fnd&pg=PR9&dq=Kaptelinin,+V.,+%26+Nardi,+B.+(2012).+Activity+theory+in+HCI:+Fundamentals+and+reflections.+Morgan+%26+Claypool+Publishers.+&ots=uu1DEmdxdf&sig=03_0f2GT5JzkuGRgUrYK2ZCTq_g
40. Kelly, M., & Barron, A. B. (2022). The best of both worlds: Dual systems of reasoning in animals and AI. *Cognition*, *225*, 105118. https://doi.org/10.1016/j.cognition.2022.105118
41. Kurzweil, R. (2014). The Singularity is Near. In R. L. Sandler (Ed.), *Ethics and Emerging Technologies* (pp. 393–406). Palgrave Macmillan UK. https://doi.org/10.1057/9781137349088_26
42. LeCun, Y. (2022). A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, *62*(1). https://openreview.net/pdf?id=BZ5a1r-kVsf&utm_source=pocket_mylist
43. Leontyev, A. N. (1981). *Problems of the development of the mind*. Progress Publishers. https://cir.nii.ac.jp/crid/1130282271226243840
44. Licklider, J. C. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, *1*, 4–11.
45. Luria, A. R. (1987). *The Mind of a Mnemonist: A Little Book about a Vast Memory, With a New Foreword by Jerome S. Bruner*. Harvard University Press. https://books.google.com/books?hl=ru&lr=&id=FXILEAAAQBAJ&oi=fnd&pg=PR9&dq=luria+small+book&ots=-2m27N113O&sig=v6WIu9D7ROW1HoT2gDPQ7xr4GN8
46. Malafouris, L. (2019). Thinking as "Thinging": Psychology With Things. *Current Directions in Psychological Science*. https://doi.org/10.1177/0963721419873349
47. Malone, T., Vaccaro, M., Campero, A., Song, J., Wen, H., & Almaatouq, A. (2023). *A Test for Evaluating Performance in Human-AI Systems*. https://doi.org/10.21203/rs.3.rs-2787476/v1
48. Mesoudi, A., & Whiten, A. (2008). The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1509), 3489–3501. https://doi.org/10.1098/rstb.2008.0129
49. Mulgan, T. (2016). Superintelligence: Paths, Dangers, Strategies. *The Philosophical Quarterly*, *66*(262), 196–203. https://doi.org/10.1093/pq/pqv034
50. Nah, F., Cai, J., Zheng, R., & Pang, N. (2023). An activity system-based perspective of generative AI: Challenges and research directions. *AIS Transactions on Human-Computer Interaction*, *15*(3), 247–267.
51. Neys, W. D. (Ed.). (2018). *Dual Process Theory 2.0*. Routledge. https://doi.org/10.4324/9781315204550
52. Pan, Y. (2016). Heading toward Artificial Intelligence 2.0. *Engineering*, *2*(4), 409–413. https://doi.org/10.1016/J.ENG.2016.04.018
53. Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22. https://doi.org/10.1145/3586183.3606763
54. Pea, R., & Cole, M. (2019). The Living Hand of the Past: The Role of Technology in Development. *Human Development*, *62*(1–2), 14–39. https://doi.org/10.1159/000496073
55. Perez, J., Léger, C., Ovando-Tellez, M., Foulon, C., Dussauld, J., Oudeyer, P.-Y., & Moulin-Frier, C. (2024). *Cultural evolution in populations of Large Language Models* (No. arXiv:2403.08882). arXiv. http://arxiv.org/abs/2403.08882
56. Piller, F. T., Nitsch, V., & Van Der Aalst, W. (2022). Hybrid Intelligence in Next Generation Manufacturing: An Outlook on New Forms of Collaboration Between Human and Algorithmic Decision-Makers in the Factory of the Future. In F. T. Piller, V. Nitsch, D. Lüttgens, A. Mertens, S. Pütz, & M. Van Dyck (Eds.), *Forecasting Next Generation Manufacturing* (pp. 139–158). Springer International Publishing. https://doi.org/10.1007/978-3-031-07734-0_10
57. Pinker, S. (2015). Thinking does not imply subjugating. In *What to Think About Machines That Think: Today's Leading Thinkers on the Age of Machine Intelligence*. Harper Perennial.
58. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

 doi:10.20944/preprints202409.1250.v

59. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. 'Sandy,' … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), Article 7753. https://doi.org/10.1038/s41586-019-1138-y
60. Rai, A., Constantinides, P., & Sarker, S. (2019). Next generation digital platforms: Toward human-AI hybrids. *Mis Quarterly*, *43*(1), iii–ix.
61. Rastogi, C., Leqi, L., Holstein, K., & Heidari, H. (2023). A Taxonomy of Human and ML Strengths in Decision-Making to Investigate Human-ML Complementarity. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *11*(1), Article 1. https://doi.org/10.1609/hcomp.v11i1.27554
62. Ren, M., Chen, N., & Qiu, H. (2023). Human-machine Collaborative Decision-making: An Evolutionary Roadmap Based on Cognitive Intelligence. *International Journal of Social Robotics*, *15*(7), 1101–1114. https://doi.org/10.1007/s12369-023-01020-1
63. Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G.-J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, *57*(2), 103174. https://doi.org/10.1016/j.im.2019.103174
64. Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. *Science*, *333*(6043), 776–778. https://doi.org/10.1126/science.1207745
65. Szathmáry, E. (2015). Toward major evolutionary transitions theory 2.0. *Proceedings of the National Academy of Sciences*, *112*(33), 10104–10111. https://doi.org/10.1073/pnas.1421398112
66. Szathmáry, E., & Smith, J. M. (1995). The major evolutionary transitions. *Nature*, *374*(6519), 227–232. https://doi.org/10.1038/374227a0
67. Taesiri, M. R., Nguyen, G., & Nguyen, A. (2022). Visual correspondence-based explanations improve AI robustness and human-AI team accuracy. *Advances in Neural Information Processing Systems*, *35*, 34287–34301.
68. Theiner, G., & Drain, C. (2017). What's the Matter with cognition? A 'Vygotskian' perspective on material engagement theory. *Phenomenology and the Cognitive Sciences*, *16*(5), 837–862. https://doi.org/10.1007/s11097-016-9482-y
69. Tomasello, M. (2022). *The evolution of agency: Behavioral organization from lizards to humans*. MIT Press. https://books.google.com/books?hl=ru&lr=&id=g0pTEAAAQBAJ&oi=fnd&pg=PP9&ots=U7yo0XBu5o&sig=kyAJwCk0Ec6yIAwg7JCqfnv5ql4
70. Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental Science*, *10*(1), 121–125. https://doi.org/10.1111/j.1467-7687.2007.00573.x
71. Turner, C. R., & Walmsley, L. D. (2021). Preparedness in cultural learning. *Synthese*, *199*(1), 81–100. https://doi.org/10.1007/s11229-020-02627-x
72. Ushakov, D. V. (2018). Anatomy of Psychological Knowledge. In *Psychological Knowledge: Current State and Development Perspectives* (p. 71).
73. Van Oudenhoven, B., Van De Calseyde, P., Basten, R., & Demerouti, E. (2023). Predictive maintenance for industry 5.0: Behavioural inquiries from a work system perspective. *International Journal of Production Research*, *61*(22), 7846–7865. https://doi.org/10.1080/00207543.2022.2154403
74. van Schaik, C. P., & Burkart, J. M. (2011). Social learning and evolution: The cultural intelligence hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1567), 1008–1016. https://doi.org/10.1098/rstb.2010.0304
75. Van Valen, L. (1977). The Red Queen. *The American Naturalist*, *111*(980), 809–810. https://doi.org/10.1086/283213
76. Voiskounsky, A. Ye. (2013). Psychology of computerization as a step towards the development of cyberpsychology. *Psychology in Russia: State of the Art*, *6*(4), 150–159.
77. Vygotsky, L. S. (2020). Eidetics. *Journal of Russian & East European Psychology*, *57*(4), 320–336. https://doi.org/10.1080/10610405.2020.1800335
78. Vygotsky, L. S., & Cole, M. (2012). *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press. https://books.google.com/books?hl=ru&lr=&id=u2PP6b0ddtoC&oi=fnd&pg=PA1&ots=vGk65BJkUC&sig=V5q-_GpwzrHqsfqkvnexwvPAVcc
79. Vzorin, G., Bukinich, A., Sedykh, A., Vetrova, I., & Sergienko, E. (2023). *Emotional Intelligence of GPT-4 Large Language Model*. https://www.preprints.org/manuscript/202310.1458
80. Wang, X., Li, X., Yin, Z., Wu, Y., & Liu, J. (2023). Emotional intelligence of Large Language Models. *Journal of Pacific Rim Psychology*, *17*, 18344909231213958. https://doi.org/10.1177/18344909231213958
81. Ward, A. F. (2013). Supernormal: How the Internet is changing our memories and our minds. *Psychological Inquiry*, *24*(4), 341–348.
82. Wilson, D. S., Madhavan, G., Gelfand, M. J., Hayes, S. C., Atkins, P. W. B., & Colwell, R. R. (2023). Multilevel cultural evolution: From new theory to practical applications. *Proceedings of the National Academy of Sciences*, *120*(16), e2218222120. https://doi.org/10.1073/pnas.2218222120

83. Wilson, D. S., & Wilson, E. O. (2007). Rethinking the Theoretical Foundation of Sociobiology. *The Quarterly Review of Biology*, *82*(4), 327–348. https://doi.org/10.1086/522809
84. Zarifhonarvar, A. (2023). Economics of ChatGPT: A labor market view on the occupational impact of artificial intelligence. *Journal of Electronic Business & Digital Economics*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/JEBDE-10-2023-0021
85. Zénon, A., Solopchuk, O., & Pezzulo, G. (2019). An information-theoretic perspective on the costs of cognition. *Neuropsychologia*, *123*, 5–18. https://doi.org/10.1016/j.neuropsychologia.2018.09.013
86. Zhou, L., Paul, S., Demirkan, H., Yuan, L., Spohrer, J., Zhou, M., & Basu, J. (2021). Intelligence Augmentation: Towards Building Human-Machine Symbiotic Relationship. *AIS Transactions on Human-Computer Interaction*, *13*(2), 243–264. https://doi.org/10.17705/1thci.00149
87. Взорин, Г. Д., Букинич, А. М., & Нуркова, В. В. (2024). Переосмысляя Google-эффект: Целесообразность забывания сохраненного на внешнем носителе материала. *Вестник Санкт-Петербургского Университета. Психология.*, *4*(3).
88. Vzorin G. D., Bukinich A. M., & Nourkova V. V. (2024) Reconsidering the “Google Effect”: reduced productivity in externalized information recognition is expedient. Vestnik of Saint Petersburg University. Psychology, 14(3) (In Russian)