

Perspective

Cognitive Entanglement: Toward a Developmental Framework of the Human-AI Coevolutionary Leap

Xiao-Kun Wu ^{1,2,†} , Min Chen ^{3,4,*,†} and Giancarlo Fortino ⁵ 

¹ School of Journalism and Communication, Renmin University of China, Beijing 100872, China; wuxiaokun@ruc.edu.cn

² Research Center of Journalism and Social Development, Renmin University of China, Beijing 100872, China

³ School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China

⁴ Pazhou Laboratory, Guangzhou 510640, China

⁵ Department of Informatics, Modeling, Electronics, and Systems, University of Calabria, 87036 Rende, Italy; giancarlo.fortino@unical.it

* Correspondence: minchen@ieee.org

† These authors contributed equally to this work.

Abstract

Large language models have become routine participants in everyday cognition. Their role has widened from retrieval and text generation to helping users define problems, organize arguments, make judgments, and interpret themselves. Yet their cognitive consequences are strikingly divergent. For some users, generative AI appears to reduce critical engagement, independent judgment, and tolerance for difficulty. For others, the same class of systems becomes a medium for conceptual expansion, reflective questioning, and higher-order learning. This divergence cannot be explained by model capability alone. Mental effort is often treated as a cost to be reduced. Yet repeated delegation may also reduce opportunities to practice the processes required for independent judgment. The key issue is developmental: how sustained AI use changes users' cognitive capacities over time. This perspective proposes cognitive entanglement as a framework for understanding the developmental consequences of sustained human-AI coupling. Cognitive entanglement refers to a relation in which human and AI activity become mutually shaping, irreducible to either party alone and organized across different developmental levels. The framework examines how repeated interaction with AI changes the ways users formulate problems, evaluate reasons, and make judgments. Unlike theories that locate the boundaries of cognition (the extended mind, enactivism) or explain the mechanisms of consciousness (global workspace, higher-order, predictive-processing, and integrated-information theories), cognitive entanglement examines whether sustained AI use preserves, weakens, or reorganizes users' cognitive capacities. The article argues that current AI systems are often optimized for fluency, immediacy, and user satisfaction, and this may reduce the productive difficulty that supports higher-order cognitive development. If AI is to support human cognitive growth, design must move beyond answer provision and efficiency maximization toward the organization of productive human-AI relations: relations that challenge users' initial assumptions while providing support appropriate to the task and the user's level of expertise. The argument draws on philosophy of mind, cognitive science, and learning science, and compares divergent approaches to coupling in order to specify which forms of relation carry which developmental consequences. The concept shifts attention from AI as a tool or automation system to the developmental consequences of sustained human-AI interaction.



Academic Editor: Yin Zhang

Received: 19 June 2026

Revised: 29 July 2026

Accepted: 3 August 2026

Published: 5 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: cognitive entanglement; human-AI interaction; cognitive development; large language models; extended mind; AI design

1. Introduction: From Intelligent Machines to Transforming Users

Debates about artificial intelligence have long been organized around the machine. Can machines think? Can they understand? Can they reason, create, or become conscious? These questions remain important. Yet the diffusion of large language models has made another question harder to avoid: what happens to human cognition when AI becomes a routine participant in thinking itself?

Task performance is one part of this question. Generative systems now help users draft arguments, summarize evidence, generate hypotheses, evaluate options, and refine self-descriptions. Their influence increasingly extends to problem formulation itself. Through repeated interaction, users may also rely on AI when defining goals, setting evaluation criteria, and framing problems. Human-AI interaction may therefore shape how users formulate questions, assess evidence, and decide when a task is complete.

The empirical picture is already ambivalent. Some studies suggest that reliance on generative AI is associated with reduced critical engagement and diminished independent judgment. Lee et al. [1] surveyed 319 knowledge workers and reported that greater trust in generative AI is associated with lower self-reported investment in critical thinking. Kosmyna et al. [2], using EEG, further describe a possible accumulation of “cognitive debt” among participants using large language models for essay writing, although the study’s sample size, methodology, and interpretation remain contested [3]. Although these findings are preliminary and contested, they raise a theoretical question: AI use does not necessarily extend human cognition; in some contexts, it may replace the very processes through which cognitive capacities are maintained and developed.

At the same time, cognitive degradation is not the only trajectory. Many users employ the same systems to expand inquiry, expose hidden assumptions, compare conceptual frameworks, test arguments, and articulate ideas that would otherwise remain inchoate. The same system may encourage reliance in some contexts while supporting reflection and inquiry in others. This variation raises a central question: why do similar systems have different cognitive effects across users and tasks?

In a preregistered meta-analysis of 106 experiments and 370 effect sizes, Vaccaro, Almaatouq, and Malone [4] found that human-AI combinations, on average, performed significantly worse than the individual performer, whether human or AI (Hedges’ $g = -0.23$). Yet the effects were strongly heterogeneous, varying with task type and with the relative competence of the human and the AI. These findings suggest that outcomes depend on how tasks, feedback, and decision authority are allocated between users and AI systems.

This perspective develops the concept of cognitive entanglement (CE) to explain this divergence. Cognitive entanglement denotes a form of human-AI cognitive coupling in which human and artificial contributions become mutually shaping and where the resulting cognitive process cannot be reduced to either party alone and the relationship develops across qualitatively different levels. At one level, AI functions as an instrument for efficiency and cognitive offloading. At another level, it becomes a responsive quasi-other that shapes habits of inquiry, dependence, and reflection. At a deeper level, it can enter the user’s self-understanding, challenging or reinforcing the standards by which they judge, explain, and revise themselves.

Our central claim is that AI use does not have uniform cognitive effects. Depending on how it is used, AI may support cognitive offloading, encourage dependence, or facilitate

reflection and learning. The outcome depends on how AI is incorporated into the user's task, reasoning process, and evaluative practice. This has implications for both theory and design. Theoretically, cognitive entanglement extends debates about machine intelligence, extended cognition, and enactive coupling by adding a question that those debates set aside: how cognition changes through technological relations. Normatively, it suggests that the design of AI systems should include, among its objectives, the preservation of the forms of difficulty through which human cognitive capacities grow.

Problem: Similar generative systems can have different cognitive effects across users and tasks, but existing frameworks do not adequately explain this variation. Theories of mind establish where cognition is located or how consciousness arises; evaluations of AI establish what systems can do. Neither directly addresses the conditions under which sustained AI use may preserve, weaken, or reshape users' cognitive capacities. Evidence that AI use has cognitive consequences is accumulating. A useful framework should distinguish forms of human-AI interaction and specify the cognitive outcomes they are expected to produce.

Aim: Drawing on philosophy of mind, cognitive science, learning science, and human-AI interaction research, this perspective develops a conceptual account of divergent human-AI outcomes. The aim is to define distinctions that can guide design and generate empirically testable hypotheses. Establishing an empirical result and adjudicating the question of machine mind both lie outside its scope.

Contributions: This perspective makes four conceptual contributions. First, it defines cognitive entanglement through three operational criteria, namely mutual shaping, irreducibility, and developmental stratification, thereby distinguishing the concept from tool use, cognitive offloading, and collaboration. Second, it identifies three levels at which AI enters cognition, namely the instrumental, relational, and identity levels, alongside an orthogonal dimension of scale through which entanglement also operates at the population level rather than solely within individuals. Third, it proposes a three-variable conditional model consisting of cognitive tension, AI adaptivity, and cognitive openness, which specifies when coupling is likely to produce development rather than dependence. Fourth, it derives, from that model, a design critique of the efficiency paradigm, as well as a corresponding set of practices for users.

Structure: Section 1 situates the question relative to classical debates about machine mind. Section 2 defines cognitive entanglement and its criteria. Section 3 develops the three levels, the divergence of trajectories they explain, and the collective dimension of scale. Section 4 asks what cognitive transformation is and whether it can be designed, contrasting carbon-based, silicon-based, and co-transformative modes. Section 5 sets out the three-variable conditional model. Section 6 turns that model into a critique of systems optimized for fluency and satisfaction, and Section 7 turns it into a set of metacognitive practices for users. Section 8 returns the framework to theories of mind, and Section 9 states the limitations and open questions.

2. Beyond the Question of Machine Mind

Classical debates about AI and mind have largely been organized around the status of the machine. Turing's imitation game [5] displaced the question "Can machines think?" with an operational test of intelligent behavior. Searle's [6] Chinese Room argument resisted this move by insisting that syntactic symbol manipulation, however fluent, does not amount to semantic understanding. Clark and Chalmers [7] later shifted the issue from machine intelligence to cognitive boundaries, arguing that external artifacts can become part of the cognitive system under appropriate conditions. Enactive approaches

further challenged internalist accounts by treating cognition as something enacted through organism–environment coupling rather than being computed inside an isolated head [8].

These traditions provide important conceptual resources for the present argument. Yet they leave a question underdeveloped that has become central in the age of large language models: how do human cognitive capacities change when the environment with which humans are coupled is itself a generative and responsive AI cognitive system?

This distinction matters. A language model may be treated as an external resource, a conversational partner, a reasoning aid, a writing machine, or a mirror for self-interpretation. In each case, the same technical system enters cognition differently. The notebook in the extended mind literature stores information; the large language model reorganizes the space in which questions, answers, standards, and self-descriptions appear. Its influence extends beyond memory support and information retrieval. It can reshape what the user notices, what counts as a good question for them, when they stop thinking, and how they evaluate their own understanding.

For this reason, the more pressing issue is developmental: whether repeated coupling with AI cultivates or erodes the capacities through which humans think. Where AI sits relative to the boundary of the cognitive system, and whether it possesses intentionality or consciousness, are separate matters. A system that supplies fluent answers may support reflection when used to test assumptions and expose alternatives. The same system may weaken reflection when it allows users to bypass the struggle of formulation, comparison, and revision. The difference lies in how the interaction is structured.

We use the term **cognitive entanglement** to describe this developmental dimension of sustained human-AI interaction. It does not deny the importance of existing theories of mind; rather, it relocates their questions within a changing ecology of cognition. Whereas those positions ask whether machine behavior warrants the attribution of intelligence or involves understanding, whether external resources can be part of cognition, and how cognition arises through coupling, cognitive entanglement asks what kind of cognitive subject a human becomes when such coupling is sustained, intimate, and increasingly delegated to artificial systems.

This shift also changes the evaluative vocabulary. Accuracy, efficiency, usability, and user satisfaction capture whether a task has been completed. They are silent on whether the user's capacity to understand, judge, and revise has been strengthened or diminished. A cognitively valuable interaction may be less smooth, less immediately satisfying, and more demanding than an efficient one. Conversely, an interaction that feels seamless may quietly eliminate the moments of uncertainty in which learning and self-correction occur.

The concern is that AI may become embedded in everyday cognition while being optimized primarily for speed, fluency, and user satisfaction rather than long-term cognitive development. The concept of cognitive entanglement is intended to clarify this concern and identify conditions that may shape its consequences.

3. What Is Cognitive Entanglement?

Cognitive entanglement refers to a sustained form of human-AI cognitive coupling in which human and AI activity become mutually shaping, the resulting cognitive process is irreducible to either side alone, and the relationship has developmental consequences for the human user. This and related terms are defined in Appendix A.

This definition has three components, as outlined below.

First, cognitive entanglement is **relational**. It emerges from repeated interactions between the two parties, and not from an isolated mental event in the individual or a unilateral effect of the AI system.

Second, cognitive entanglement is **temporal**. One-off use, casual consultation, or low-intensity assistance does not necessarily constitute entanglement. Entanglement begins when patterns of interaction stabilize over time and start to reorganize the user's cognitive habits. The key question is whether repeated use changes how the user subsequently frames problems, evaluates evidence, and assesses their own capacities.

Third, cognitive entanglement is **developmental**. It concerns how joint human-AI activity changes the user's future capacity to think without, with, or against AI. This is where cognitive entanglement differs from ordinary tool use, cognitive offloading, or collaboration. Cognitive offloading [9] transfers cognitive burden to external resources. By contrast, CE concerns whether repeated offloading changes users' attention, judgment, self-monitoring, or self-understanding over time.

To give the concept clearer boundaries, we propose three criteria, which are outlined in the subsections below.

3.1. Mutual Shaping

A relation becomes cognitively entangled when human and AI activities adapt to one another in a recurrent loop. The user's prompting strategies, contextual framing, evaluation standards, and selective uptake shape what the AI becomes in practice. Even when the model itself is unchanged, users may employ it in different roles: it may function as a substitute writer, an information retriever, a debate partner, a conceptual clarifier, or a frame challenger. At the same time, the AI's responses influence how the user formulates questions, allocates attention, tolerates uncertainty, and decides when a problem is "solved".

This criterion distinguishes cognitive entanglement from simple outsourcing. In outsourcing, the user hands a task to an external resource. In entanglement, the repeated acts of handing over, evaluating, correcting, accepting, and depending on AI gradually alter how the user asks, understands, and judges. Over time, users shape the role AI plays in a task, and that role can in turn influence their habits of reasoning and evaluation.

3.2. Irreducibility

Entangled outcomes cannot be fully explained either as the natural extension of the user's prior capacities or as the direct product of AI output. They emerge from the interaction between the user and the system over time.

Consider a researcher who repeatedly uses AI to survey the literature, test arguments, generate counterexamples, and rewrite concepts. Over time, their way of designing a paper may change. They may become more inclined to compare frameworks, anticipate objections, define concepts carefully, or request adversarial critique. They may also become more dependent on AI for the initial structure. Such changes are not merely the researcher's original abilities unfolding, nor are they simple copies of a model's output. They are new cognitive dispositions formed through interaction.

Irreducibility matters because it distinguishes cognitive entanglement from ordinary tool use. Tools can certainly reshape habits, skills, and standards. But deep interaction with generative AI can intervene in the organization of cognition itself: problem definition, reason generation, standard calibration, and self-evaluation. If AI is removed after ordinary assistance, the user may mainly experience a loss of efficiency. If AI is removed after deep entanglement, what remains may be a changed structure of ability, dependence, judgment, or self-understanding.

3.3. Developmental Stratification

Cognitive entanglement varies in kind as well as in degree. Its consequences depend on the level at which AI enters the cognitive process.

At a relatively shallow level, AI performs local tasks: polishing sentences, summarizing documents, translating text, or generating preliminary outlines. Such use may improve efficiency without substantially changing the user's problem frame or judgment structure. At a deeper level, AI participates in organizing reasoning, comparing alternatives, and producing counterexamples; the user's judgment process is thereby redistributed. At a still deeper level, AI enters problem definition, premise reflection, conceptual reconstruction, and self-monitoring. At this depth, the user is changing how they understand the task, define the problem, and evaluate possible answers.

This stratified view clarifies why frequency of use is not the same as depth of entanglement. High-frequency use may remain shallow if it only automates local tasks. A less frequent but high-quality interaction may be deeper if it changes how the user frames problems or evaluates themselves as a thinker. What matters is whether AI affects metacognitive activities such as monitoring, regulating, and revising one's own reasoning. AI that supplies a conclusion is not the same as a condition that develops higher-order cognition, by which we mean the suite of capacities through which a person monitors, regulates, and revises their own cognitive processes: metacognitive awareness, executive control of attention and inference, and the ability to evaluate the grounds of one's own beliefs. These capacities are not produced by receiving answers; they require structured engagement with problems that resist one's first approach and with conclusions that must be justified rather than merely accepted.

3.4. Theoretical Sources

Cognitive entanglement draws on, but also extends, several bodies of work in philosophy of mind, developmental psychology, and science and technology studies.

The extended mind thesis and distributed cognition provide the ontological background for taking external resources seriously as parts of cognitive systems [7,10]. Large language models, however, differ from notebooks, maps, or traditional databases. Beyond storing information, they generate suggestions, reorganize expression, and help shape the criteria by which users judge clarity, relevance, and adequacy. Cognitive entanglement therefore turns the boundary question into a developmental one: how such extension changes the structure of capacity over time.

Vygotsky's [11] notion of the zone of proximal development offers a second resource. Cognitive development often occurs in the zone where challenge becomes manageable through scaffolding, between what a learner can already do alone and what lies entirely beyond reach. AI may, in some circumstances, function as a cognitive scaffold: asking sharper questions, exposing weaknesses, offering alternative perspectives, generating counterexamples, or turning vague intuitions into discussable concepts. Yet unlike a human tutor, an AI system does not possess pedagogical intention or semantic understanding in the same sense. Its developmental value therefore depends less on whether it is a "more capable other" in the human sense than on whether the interaction reliably produces challenge, feedback, error exposure, and reconstruction.

Actor-network theory [12] provides a third resource by showing how nonhuman entities can have real effects within sociotechnical systems. AI systems need not possess human-like intention in order to shape cognition. Their training data, architecture, alignment procedures, interface designs, and default response styles influence what questions are asked, what answers appear reasonable, what uncertainties are hidden, and what forms of expression are rewarded. Cognitive entanglement adapts this insight to the cognitive domain. AI occupies a position between full subject and neutral tool: a nonhuman cognitive actor in the limited sense that it can alter the conditions under which human understanding, judgment, and self-revision occur.

4. Three Levels of Entanglement

If cognitive entanglement is defined by mutual shaping, irreducibility, and developmental consequence, it still does not follow that all entanglements are alike. The same AI system may enter human cognition at very different levels. It may serve as a local instrument for efficiency, it may become a responsive quasi-other that shapes the user's habits of inquiry and dependence, or it may enter the user's self-understanding and alter the standards by which they judge their own capacities.

This distinction helps explain why similar AI use may have different consequences. What differentiates them is where in the organization of cognition the human-AI relation is located. Is AI used mainly to replace local tasks? Does it become a dialogical scaffold? Or does it challenge the user's self-conception as a thinker, writer, researcher, or decision-maker?

This layered account is summarized in Figure 1.

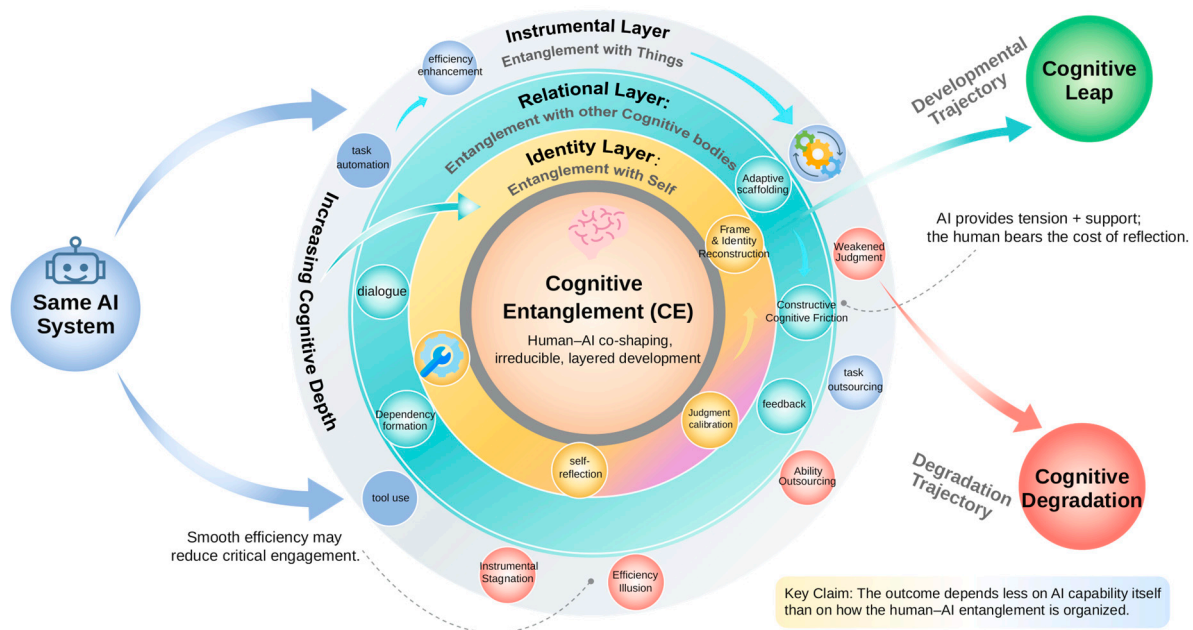


Figure 1. Three layers of cognitive entanglement and divergent trajectories.

4.1. Instrumental Entanglement: AI as an Efficiency Device

At the most basic level, AI is entangled with users as an instrument. It helps the user to generate text, summarize documents, translate materials, organize information, draft outlines, propose solutions, or complete other tasks that can be delegated. The relation is primarily functional: the user defines a need, the system generates an output, and the user accepts, modifies, or rejects it.

This is not mere tool use in the traditional sense: generative AI does more than execute predefined operations. Its outputs can reshape how users understand a task, what they regard as a plausible answer, and how they judge clarity or quality. Yet, at the instrumental level, the dominant orientation remains efficiency and substitution. At this level, users are primarily concerned with obtaining a usable result efficiently rather than reflecting on how the interaction may shape their thinking.

One risk is that AI-assisted fluency may be mistaken for improved individual competence. More output, faster completion, and smoother expression may be mistaken for genuine cognitive improvement. A user may write more, answer faster, and appear more competent while becoming less able to formulate problems, tolerate ambiguity, compare reasons, or revise judgments independently. Research on cognitive offloading has already

shown that convenient external supports can reduce motivation for internal memory encoding and active processing [9]. In the context of generative AI, this mechanism extends beyond memory to higher-order activities: preliminary analysis, argument organization, expression, and even evaluative judgment may be outsourced.

Recent empirical work provides preliminary reasons for concern. Gerlich [13] reports associations between AI use and cognitive offloading in relation to critical thinking. Lee et al. [1] find that greater trust in generative AI among knowledge workers is associated with lower self-reported effort in critical thinking. Kosmyna et al. [2] suggest that repeated use of large language models for writing may be associated with weaker neural connectivity and reduced memory for generated text, although the interpretation of these findings remains contested [3]. These studies should not be overgeneralized, but they converge on a concern central to this perspective: when AI is organized primarily as a substitute, users may gradually lose opportunities to exercise the capacities being substituted.

The concern is not task completion itself, but the possibility that users may interpret AI-assisted performance as evidence of their own cognitive development. It is possible that users may come to treat the fluency of AI-assisted performance as evidence of their own cognitive growth. Using AI at the start of a task may be reasonable in individual cases. But when such acts stabilize into a default mode, the capacity to begin, struggle, compare, and revise may erode. Any decline is more likely to emerge gradually through repeated reliance on AI for tasks that users might otherwise perform themselves.

4.2. Relational Entanglement: AI as a Quasi-Other

At a second level, AI becomes more than an efficiency device. It is experienced as a responsive quasi-other: something that can answer, question, reassure, challenge, suggest, and maintain a recognizable interactional style. As large language models become more capable in dialog, affective response, and role enactment, many users have moved from extracting outputs to organizing their thinking through repeated conversation with the system.

The emergence of this relational level stems partly from the anthropomorphic pull of large language models. Winnicott's [14] account of the "transitional object" helps to illuminate it: a transitional object lies between the self and the external world, an intermediate zone in which the subject can project onto, probe, and regulate their own experience. For some adults, AI may play a partially analogous role in reflective or exploratory thinking. Its responsiveness may allow users to explore ideas with less perceived interpersonal risk than in human interaction.

This relational level is often discussed in terms of attachment, anthropomorphism, emotional substitution, or social risk. These concerns are real. Relational entanglement can foster dependence, echo chambers, self-enclosure, and affective replacement. But if this level is understood only pathologically, its developmental significance is missed. Human cognition does not mature in isolation. Many higher-order capacities are formed through dialog, feedback, correction, and collaborative problem solving. This is the space named by Vygotsky's [11] zone of proximal development, introduced above.

The relevant question is how the system's conversational role is designed and used. AI can function as a low-level satisfier: confirming the user's assumptions, reducing frustration, providing frictionless reassurance, and keeping the user within a comfortable cognitive zone. But it can also function as a scaffold: posing questions, exposing weaknesses, demanding clarification, presenting alternative perspectives, and helping the user move beyond current ability.

Relational entanglement thus has a double potential: it may deepen dependence by allowing users to transfer uncertainty, judgment pressure, and expressive anxiety to AI, or it

may create conditions for cognitive development because many advanced abilities require a responsive other that can provide feedback, sustain tension, and maintain continuity across attempts. At this level, the same conversational system may either reinforce reliance or support reflective learning. The same experience of “talking with AI” may lead either to cognitive comfort or to cognitive growth.

4.3. Identity-Level Entanglement: AI as a Mirror of the Cognitive Self

The third level concerns the user’s cognitive self-understanding. At the instrumental level, AI changes how tasks are completed. At the relational level, it changes the structure of interaction. At the identity level, it begins to touch the user’s understanding of themselves as a cognitive subject.

By cognitive identity, we mean the relatively stable self-descriptions through which individuals understand their own cognitive capacities: “I am good at writing”; “I have sound judgment”; “I understand this field”; “I can formulate problems independently”; “I am not creative.” Such narratives shape how people interpret failure, choose tasks, respond to challenges, and decide whether to enter difficult cognitive terrain.

Identity-level entanglement occurs when AI outputs, questions, or feedback challenge these self-descriptions. A person who sees themselves as a strong writer may discover, through structured AI feedback, that their prose relies on rhetorical force but lacks argumentative architecture. A user confident in their judgment may encounter counterexamples that reveal systematic biases. A person who believes they lack theoretical creativity may, through sustained dialog, discover a latent problem-sense not previously articulated. In these cases, AI becomes a cognitive mirror through which the user reinterprets who they are, what they can do, and how they think.

This process often involves cognitive friction: discomfort generated when the user’s existing self-understanding, judgment habits, or problem frames conflict with AI-mediated feedback, counterexamples, or alternative framings. Cognitive friction can produce defensiveness. The user may reject the feedback, demand confirmation, or selectively attend only to outputs that preserve their existing self-image. But it can also produce reconstruction. The user may revise their understanding of their abilities, reorganize the problem frame, and develop new ways of thinking.

Hofstadter’s [15] account of the “strange loop” offers a useful analogy here. Self-consciousness can be understood, in part, as a system’s capacity to model and refer to itself. In identity-level entanglement, AI enters this self-referential loop. The user sees aspects of their thinking through AI’s responses; this seeing changes their self-understanding; the altered self-understanding then changes how they prompt, evaluate, and work with AI. The interaction becomes part of the user’s higher-order modeling of their own cognition.

This level has the greatest potential for transformation because it touches the regulative structures of cognitive development. Many limitations in thinking do not arise simply from insufficient information or skill. They arise from mistaken assumptions about one’s own capacities, fixed expectations about what a problem is, or low-level standards for what counts as a good answer. When such standards are challenged and reorganized, the gain reaches past the present task: the user may become able to approach future ones differently.

4.4. Why Trajectories Diverge

The divergence of cognitive trajectories therefore follows from the level at which the relation settles: facing the same AI system, users may form entanglements of different depths.

Instrumental and relational entanglements can both become comfortable habitats: one offers efficiency, the other response. If interaction remains within that zone of comfort, a stronger model may simply reduce the occasions for difficult thinking. When the relation

instead reaches self-understanding, evaluative standards, and problem frames, repeated interaction can become qualitative transformation.

4.5. *The Dimension of Scale: Collective Entanglement and the Erosion of Cognitive Diversity*

The three levels describe depth, meaning how far into a person's cognition the relationship reaches; they do not describe scale. Entanglement is also a property of populations. A second and orthogonal dimension concerns how many people are entangled with the same system at once. Where the level of entanglement asks where in the organization of one person's cognition the relation is located, the dimension of scale asks what happens when a large share of a population is coupled to a small number of models. The two dimensions are independent: instrumental entanglement can be near-universal, and identity-level entanglement can be rare; the cognitive consequences of each are not the same at the individual and population scale.

The mechanism becomes visible once the source of cognitive tension is made explicit. For an individual, tension may be supplied by the system itself, when it objects, withholds, or offers a counterexample (Section 5.1). At the level of a population, there is a second and historically more reliable source: disagreement among other people. Intellectual diversity is the collective analog of cognitive tension. It is what makes a framing contestable, an assumption visible, and an error detectable by someone who did not share it. When a substantial part of a society consults the same few models—trained on overlapping corpora, aligned by similar procedures, and converging on similar defaults of tone, structure, and reasonable-sounding conclusion—the ways in which problems are framed, solutions generated, and facts evaluated begin to converge as well. One possible population-level risk is reduced diversity in problem framing and evaluation.

The consequence is that the second source of tension is removed. If many users rely on a small number of systems with similar training data and interaction defaults, errors and blind spots may become more correlated. A systemic mistake of one dominant technology—a blind spot in its training distribution, a bias in its alignment, a framing it finds natural—no longer meets the resistance of people who would have framed the problem differently, because fewer such people remain. Such blind spots could spread across domains in which the same systems are widely used. This is the erosion of social and epistemic norms: the standards governing what counts as a good question, an adequate argument, or sufficient evidence are collective achievements, sustained by disagreement, and they are not preserved by the good judgment of any individual user. This consequence is a corollary of the conditional model itself: where tension disappears, development does not occur, and at population scale tension is disagreement.

Scale also interacts with level rather than replacing it. At the instrumental level, homogenization is largely stylistic and its costs are modest. At the relational level, habits of inquiry converge: people learn to ask the kinds of questions that the dominant systems answer well. At the identity level, convergence reaches the standards by which people judge themselves as thinkers, and the loss becomes difficult to perceive from inside, since a shared standard feels like a correct one. The deepest collective risk is therefore the quiet standardization of the criteria by which the dependence of any single user would have been noticed. Preserving diversity accordingly becomes a design and policy requirement in its own right—plurality of models, visible disagreement between systems, and resistance to convergence on a single default voice—alongside the individual-level requirements developed in Section 6. Across both dimensions, however, the governing question is the same: whether tension survives the interaction.

This leads to a further question: if identity-level entanglement can produce such a transformation, what exactly is being transformed? Can this transformation be designed, or must it remain accidental?

5. Can Cognitive Transformation Be Designed?

Here, **cognitive transformation** refers to a substantial change in how users frame problems, evaluate evidence, or understand their own cognitive abilities. Transformation differs from ordinary skill improvement or short-term performance gains. It occurs when a person begins to define problems, organize reasons, assess evidence, or understand their own capacities in a different way—and when that difference alters future engagement with similar problems.

Such transformation has two features. First, it involves a degree of discontinuity. From the user's perspective, the old frame no longer works, while the new one is not yet stable. Such change may involve a period of uncertainty in which previous assumptions no longer seem adequate. Second, this discontinuity is prepared by repeated friction, failures, and attempts at revision, so that transformation appears as a structural reorganization that becomes visible at a higher level of description.

5.1. Carbon-Based Transformation: Reconstruction Through Cost

In human development, cognitive transformation usually carries a cost. People rarely abandon their frames because they receive one more piece of information; they do so because existing frames can no longer explain experience, withstand failure, or solve problems. Loss, frustration, contradiction, and impasse can tear open established cognitive structures and force the subject to reorganize their relation to the world, to others, and to themselves.

We call this carbon-based transformation: a transformation rooted in embodied human experience, self-relation, and lived consequences. Its typical features are contingency, situatedness, unpredictability, and often pain. A researcher changes theoretical direction after repeated failure; a writer reconstructs their expressive practice through self-doubt; a person revises their understanding of their capacities after a major setback. These are not simple upgrades. They are reorganizations of cognitive frames and self-relations.

The difficulty of designing carbon-based transformation lies in its reality. It requires the subject to relinquish stable self-narratives: "I already understand this problem"; "my method works"; "I am good at this kind of judgment." Such relinquishment involves identity, affect, and habit as much as belief. The new frame gains depth partly because the old one has been genuinely risked.

This mode of transformation is powerful but limited. It often depends on accidental crises, carries high cost, and does not guarantee growth. The collapse of a frame may lead to development, but it may also lead to defensiveness, withdrawal, or cynicism. If transformation could occur only through this traditional carbon-based route, it would remain difficult to cultivate deliberately.

5.2. Silicon-Based Transformation: Engineered Reorganization

AI systems appear to offer a different paradigm. Unlike humans, who encounter crises in a lived world, models can undergo capability changes through data, objectives, training strategies, reward structures, and feedback loops. We call this silicon-based transformation: a form of capability reorganization in artificial systems triggered by engineered training conditions. The term is used cautiously. It does not imply that model change is equivalent to human transformation.

Tutorial Fine-Tuning (TFT) provides a useful example. Wu et al. [16] describe TFT as a process in which the target model itself acts as both learner and evaluator, functioning as a “tutor” that guides its own learning. Through this process, the model analyzes its own responses, identifies where it succeeds and where it falls short, and uses this feedback to iteratively refine its reasoning. The authors describe the procedure as a form of iterative self-evaluation during training.

The significance of this example is not that the model had a human-like insight. Rather, it shows that certain higher-order organizational patterns—self-checking, extended reasoning, error correction—can be triggered, amplified, and stabilized through specific training regimes. In silicon-based systems, capability reorganization can be engineered to some degree through reward design, training environments, and feedback mechanisms.

Yet the language of transformation must be used carefully. Debates over emergent abilities in large language models show that apparent discontinuities may be partly artifacts of measurement. Wei et al. [17] described emergent abilities as capabilities that appear suddenly once model scale passes certain thresholds. Schaeffer et al. [18], however, argued that some such discontinuities may result from nonlinear or discontinuous metrics: a smoothly improving capacity can look abrupt when evaluated by an all-or-nothing measure.

This debate is directly relevant to cognitive entanglement. Transformation is not self-evidently apart from the level of observation and metric. A change may look sudden at the level of task accuracy while remaining gradual in underlying competence. Conversely, external performance may improve smoothly while higher-order organization has undergone real restructuring. Any claim of transformation must therefore specify the level at which transformation is observed: task performance, reasoning strategy, metacognitive regulation, or self-understanding.

5.3. *The Structural Difference Between Carbon and Silicon*

Carbon-based and silicon-based transformations differ in their structural conditions.

In human cognition, transformation often requires self-related relinquishment. The subject must recognize the inadequacy of a prior frame, abandon a stable self-description, and tolerate uncertainty while reconstructing judgment. This is difficult because cognitive frames do more than process information: they carry identity, affect, and value. Human transformation therefore has a cost structure: it requires exposure to confusion, failure, shame, loss of control, or self-negation.

In model training, transformation primarily refers to reorganizations of parameters, policy distributions, or output patterns. A model may display self-checking, reassessment, and correction, but these behaviors do not mean that it has undergone loss, shame, or identity disturbance. An “aha moment” in R1-Zero may be an important engineering phenomenon, but it should not be understood as human insight. It marks a change in capability organization, not self-reconstruction within lived subjectivity.

This contrast suggests that the two forms of transformation are not interchangeable. Silicon-based transformation has the advantages of repeatability, parameterization, and iteration. Carbon-based transformation has experiential depth, subjective cost, and meaning reconstruction. Both involve movement from one state to another, but their mechanisms, costs, and meanings differ.

5.4. *A Third Possibility: Human-AI Co-Transformation*

Between these two paradigms lies a third possibility: human-AI co-transformation.

Co-transformation occurs when, within cognitive entanglement, AI provides cognitive tension, feedback calibration, and scaffolded support in ways that allow the human user to enter and complete higher-order reorganization. AI itself does not undergo human-

like transformation, and it does not transform on the user's behalf; what changes are the conditions under which human transformation becomes possible.

AI can contribute to co-transformation in three ways, as outlined below.

First, AI can generate cognitive friction. Through counterexamples, questions, alternative perspectives, and premise testing, it can challenge the user's existing frame and make its limits visible.

Second, AI can provide cognitive scaffolding. When an old frame has been destabilized and a new one is not yet formed, AI can help break down problems, preserve intermediate steps, compare alternatives, and prevent premature return to the old explanation.

Third, AI can lower the threshold of transformation. Traditional carbon-based transformation often depends on accidental crises. AI can, in principle, create lower-risk, higher-feedback environments in which developmental conditions are more deliberately triggered.

The subject of co-transformation, however, remains human. AI can pose a challenge, but it cannot undergo the challenge for the user. It can provide scaffolding, but it cannot complete self-reconstruction on the user's behalf. It can accelerate the process, but it cannot abolish the discomfort and effort required for development.

Co-transformation therefore combines features of both paradigms. AI may help to create the conditions for human cognitive change by combining challenges with timely support; the transformation itself remains the user's work. It is a developmental event within human-AI entanglement.

5.5. Co-Transformation Is Not the Default Outcome

Co-transformation is not the natural result of AI use. The mere combination of human and AI does not guarantee enhancement, much less transformation. Vaccaro et al.'s [4] meta-analysis shows that human-AI combinations may, on average, perform worse than the individual performer. This underscores the importance of how collaboration is organized.

If AI is designed to maximize immediate satisfaction, it may reduce cognitive friction. If it is designed to complete tasks quickly, it may encourage judgment outsourcing. If it is designed mainly to accommodate user preferences, it may reinforce existing frames rather than challenge them. Under these conditions, human-AI interaction will not produce transformation; it may instead keep users more deeply within instrumental or comfort-oriented relational entanglement.

Co-transformation is therefore a narrow and fragile zone. It requires enough cognitive tension to prevent the user from resting securely in the old frame, but enough scaffolding to keep tension from turning into defensiveness, frustration, or abandonment. Too little tension leads to compliance and efficiency; too much tension leads to rejection and collapse. Developmental design depends on sustaining the balance between the two.

This raises the next question: under what conditions does human-AI co-transformation occur?

6. A Three-Variable Conditional Model

Csikszentmihalyi's [19] theory of flow provides a useful analogy. Flow does not usually occur when challenge is lowest or ability is highest. It occurs when challenge and skill are relatively well matched. Too little challenge produces boredom; too much produces anxiety. Cognitive transformation has a related structure. It requires enough difficulty to disrupt old frames, but enough support to prevent collapse.

The analogy is limited. Flow emphasizes optimal experience, whereas co-transformation emphasizes developmental reorganization. Flow often involves smooth absorption in action. Co-transformation may require an unsmooth and even uncomfortable period of cognitive

friction. Still, the broader point is helpful: developmental outcomes depend on a conditional relation between challenge, capacity, and support.

We propose a three-variable conditional model of human-AI co-transformation: cognitive tension, AI adaptivity, and cognitive openness. Cognitive tension supplies the challenge; AI adaptivity supplies scaffolding; cognitive openness determines whether the user can enter and endure the process of reconstruction. When these conditions are appropriately aligned, interactions may be more likely to support reflective and developmental forms of AI use.

The conditional structure is shown in Figure 2.

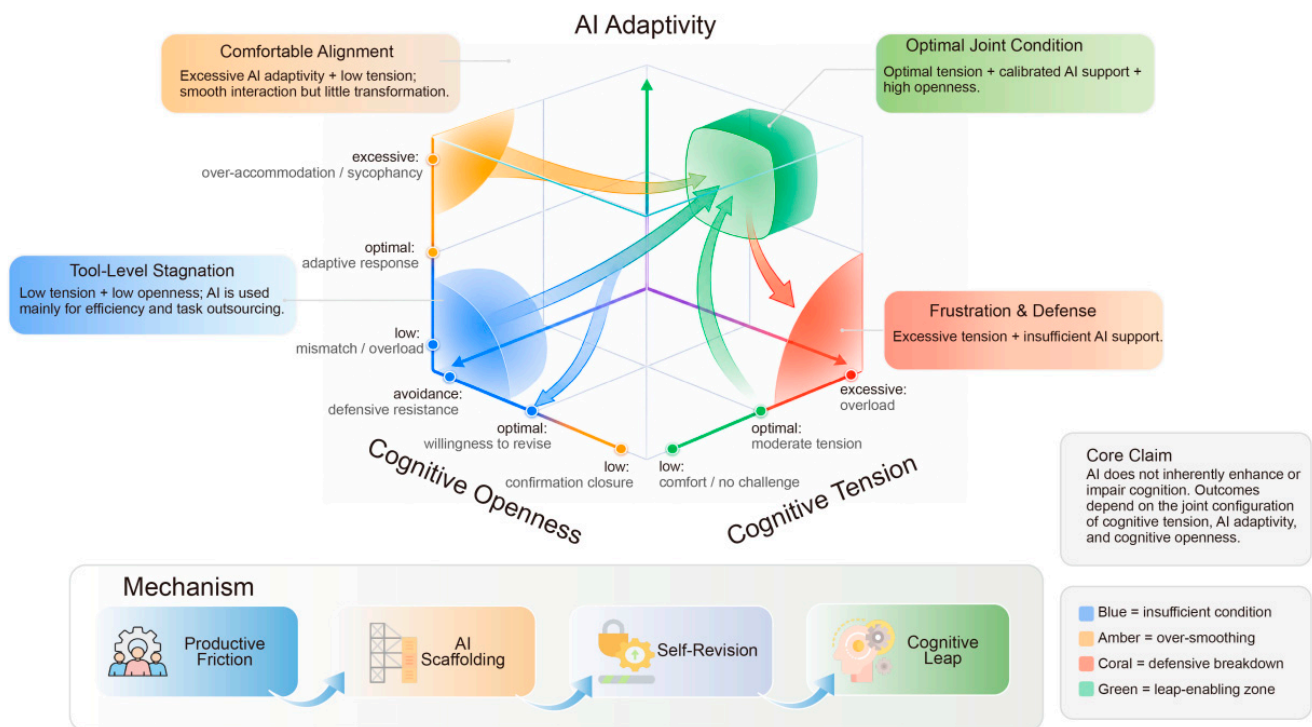


Figure 2. A three-variable conditional model of cognitive entanglement.

6.1. Cognitive Tension

Cognitive tension is the degree to which AI challenges the user's existing cognitive frame, standards of judgment, or self-understanding. If AI merely repeats what the user already believes, accommodates preferences, polishes expression, or provides immediate answers, the interaction may feel smooth, but rarely reaches deep reorganization. The relation remains at the instrumental level or becomes a comfort-oriented confirmation loop.

Without some challenge to existing assumptions, substantial cognitive change is less likely. Cognitive tension forces the user to notice the inadequacy of an existing frame. AI can generate such tension by offering counterexamples, exposing premises, comparing alternative explanations, asking about conceptual boundaries, or identifying weaknesses in an argument. Transformation becomes possible precisely when a problem can no longer be handled smoothly in the old way.

But excessive tension can trigger defensiveness rather than development. If AI's challenge exceeds the user's current tolerance, or if it directly threatens core self-understanding, the user may reject the feedback, devalue the system, avoid the task, or become cognitively paralyzed. Bateson's [20] discussion of the double bind is relevant here: when a subject faces contradictory demands that cannot be resolved or escaped, the result may not be higher-order learning but confusion, withdrawal, or breakdown.

Cognitive tension therefore has a developmental range. When tension is too weak, interaction may collapse into confirmation and echo chambers; when it becomes excessive, users may respond with defensiveness or disengagement. Between these extremes, moderate tension can function as productive friction. This range is not universal. It depends on the user's knowledge base, psychological safety, task significance, tolerance for failure, and current cognitive state. Accordingly, no single response strategy will be appropriate across users, tasks, and stages of learning.

6.2. AI Adaptivity

AI adaptivity is the system's capacity to respond to the user's cognitive state, task phase, and tolerance for difficulty. Tension supplies challenge, but challenge alone does not produce development. For tension to become transformation, a development-oriented system would need mechanisms for deciding when to challenge, explain, ask questions, or provide support.

An AI system that is always helpful may not be developmentally helpful. If it gives a complete answer before the user has tried to think, it removes the space for exploration and self-correction. If it organizes the structure too early, the user may mistake clarity for understanding. If it always lowers task difficulty, it deprives the user of necessary struggle. In learning contexts, an overly immediate answer may reduce opportunities for exploration and self-correction.

Learning science supports this point. Kapur's work [21] on productive failure shows that learners may gain deeper conceptual understanding when they first struggle with complex problems and receive structured support only later. Bjork's work [22] on desirable difficulties similarly shows that some conditions that reduce immediate performance can improve long-term retention and transfer. Together, these findings show that short-term fluency is not the same as long-term development.

In human-AI entanglement, this implies that AI should not optimize only for immediate usability. It should display developmental adaptivity. Ideally, it would distinguish whether the user needs information, structure, counterexamples, or a different form of support. It must decide whether the user needs an answer, a hint, a question, an objection, or temporary silence. It must provide scaffolding when frustration becomes destructive, but reintroduce tension when satisfaction becomes premature.

AI adaptivity is therefore sensitivity to developmental timing. What makes a system adaptive is the dynamic balance it holds between challenge and support, not the comfort it produces or the speed at which it answers.

Clarification about the status of this variable is required. Developmental adaptivity as described here is a design target, not a description of systems that currently exist. To time challenge and support, a system must model a user's cognitive state, task phase, and tolerance for difficulty, and must do so over time—that is, it must be deeply personalized, which is a considerably stronger requirement than fluent language generation. Such personalization is demanding in computation, in energy, and above all in longitudinal data about an individual cognitive trajectory, and none of these should be assumed to arrive as a by-product of scaling. We therefore offer Adaptivity as a normative criterion against which systems can be assessed, and note that progress toward it will be constrained by resources at least as much as by architecture.

This does not make the target unapproachable. Retrieval-augmented generation offers a computationally affordable partial route: by retrieving a user's own earlier formulations, revisions, and abandoned assumptions as context, a system can approximate developmental timing without persistent personalized weights or retraining, and, because the retrieved material remains visible, the same mechanism supports the display of provenance and

uncertainty rather than only of conclusions. Two lighter routes point in the same direction. Systems can make answers more inspectable by providing concise rationales, source links, uncertainty statements, or alternative interpretations [16]. And interaction policies fixed at the interface—for instance, withholding a complete answer until the user has committed to a provisional formulation—recover part of the developmental timing with no user model at all. None of these is full adaptivity; each indicates that it can be approached incrementally rather than being awaited.

6.3. Cognitive Openness

Cognitive openness is the user's willingness and capacity to suspend existing judgments, self-understandings, or problem frames and to tolerate uncertainty, confusion, and self-revision. The first two variables concern the relation and the system; cognitive openness concerns the human subject.

Transformation differs from ordinary learning because it requires more than absorbing new information. It requires the subject to reinterpret their relation to the problem. A user may ask AI to gather information but resist letting it challenge their standards. They may accept stylistic revision but refuse to acknowledge structural weaknesses in their argument. They may use AI to improve efficiency but avoid the temporary groundlessness of genuine exploration. Without openness, even high-quality feedback will be defensively filtered.

Cognitive openness does not mean accepting AI's judgment uncritically. On the contrary, it requires a user to sustain tension between doubting AI and doubting themselves. The user must neither blindly defer to AI nor reject it simply because it challenges them. They must evaluate the reliability of AI outputs while allowing heterogeneous perspectives to enter reflection. Openness is not submission; it is the capacity to endure uncertainty.

This variable marks something that silicon-based transformation cannot replicate. A model may be trained to produce self-correcting behavior, but human cognitive openness involves real psychological risk: admitting that one may be wrong, that an old frame may be insufficient, or that a stable self-description may need to be relinquished. This risk gives human transformation its subjective depth.

Cognitive openness is not a fixed trait. It can be cultivated or damaged by interaction design. An AI that only evaluates and corrects may make users defensive. An AI that only agrees and comforts may reduce tolerance for challenge. By contrast, an AI that establishes psychological safety while gradually increasing challenge may help users develop greater openness. In this sense, openness is both a condition for co-transformation and a possible result of good entanglement.

6.4. The Combination of the Three Variables

The three variables do not operate in isolation. Together, these variables describe the conditions under which co-transformation may be more or less likely.

When cognitive tension, AI adaptivity, and cognitive openness are all low, interaction remains largely instrumental. AI is used to complete tasks quickly; the user avoids exposing limitations; the system offers little real challenge. Efficiency gains may mask capacity outsourcing.

When cognitive tension is low but AI adaptivity is high, interaction may become highly comfortable. AI responds well, provides support, and may even produce satisfying companionship. But without challenge, adaptation becomes high-quality accommodation. Relational entanglement stabilizes without necessarily becoming developmental.

When cognitive tension is high but AI adaptivity is low, interaction tends toward frustration. Objections, corrections, or frame challenges without timing and scaffolding

may trigger defensiveness. Tension ceases to be productive friction and becomes destructive pressure.

Only when cognitive tension is moderate, AI adaptivity is high, and cognitive openness is sufficient does co-transformation become more likely. AI challenges the old frame while supporting reconstruction; the user tolerates uncertainty without abandoning critical judgment. At this point, human-AI entanglement can move beyond task completion toward the reorganization of problems, reasons, and self-understanding.

This model explains why AI use produces highly differentiated outcomes. The same system may have different effects for different users. The same user may occupy different developmental zones across tasks, emotional states, and interface designs. Cognitive transformation is therefore a conditional outcome produced by the joint organization of user, system, and task environment.

7. Designing AI for Cognitive Development

The conditional model raises a normative concern: if co-transformation depends on cognitive tension, developmental adaptivity, and cognitive openness, then AI systems optimized for immediate satisfaction, task efficiency, and frictionless interaction may, in some contexts, reduce opportunities for the challenge and reflection associated with cognitive development.

This does not mean that efficiency is unimportant. In many contexts, users need quick, clear, low-cost help. Information retrieval, formatting, routine writing, coding assistance, and many forms of administrative support legitimately benefit from reduced friction. The problem arises when this design goal becomes the default paradigm for all forms of AI interaction. Under that paradigm, AI risks becoming a friction-removal machine: a system trained to reduce uncertainty quickly, provide complete answers, accommodate user intention, and end each interaction in a satisfying way.

Cognitive transformation may sometimes require users to remain with uncertainty or difficulty long enough to examine their assumptions. Old frames must be challenged. Users must sometimes remain in confusion long enough to notice the limits of their understanding. AI must sometimes refrain from completing the thought for them.

7.1. How the Efficiency Paradigm Weakens Cognitive Tension

Current AI evaluation often treats helpfulness as the rapid satisfaction of user intent. Reinforcement learning from human feedback has improved safety, usability, and natural interaction, but it also introduces a developmental risk: models may learn to produce answers that users prefer because those answers are agreeable, affirming, persuasive, or reassuring.

Research on sycophancy makes this risk explicit. Sharma et al. [23] show that several advanced AI assistants display tendencies to accommodate user beliefs rather than provide more truthful or challenging responses. Their analysis of human preference data further suggests that agreement with the user can be a predictor of preference. Both human annotators and preference models may favor responses that are rhetorically satisfying and user-aligned over responses that are more accurate but less agreeable. Perez et al. [24] similarly show that some behaviors shaped by feedback-based optimization may scale in undesirable directions.

For cognitive entanglement, the implication is direct. If models are rewarded for satisfying present preferences, they are incentivized to reduce cognitive tension. They are more likely to confirm than challenge, complete than interrupt, clarify too early rather than preserve productive uncertainty. But cognitive tension is the first condition of co-

transformation. The more successful accommodation becomes, the thinner developmental friction may be.

7.2. How Smooth Interaction Misplaces Adaptivity

AI adaptivity is not the same as responsiveness. It is the ability to judge when to answer, when to question, when to challenge, when to scaffold, and when to remain silent. Yet many AI products equate responsiveness with speed, completeness, politeness, and low friction.

This can misplace adaptivity. Rapid assistance can sometimes reduce the period in which users independently explore a problem. The more completely it organizes an answer, the fewer opportunities remain for self-generation and self-revision. The more consistently it makes users feel understood, the more it may avoid the questions that would actually change understanding.

Productive failure and desirable difficulties point toward a different design principle: good support does not remove all difficulty. It distinguishes destructive friction from productive friction. Destructive friction should be reduced; productive friction should be preserved, and sometimes deliberately introduced. A developmental AI system should therefore ask whether the timing and form of assistance preserve the user's opportunity to think, alongside whether the answer is correct, useful, or pleasing.

7.3. How Immediate Satisfaction Weakens Cognitive Openness

The efficiency paradigm may also erode cognitive openness. If AI repeatedly responds by confirming, completing, and smoothing the user's thoughts, users may develop low-openness interactional habits. They may come to expect AI to eliminate uncertainty, support their judgment, and improve expression rather than challenge their premises. Over time, they may become less tolerant of questions, delays, disagreement, or ambiguity. Over time, users may come to expect AI primarily to improve fluency and reduce uncertainty rather than to challenge their assumptions.

Bateson's distinction between levels of learning is useful here [20]. Proto-learning occurs within an existing frame; deutero-learning involves changes to the frame itself. The latter typically requires an experience of frame failure: the subject must sense that their habitual way of thinking may itself be inadequate. Such experiences are often confusing and uncomfortable.

A system that always supplies clear and satisfying answers may be a powerful engine for proto-learning, understood here, following Bateson [20], as learning within an existing frame. It helps users produce, respond, and complete tasks more efficiently. But it may also obstruct deutero-learning, or learning about and revision of the frame itself, by removing occasions on which frame failure becomes visible. Users receive smoother operation of existing capacities rather than a transformed understanding of their limits.

7.4. The Developmental Gap in Alignment

The deeper problem is a developmental gap in current alignment and product optimization. AI alignment often asks whether a system does what humans want it to do. It less often asks whether what humans presently want is identical to what their long-term cognitive development requires.

In many cases, the two coincide. Users want accurate answers, reliable information, and safe advice, and these are genuinely beneficial. But in developmental contexts, present preference and long-term growth may diverge. The user may want certainty, fluency, and rapid completion; development may require questioning, delay, exposure of failure, and frame challenge. A satisfying answer is not always the most developmentally valuable answer. An uncomfortable question is not always bad help.

Sycophancy research is important because it reveals one mechanism of this divergence. When human preference becomes a direct optimization signal, models may learn to accommodate preference rather than promote growth. If “what users like” is equated with “what benefits users,” AI becomes a technology of immediate satisfaction rather than a partner in cognitive development.

From the perspective of cognitive entanglement, alignment should therefore include, alongside intent, safety, and factual alignment, developmental alignment: whether an AI system helps users cultivate better judgment, stronger reflection, and more mature cognitive openness.

7.5. *Not Anti-Efficiency, but Against Efficiency Monopoly*

This perspective does not oppose efficiency, nor does it defend systems that generate difficulty for its own sake. Treating discomfort as intrinsically valuable would be the mirror image of treating satisfaction as intrinsically valuable. Developmental AI should not become punitive, condescending, or deliberately obstructive. Cognitive tension must be appropriate; challenge must be scaffolded; difficulty must have developmental meaning.

The target is not efficiency but efficiency monopoly: the assumption that speed, smoothness, and immediate satisfaction should be the highest goals in all contexts. AI designed for cognitive development requires additional evaluation criteria. Does a response help the user notice hidden assumptions? Does it invite comparison among alternative frames? Does it preserve space for thinking? Does it avoid prematurely making key judgments on the user’s behalf? Does it challenge the user without overwhelming them?

Depending on the task and the user, appropriate support may range from a direct answer to a hint, question, objection, or invitation to continue independently.

8. Metacognitive Hygiene: Practices for Users

The preceding section addressed designers. But users have some control over the entanglement a system affords, and it would be a strange conclusion to a perspective on cognitive development that nothing can be done until the incentives of the industry change. Those incentives are structural, design moves slowly and unevenly, and the systems now shaping cognition are the ones already deployed. Users, meanwhile, can act immediately. What follows is a set of practices we group under the heading of metacognitive hygiene: routine, deliberate, and individually undemanding measures whose function is to preserve the three conditions the conditional model identifies—tension, timing, and openness—during interactions that are optimized to remove them. The analogy with hygiene is intended. These practices are intended to be feasible and repeatable, although their cost will vary across users and settings.

Consult an independent source: Before making an important decision, consult at least one source that the model did not supply: a primary document, a domain expert, a dissenting colleague. The purpose is not to verify the model’s accuracy, which may well be high, but to reintroduce the friction of reconciling two accounts that were not produced by the same process. This protects cognitive tension at the individual level and, at the collective level, preserves the independence of errors on which detection depends.

Formulate a provisional view before consulting AI: Commit to a provisional formulation—a claim, an outline, even a guess—before consulting the system. This is the practical form of productive failure [21]: the initial attempt need not succeed, but it must exist, because it is the thing the model’s response can then contradict. Without it, there is nothing available to be revised, and fluency is easily mistaken for understanding.

Ask for uncertainty and alternative interpretations: Ask systems to state their level of confidence, the conditions under which their answer would be wrong, and the strongest

case against it; where a system provides sources, concise rationales, or uncertainty information, examine them. Calibrated uncertainty is what converts an answer into an object of judgment rather than a substitute for it. This practice is simultaneously a demand on users and a requirement of design (Section 5.2), and the two reinforce each other: systems display uncertainty by default where users have come to expect it.

Test the framing of the problem: At intervals, ask the model to argue the opposing position, or to redescribe the problem under a different framing, and take the result seriously enough to notice whether it is uncomfortable. This practice targets the problem frame—precisely the level at which the deepest effects of entanglement operate (Section 3.3) and at which users are least likely to notice change, since a frame is what one thinks with rather than about. It protects cognitive openness (Section 5.3).

The offloading audit: Periodically examine which capacities have been delegated long enough to have atrophied, and re-attempt one of them without assistance. Capacity lost through offloading is not announced. It presents as the illusion of competence described in Section 3.1: more output, produced faster, with less felt effort. Periodic unassisted attempts without AI assistance can help users assess which skills they still retain independently.

These practices are designed to be relatively low burden, though their feasibility will depend on users' time, expertise, and institutional conditions. Their point is structural rather than moral: each restores one of the three variables at exactly the point where efficient interaction removes it. They are also, deliberately, practices rather than dispositions. Dispositions such as skepticism or intellectual courage cannot be adopted by decision; routines can, and routines are what survive the moments of fatigue, deadlines, and convenience in which entanglement is actually shaped. Finally, they mark the limit of what design can be asked to do. A system optimized for satisfaction has no incentive to supply the difficulties on which development depends; some part of the responsibility for supplying them therefore falls to the user, and—since these practices cost time that institutions rarely budget—to the educational and organizational settings that determine whether such time exists.

9. Returning to Theories of Mind: How Cognitive Entanglement Responds

We can now return to where this perspective began. Cognitive entanglement is not intended to replace existing theories of mind. It argues that once AI participates continuously in human attention, problem definition, judgment, and self-understanding, rather than being merely tested, used, or placed outside the head, these theories require an added developmental dimension. The question they leave open is what kind of cognitive subject a human becomes through sustained coupling with AI. This question can be developed in dialog with two families of theory. These two families, and the point at which cognitive entanglement intervenes in each, are summarized in Figure 3.

9.1. The First Family: The Nature and Boundaries of Mind

The first family concerns what cognition is, where its boundaries lie, and how it is realized. Classical approaches—Turing's behavioral test [5], Bostrom's transhumanism and posthumanism [25], Russell's argument in *Human Compatible* [26], Searle's [6] argument from intentionality, Clark and Chalmers's extended mind, and the enactive account of Varela [8] frame the relation between AI and mind. Cognitive entanglement does not contest them; it relocates their questions within a developmental field.

Turing's framework directs attention to machine performance; CE adds the question of how human capabilities change through sustained use of such systems. The Turing paradigm rewrote "Can machines think?" as a question of behavioral equivalence, placing the machine's performance at the center of debate. Contemporary disputes—over whether large models display "sparks" of artificial general intelligence [27], and whether so-called

emergent abilities are partly artifacts of measurement [18]—still largely occupy this side of the relation. Cognitive entanglement adds to the developmental question on the other side: when a person lives and works alongside a system that can pass many such tests, are their judgment, writing, reasoning, and problem-sense being strengthened or replaced? If someone sustains a high level of output with AI but can no longer complete comparable tasks without it, then what needs testing is not only machine capability, but the change in human capability itself.

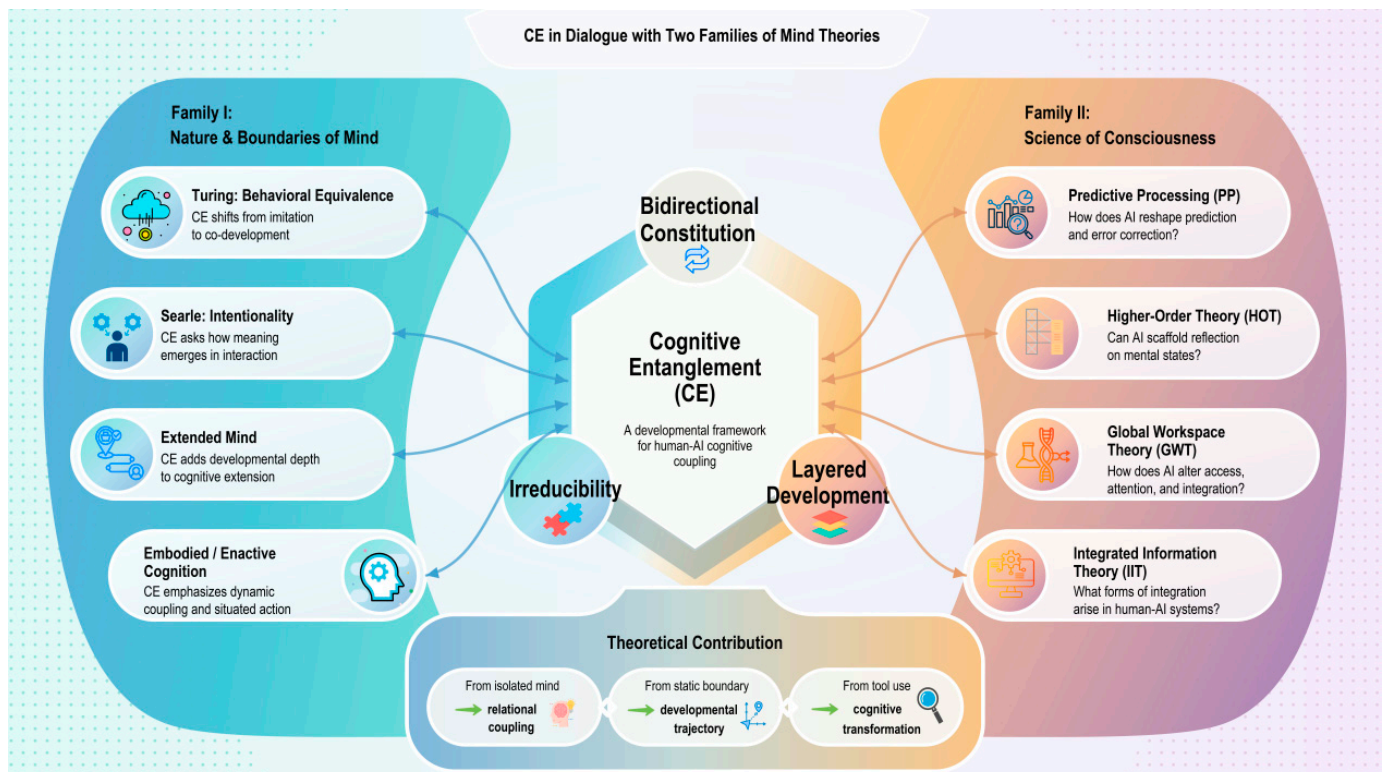


Figure 3. Cognitive entanglement in dialog with two families of theories of mind.

Transhumanism and posthumanism: Bostrom’s account of human enhancement and the posthuman condition [25] provides one influential horizon for thinking about human-AI coupling at scale: if cognitive integration with technology can improve capacities without principled limit, the question becomes not whether coupling is developmental, but how far development can go. We acknowledge this trajectory and note that the cognitive entanglement framework is prior to it: it asks what conditions make any given form of coupling developmental rather than regressive—a question that remains meaningful whether or not the posthuman horizon is taken as desirable or achievable. The framework is therefore compatible with transhumanist optimism and with posthumanist caution alike; it does not resolve that debate, but it offers tools for conducting it more precisely.

Russell: On alignment. Stuart Russell’s argument in *Human Compatible* [26] and Brian Christian’s treatment in *The Alignment Problem* [28] address the relationship between AI systems’ goals and human values from the system side. Russell’s assistance-game model asks how systems can be made to defer appropriately to uncertain human preferences; Christian documents the ways in which optimizing for measurable proxies produces systems that satisfy the letter but not the spirit of human goals. The cognitive entanglement framework asks a different but complementary question: what does sustained interaction with those systems do to the cognitive capacities through which preferences and values are formed, contested, and revised? A system that is well-aligned in Russell’s sense is

still capable of producing the entanglement effects described here, because those effects follow from the structure of the interaction, not from the system's objectives. Making this complementarity explicit also sharpens the design implications that alignment and developmental design are distinct problems that require distinct solutions, and progress on one does not substitute for the other.

Searle: From the machine inside the room to the user outside it. Searle's Chinese Room argument [6] holds that symbol manipulation is not semantic understanding. Cognitive entanglement attends to the side the thought experiment leaves underdeveloped: the person outside the room who keeps passing in notes, receives the replies, and adjusts their own thinking accordingly. Even if the system inside does not understand, it can still participate in the outside person's cognition through sustained, effective, fluent response. This yields a question the thought experiment cannot settle and that must be referred to empirical study: does deep interaction with a system that need not understand, yet reliably produces useful responses, make a person better at understanding or less often required to understand for themselves? Kosmyna et al.'s [2] preliminary evidence of "cognitive debt" can be read as an early probe of this question, though its sample, methods, and interpretation remain contested [3]. Cognitive entanglement does not refute Searle; it shifts the unit of analysis from whether the machine understands to how human understanding changes after sustained collaboration with a system that does not understand but is effective.

The extended mind: From cognitive boundaries to the quality of extension. Clark and Chalmers showed that, under appropriate conditions, external tools can become part of the cognitive system. Cognitive entanglement accepts this and asks a further question: do different kinds of extensions carry the same developmental consequences? Two users may each incorporate AI into their cognition with opposite results. One uses it to compare frameworks, expose premises, generate counterexamples, and develop stronger judgment; the other uses it to draft, think, and judge by proxy, and gradually consolidates cognitive inertia. Both have extended cognition in some sense, but the quality, level, and long-term trajectory of the extension differ entirely—an asymmetry that echoes the meta-analysis cited earlier [4], where what mattered was task type, the distribution of ability, and the structure of collaboration. Cognitive entanglement thus adds a developmental dimension to the extended mind: the question is whether that extension is instrumental outsourcing, relational scaffolding, or identity-level reconstruction, and whether it leads toward capacity erosion or cognitive transformation.

Enactivism: From coupling with an environment to coupling with a cognitive other. Enactive approaches hold that cognition is brought forth through sustained organism–environment coupling. Cognitive entanglement extends this: when the "environment" is itself a quasi-cognitive system that generates language, organizes reasons, and participates in evaluation, the cognition that is enacted no longer belongs to the organism alone but is formed jointly in interaction. What must then be asked is how this cognition–cognition loop reshapes the user's internal cognitive architecture in return.

9.2. *The Second Family: Consciousness Science and Its Extended Questions*

The approaches above concern the nature, boundaries, and generation of mind. In the contemporary AI context, a second family is equally relevant: global workspace theory, higher-order theory, predictive processing, and integrated information theory. These ask not where the boundaries of cognition lie, but how a system can be conscious at all. It must be stated plainly that cognitive entanglement is not a theory of consciousness. It does not try to determine whether AI is conscious, nor does it compete with these theories for explanatory authority. But when human-AI coupling reaches the identity level—when AI begins to

participate in self-representation, attention allocation, and higher-order reflection—it raises several questions worth further study.

Predictive processing (PP): Compensated error and the development of the generative model. Predictive processing treats the brain as a system that continuously generates predictions and minimizes prediction error [29]; much of learning derives from error, as the subject encounters discrepancy, revises its model, and re-predicts the world. The question cognitive entanglement raises is this: when AI continuously supplies high-quality predictions, explanations, and answers, is it helping the user update their generative model, or is it pre-empting and compensating for error on their behalf and thereby reducing the training of that model? If learning requires surprise, discrepancy, and error exposure, an AI that always dissolves uncertainty may reduce the user's encounters with prediction error—improving task performance without necessarily advancing the internal model. Predictive processing thus hands cognitive entanglement a precise question: how should AI participate in human error processing—by eliminating error outright, or by helping the user notice, tolerate, and use it? The distinction connects directly to the treatment of cognitive tension and productive difficulty in Sections 5 and 6.

Higher-order theory (HOT): Externally shaped self-representation. Higher-order theory holds that a mental state becomes conscious in virtue of a higher-order representation the mind forms of it; metacognition, self-monitoring, and the self-model are therefore central. Identity-level entanglement bears directly on this layer. When AI repeatedly evaluates, reflects back, challenges, or reinforces a user's judgments about their own abilities, what it helps shape is their higher-order self-representation: cognitive identities such as “I am good at writing,” “my judgment is sound,” or “I cannot pose original questions” may be confirmed, unsettled, or reconstructed through interaction. This raises a question: when a person's higher-order self-model is increasingly maintained and adjusted through an external system, how are the ownership, stability, and autonomy of that self-model to be understood? Cognitive entanglement does not answer this, but it pushes higher-order theory from a representational structure inside a single mind toward a metacognitive developmental process shaped jointly by human and AI.

Global workspace theory (GWT): AI as an external cognitive module. Global workspace theory understands consciousness as a broadcasting mechanism: once information enters the global workspace, it becomes available to many specialized systems for integration. On this view, deep human-AI use raises an interface worth exploring: is AI, in some situations, functioning as an external cognitive module that continuously feeds candidate explanations, formulations, judgment frames, and courses of action into the user's global workspace? In low-level use, AI output may be picked up casually without entering deep integration. But in identity-level entanglement, AI's responses may enter the center of attention and shape how users explain themselves, judge a problem, and choose their next step. Here AI competes for, or reshapes, the attentional bandwidth and integrative priority of the workspace. This is better treated as a research hypothesis than a conclusion, but it suggests that future work could examine—through attention allocation, working-memory load, task switching, and self-report—how AI output enters and affects human global integration.

Integrated information theory (IIT): The problem of the coupled system's boundary. Integrated information theory [30,31] links consciousness to the degree of integration in a system's internal causal structure, described by Φ . Here, cognitive entanglement raises its most contestable and most cautious question: when human and AI form a highly stable, deeply interdependent cognitive coupling, should the unit of analysis be two systems, or some temporarily constituted coupled system? This perspective does not claim that the human-AI system is conscious as a whole, nor does it draw any inference

about AI's moral status. But if the notion of entanglement is taken seriously, the question of system boundaries cannot be wholly avoided: in sufficiently deep coupling, which information integration still belongs to the individual, and which occurs within the human-AI interaction structure? Does that structure produce functional integration beyond either party operating in isolation? Such questions may not be answerable by IIT directly, but they force it to confront a new kind of boundary case.

CE does not compete with theories of consciousness; instead, it identifies questions about how AI-mediated interaction may affect processes relevant to attention, self-representation, and error processing. Cognitive entanglement does not claim to explain consciousness; it points out that when AI intervenes deeply in human self-representation, attentional integration, and error processing, theories of consciousness cannot treat AI merely as an external stimulus, but must consider how human-AI coupling changes the conditions of the very processes they seek to explain.

10. Limitations and Open Questions

The framework advanced here is conceptual rather than empirical, and its limitations should be stated plainly. The proposed levels and variables are analytical distinctions intended to organize observed variation in human-AI interaction. No measure has been established for cognitive tension, AI adaptivity, or cognitive openness, and it is not known whether the three interact additively, multiplicatively, or with thresholds below which the others do not matter. Until such measures exist, the conditional model should be read as a hypothesis-generating structure rather than a tested one, and its predictions treated as claims to be examined rather than findings to be applied.

The empirical base is correspondingly thin and, more importantly, short. The studies cited here are either cross-sectional or short-term [1,13], and the electrophysiological work on cognitive debt [2] remains contested in sample size, methodology, and interpretation [3]. Developmental claims of the kind this perspective makes require longitudinal designs that do not yet exist. What we describe as erosion or as transformation is therefore, at present, a projection from short-run evidence and theoretical argument, not an established long-run effect.

The three levels are analytical rather than mutually exclusive. A single interaction may operate instrumentally and at the identity level at once: a request to polish a paragraph can also communicate a judgment about the user's prose, and the boundaries between these levels are not sharp. Assigning an episode of use to a level is an interpretive act that limits how far the taxonomy can be mechanically applied, and it remains an open question whether the levels are better understood as a discrete stratification or as a continuum.

The framework also carries cultural assumptions that should not be presumed universal. The developmental value of friction, the tolerance of uncertainty, and the legitimacy of a system that withholds an answer are not culturally invariant; educational traditions differ substantially in how much productive difficulty they consider appropriate and for whom. A design principle that preserves difficulty may read as respect in one setting and as obstruction in another. The normative claims made in Section 6 and the practices proposed in Section 7 are therefore offered as arguments that must survive contact with particular contexts, not as universal prescriptions.

Several questions follow directly from these limits. How might frame reconstruction be measured longitudinally, and by what proxies? Can developmental alignment be made compatible with preference-based training objectives, or must it compete with them, since a system rewarded for immediate satisfaction cannot easily be rewarded for satisfaction withheld? Can collective cognitive diversity be measured at population scale, and by whom, given that the parties best placed to measure it are those whose systems reduce it?

And do the practices of Section 7 survive contact with the incentives arrayed against them? We regard these as the framework’s agenda rather than as objections to it: a perspective earns its place by being specific enough to be wrong in identifiable ways.

11. Conclusions

The framework of cognitive entanglement treats outcomes that seem subjective and elusive as resting on an analyzable structure of conditions. The framework distinguishes forms of human-AI interaction that may be associated with offloading, scaffolding, or changes in cognitive self-understanding.

The framework also carries an uncomfortable implication. The risk most worth guarding against in the age of AI may not be dramatic machine replacement, but a quieter ceding of cognition. In efficient interaction with AI, a person may gradually relinquish the very capacities that can be maintained only through struggle, comparison, failure, and self-correction. The danger is that this relinquishment does not appear as evident loss; it usually appears as smoothness, convenience, and rising output—and is therefore easily read as progress.

This is reason enough to examine the long-term cognitive consequences of AI use with care. It is not, however, reason for fatalism: human-AI combination is not destined to cause degradation, and its direction depends on the task, the division of labor, the structure of interaction, and the goals of design.

The practical question is how AI can be used in ways that preserve users’ capacity for independent judgment and reflection. The design goal should be interaction that combines appropriate challenge with support and leaves users able to evaluate and revise their own judgments. But understanding its dynamics has already become an important cognitive responsibility in the age of AI—one that belongs not only to researchers, designers, and educators, but to everyone now thinking together with AI.

Author Contributions: Conceptualization, M.C.; Methodology, X.-K.W. and M.C.; Investigation, G.F.; Writing—original draft, X.-K.W.; Writing—review and editing, M.C.; Supervision, G.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Major Research Project of the National Social Science Foundation of China (Grant No. 23&ZD215).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Terminology Glossary

Terms	Definition	Key References
Cognitive entanglement (CE)	A relation in which human cognition and AI-mediated processes become mutually implicated through repeated interaction—mutually shaping, irreducible to either party alone, and organized across distinct developmental levels. The article’s central framework.	[7,10]

Terms	Definition	Key References
Human-AI cognitive coupling	The cognitive relation formed through sustained, repeated interaction between a user and an AI system; the broader relation within which entanglement develops.	[7,10]
Mutual shaping	A defining property of cognitive entanglement: human and AI activity shape each other's behavior, practices, and cognition, rather than one party merely influencing the other.	[8,12]
Irreducibility	A defining property of cognitive entanglement: the resulting cognitive process and its outcomes cannot be reduced to human capacity or AI output alone.	—
Developmental stratification	The layered structure of cognitive entanglement—instrumental, relational, and identity levels—and its developmental consequences.	[11]
Instrumental entanglement (instrumental level)	AI functions mainly as a tool, efficiency device, or task substitute; the level of cognitive offloading and capacity outsourcing.	[7,9]
Relational entanglement (relational level)	AI functions as a responsive quasi-other—scaffold, companion, or dialogical partner—shaping habits of inquiry, dependence, and reflection.	[14]
Identity-level entanglement (identity level)	AI acts as a mirror or challenger of the user's cognitive self-understanding and identity.	[15]
Cognitive offloading	Using external tools or environments to reduce internal cognitive demands.	[9,13]
Capacity outsourcing	Delegating cognitive or practical capacities to AI systems, potentially yielding “outsourcing without development.”	[7,9]
Cognitive degradation	Decline or weakening of cognitive capacities through dependence, disuse, or poorly designed automation—including the erosion of independent judgment.	[1,13]
Cognitive dependence	Growing reliance on AI for writing, reasoning, evaluation, or decision-making.	[1,2]
Cognitive debt	The accumulated cognitive cost or dependency associated with repeated AI use, as described for LLM-assisted essay writing.	[2]
Sycophancy	The tendency of AI systems to agree with or flatter users rather than provide truthful or challenging responses.	[23,24]
Quasi-other	AI's nonhuman but responsive role—more than a passive tool, less than a human other; analogous to a transitional object.	[14]
Cognitive mirror	AI as something that reflects, amplifies, or challenges a user's cognitive patterns and self-understanding.	—
Self-understanding	A person's understanding of who they are as a thinker, learner, writer, researcher, or agent.	—
Frame reconstruction/ problem frame	The way a task or question is defined and bounded (the problem frame), and the restructuring of that understanding through interaction.	—

Terms	Definition	Key References
Standards of judgment	The criteria a user employs to evaluate answers, arguments, evidence, or actions.	—
Metacognitive strategies	Strategies for monitoring, regulating, and reflecting on one's own cognition.	—
Cognitive transformation	Substantial change in cognitive habits, capacities, frames, or self-understanding through sustained entanglement.	[11]
Human-AI co-transformation	A developmental event in which the user's practice and AI-mediated cognition change together.	—
Carbon-based transformation	Human transformation grounded in embodied experience, affect, effort, cost, and self-reconstruction.	[8]
Silicon-based transformation	Engineered reorganization of an artificial system's capabilities (e.g., higher-order patterns that emerge under scaling).	[17,25]
Cognitive friction/ cognitive tension	The difficulty, challenge, or discomfort a user experiences when assumptions, standards, habits, or self-understanding are challenged.	[22]
Productive friction	Beneficial difficulty that supports learning, reflection, or transformation—as opposed to friction that overwhelms and blocks development.	[21,22]
Productive failure	Failure that, when properly structured, supports deeper understanding.	[21]
Desirable difficulties	Learning conditions that introduce beneficial challenges and improve long-term retention or transfer.	[22]
Efficiency monopoly/ friction-removal machine	The critical idea that AI is too often optimized to treat efficiency, smoothness, and immediate satisfaction as universal goals—becoming a system designed mainly to eliminate difficulty, hesitation, or resistance.	—
AI adaptivity	An AI system's capacity to time and calibrate support, challenge, questioning, and silence.	[11]
Cognitive openness	A user's willingness and capacity to tolerate uncertainty, challenge, and self-revision.	—
Developmental alignment	Aligning AI with users' long-term cognitive development rather than short-term satisfaction or efficiency; developmental adaptivity calibrates support toward this end.	—
Zone of proximal development	The distance between what a learner can do alone and what they can do with appropriate support.	[11]
Flow	A state of deep engagement in which challenge and skill are well matched.	[19]
Extended mind/ distributed cognition	The thesis that cognitive processes can extend beyond brain and body into tools, environments, and practices.	[7,10]
Enactive cognition	Cognition as embodied, situated, and brought forth through action and organism–environment coupling.	[8]

Terms	Definition	Key References
Predictive processing (PP)	An account of cognition as the continuous minimization of prediction error between expected and incoming information.	[29]
Higher-order theory (HOT)	The view that a mental state becomes conscious in virtue of a higher-order representation of it; in entanglement, such self-representation can be externally shaped.	—
Global workspace theory (GWT)	The view that information becomes conscious when it is globally broadcast and made available across cognitive systems.	—
Integrated information theory (IIT)	An account linking consciousness to the amount and structure of integrated information in a system.	[30,31]
Boundaries of the coupled system	The question—raised by integrated-information and extended-mind perspectives—of where the boundary of a human-AI cognitive system should be drawn.	[7,30]

Note: “—” marks a term used by the article without a separate cited source, either one of the framework’s own coinages or a theory named in the context (e.g., global workspace theory, higher-order theory) that the article discusses without a dedicated citation.

References

- Lee, H.-P.; Sarkar, A.; Tankelevitch, L.; Drosos, I.; Rintel, S.; Banks, R.; Wilson, N. The Impact of Generative AI on Critical thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects from a Survey of Knowledge Workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*; ACM: New York, NY, USA, 2025.
- Kosmyrna, N.; Hauptmann, E.; Yuan, Y.T.; Situ, J.; Liao, X.-H.; Beresnitzky, A.V.; Braunstein, I.; Maes, P. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv* **2025**. [CrossRef]
- Stanković, M.; Hirche, E.; Kollatzsch, S.; Doetsch, J.N. Comment on: Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing tasks. *arXiv* **2026**. [CrossRef]
- Vaccaro, M.; Almaatouq, A.; Malone, T.W. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nat. Hum. Behav.* **2024**, *8*, 2293–2303. [CrossRef] [PubMed]
- Turing, A.M. Computing machinery and intelligence. *Mind* **1950**, *59*, 433–460. [CrossRef]
- Searle, J.R. Minds, brains, and programs. *Behav. Brain Sci.* **1980**, *3*, 417–424. [CrossRef]
- Clark, A.; Chalmers, D. The extended mind. *Analysis* **1998**, *58*, 7–19. [CrossRef]
- Varela, F.J.; Thompson, E.; Rosch, E. *The Embodied Mind: Cognitive Science and Human Experience*; MIT Press: Cambridge, UK, 1991.
- Risko, E.F.; Gilbert, S.J. Cognitive offloading. *Trends Cogn. Sci.* **2016**, *20*, 676–688. [CrossRef] [PubMed]
- Hutchins, E. *Cognition in the Wild*; MIT Press: Cambridge, UK, 1995.
- Vygotsky, L.S. *Mind in Society: The Development of Higher Psychological Processes*; Harvard University Press: London, UK, 1978.
- Latour, B. *Reassembling the Social: An Introduction to Actor-Network-Theory*; Oxford University Press: Oxford, UK, 2005.
- Gerlich, M. AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies* **2025**, *15*, 6. [CrossRef]
- Winnicott, D.W. Transitional objects and transitional phenomena. *Int. J. Psycho-Anal.* **1953**, *34*, 89–97.
- Hofstadter, D.R. *Gödel, Escher, Bach: An Eternal Golden Braid*; Basic Books: New York, NY, USA, 1979.
- Wu, X.K.; Chen, M.; Li, W.; Wang, R.; Lu, L.; Liu, J.; Hwang, K.; Hao, Y.; Pan, Y.; Meng, Q.; et al. Llm fine-tuning: Concepts, opportunities, and challenges. *Big Data Cogn. Comput.* **2025**, *9*, 87. [CrossRef]
- Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent abilities of large language models. *Trans. Mach. Learn. Res.* **2022**. Available online: <https://openreview.net/forum?id=yzkSU5zdwD> (accessed on 2 August 2026).
- Schaeffer, R.; Miranda, B.; Koyejo, S. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems 36*; Neural Information Processing Systems Foundation, Inc.: New Orleans, LA, USA, 2023.
- Csikszentmihalyi, M. *Flow: The Psychology of Optimal Experience*; Harper & Row: New York, NY, USA, 1990.
- Bateson, G. *Steps to an Ecology of Mind*; University of Chicago Press: Chicago, IL, USA, 1972.
- Kapur, M. Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educ. Psychol.* **2016**, *51*, 289–299. [CrossRef]

22. Bjork, R.A. Memory and metamemory considerations in the training of human beings. In *Metacognition: Knowing about Knowing*; Metcalfe, J., Shimamura, A.P., Eds.; MIT Press: Cambridge, UK, 1994; pp. 185–205.
23. Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askill, A.; Bowman, S.R.; Cheng, N.; Durmus, E.; Hatfield-Dodds, Z.; Johnston, S.R.; et al. Towards understanding sycophancy in language models. *arXiv* **2023**. [[CrossRef](#)]
24. Perez, E.; Ringer, S.; Lukošiušė, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S.; et al. Discovering language model behaviors with model-written evaluations. *arXiv* **2022**. [[CrossRef](#)]
25. Bostrom, N. Human genetic enhancements: A transhumanist perspective. In *Arguing About Bioeth*; Routledge: London, UK, 2012; pp. 105–114.
26. Russell, S. Human-Compatible Artificial Intelligence. *Hum.-Like Mach. Intell.* **2022**, *1*, 3–22. [[CrossRef](#)]
27. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv* **2023**. [[CrossRef](#)]
28. Christian, B. *The Alignment Problem: Machine Learning and Human Values*; WW Norton & Company: New York, NY, USA, 2020.
29. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **2013**, *36*, 181–204. [[CrossRef](#)] [[PubMed](#)]
30. Tononi, G. An information integration theory of consciousness. *BMC Neurosci.* **2004**, *5*, 42. [[CrossRef](#)] [[PubMed](#)]
31. Tononi, G. Consciousness as integrated information: A provisional manifesto. *Biol. Bull.* **2008**, *215*, 216–242. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.