Ifenthaler, D. (2026). What is hybrid intelligence? Toward a multi-layered model for complex problem-solving. *International Journal of Technology in Teaching and Learning, 22(1),* 1-14.

What is Hybrid Intelligence? Toward a Multi-Layered Model for Complex Problem-Solving

Dirk Ifenthaler
University of Mannheim
Curtin University
University of Eastern Finland

The dominant automation paradigm, which measures artificial intelligence by its capacity to act independently of humans, is increasingly inadequate for complex problem-solving. This article develops the concept of Hybrid Intelligence, defined as an adaptive, self-regulating system arising from the synergistic integration of human and artificial intelligence to enhance cognitive abilities. Revisiting the cognitive-tools tradition through an information-processing lens, it argues that generative AI dissolves the safeguard that once ensured tools amplified rather than supplanted human thinking. It then specifies a layered, multi-agentic architecture in which AI agents and agentic AI are embedded within contextual and collaborative human layers. It applies this architecture to the phases of complex problem-solving. The overriding argument is that sustainable Hybrid Intelligence requires enduring theoretical foundations that outlast individual model releases, ensuring that human–AI collaboration genuinely enhances, rather than inadvertently constrains, human cognitive agency

Keywords: Hybrid intelligence, Problem-solving, Information processing, Artificial intelligence.

INTRODUCTION

Human beings have repeatedly externalised cognitive functions onto artefacts to think further, remember more reliably, and reason about problems that exceed unaided cognitive capacity. This phenomenon is not a sign of cognitive weakness but a defining feature of human intelligence. Humans are distinctive precisely because they can offload, distribute, and scaffold thinking across tools, material, and social environments. Hence, the cognitive psychologist's account of the bounded human mind, which assumes limitations in working memory, susceptibility to bias, and constrained attention (Simon, 1955), has always been accompanied by a complementary account of how external resources expand those bounds (Ifenthaler & Schumacher, 2016). This need for support for cognitive processing in a world of volatility, uncertainty, complexity, and ambiguity has intensified.

Dirk Ifenthaler is the Chair of Learning, Design and Technology at the University of Mannheim, Germany; the UNESCO Co-Chair on Data Science in Higher Education Learning and Teaching at Curtin University, Australia; and a Distinguished Professor in the Department of Applied Educational Science and Teacher Education at the University of Eastern Finland, Finland, and can be reached at dirk@ifenthaler.info.
DOI: https://doi.org/10.37120/ijttl.2026.22.1.01

Contemporary problems in a digitally saturated world are characterised by vast volumes of data, rapid rates of change, and high degrees of interdependence that overwhelm individual and even collective human capacity. When problems are dynamic, ill-structured, and distributed across many interacting stakeholders (Jonassen, 1997), systems for cognitive expansion are no longer passive external memories or calculating devices; they are data analytics, machine learning algorithms, and artificial intelligence (AI). These systems can process, structure, and act on data at scales and speeds beyond human reach (LeCun et al., 2015), detecting statistical regularities in noisy data and generating predictions infeasible for unaided cognition. Their significance lies less in raw computational power than in their potential to reconfigure how humans perceive, reason about, and respond to complexity (Dörner, 1987).

The essential inquiry, therefore, is not *whether* cognition should be supported by technology, but rather *how* these technologies should be integrated into human thought processes to enhance human agency rather than displace it. Much of the current discourse rests on what may be termed the *automation* paradigm, in which an AI system's success is measured by its ability to operate independently of human involvement and perform tasks traditionally done by humans (Richardson, 2026). An alternative paradigm, known as *augmentation* (Meier et al., 2021), considers the combined human–machine system rather than the autonomous machine itself, exploring how such systems can be designed so that the whole surpasses the sum of its parts (Daugherty & Wilson, 2018; De Laat et al., 2020). This contrast has gained renewed urgency with the advent of large-scale generative and agentic AI (Kramer, 2026). The augmentation case is based not only on historical precedent but also on the principle of *complementarity* (Dellermann et al., 2019).

Thus, the concept of *Hybrid Intelligence* emerges, defined as an adaptive system resulting from the synergistic integration of human and artificial intelligence to enhance cognitive capabilities. Clearly, the human role does not disappear; it becomes more strategic, repositioned at the level where context, value, and judgment are decisive.

This article begins with a critical re-examination of the cognitive-tools tradition, then defines Hybrid Intelligence and specifies a layered, multi-agentic architecture in which AI agents and agentic AI are embedded within a collaborative human layer. It then applies this architecture to the demands of complex problem-solving and considers the implications for research, practice, and preserving human cognitive agency.

COGNITIVE TOOLS REVISITED

The augmentation paradigm derives its theoretical depth from a long-standing tradition in the learning sciences on the relationship between cognition and the technologies that support it (Clark, 1994; Kozma, 1991). A foundational pillar of this field is the information-processing metaphor, which conceptualises the human mind as a system that perceives, processes, stores, and retrieves information through a series of cognitive operations such as attention, perception, memory encoding, storage, and retrieval (Baddeley, 1986; Baddeley et al., 2001). Further, it is assumed that external objects of the world are internalised as mental representations and, in turn, re-represented back into the world (Ifenthaler & Seel, 2013). Hence, knowledge representation can be understood as a cyclical exchange among an object zone, a knowledge zone, and an intermediary representation zone, governed by an internalising function (*fin*) and an externalising function (*fout*) (Ifenthaler, 2010; Seel, 1991). The relevance of this tradition to the design of cognitive technologies is direct (Shavelson, 1972): if learning is understood as the internalisation of external information and its subsequent re-representation, then the central question for any educational technology becomes which of these processing operations the technology assumes, and which it leaves for the learner to perform (Jonassen et al., 1993). Along this axis, the

Mindtools framework drew its fundamental distinction (Jonassen, 1999; Jonassen & Cho, 2008).

On the one hand, technologies may be used to instruct, i.e., to convey information, pose queries, judge responses, and deliver feedback, with relevant information effectively stored in the technology and the learner positioned as its recipient (Jonassen, 1995). In information-processing terms, instructional technology assumes responsibility for the externalisation function, i.e., it pre-packages representations and transmits them, leaving the learner to perceive and understand the information stored in the technology as they interact with it (Jonassen, 1999). However, such systems may remove meaningful control of the learning process from learners and teachers alike. The learner's cognitive contribution is confined to encoding and recall, the very operations the information-processing framework regards as the least generative, because they engage neither the structural analysis nor the re-representation through which sustainable understanding is built. On the other hand, technologies may serve as cognitive tools that learners learn with, not from. Such systems may engage them in critical thinking about the information they are studying. Here, the load of externalisation is deliberately returned to the learner, with the tool supplying a formalism. However, the learner must perform the fout operation of articulating, structuring, and visualising what they know.

The theoretical foundation for this stance is based on the distinction between the effects of technology and the effects with technology. Learning with technology involves the learner engaging in an intellectual partnership with the system, wherein there exists the potential to qualitatively enhance the performance of the joint system comprising the learner, the teacher, peers, and the technology. Importantly, this enhancement is not achieved merely by reducing the cognitive load associated with information processing (Kirschner et al., 2011). Mindtools are not necessarily intended to simplify learning, rather, they often require learners to engage in more complex thinking about the subject matter being studied, generating thoughts that would be impossible without the tool (Jonassen, 1999; Jonassen & Cho, 2008). The pedagogical importance lies not in alleviating the demands on memory encoding or representation but in intensifying them, thereby stimulating deeper re-representation and leading to the reorganisation of knowledge structures (Shavelson, 1972).

At the core of the framework is a deliberate assignment of cognitive workload, which makes this argument both compelling and open to revision. Mindtools are described as unintelligent systems that rely on learners to supply intelligence rather than on the computer (Jonassen, 1999). The pedagogical principle assigns cognitive responsibility to learners for tasks they handle best, while technology manages the operations it performs most efficiently. This division aligns well with the information-processing model: learners focus on higher-order skills such as recognising, judging patterns, and organising information, while systems execute more quickly and accurately. This division is static and, importantly, safe because the machine's limitations define the boundary. For instance, a spreadsheet can calculate but not interpret, and a semantic networking tool can generate a concept map but cannot analyse the structural relationships that underlie it. Learners are protected from cognitive abdication because the system cannot perform the learner's generative functions.

This division of labour clarifies how such systems transform the learner from a reproducer of information into a knowledge builder (Jonassen, 2009). Generative AI dissolves this safeguard, and the framework must be critically revisited rather than simply reapplied. Recent work argues that AI no longer merely stores and retrieves but actively participates in both the internalising and externalising functions, inserting an interaction layer that intersects the object, knowledge, and representation zones alike (Ifenthaler, 2025). Large language models internalise vast corpora into sub-symbolic representations and externalise them as fluent, human-like re-representations, thereby performing the very *fout*

operations the Mindtools tradition reserved for the learner (Yan et al., 2025). Hence, the pedagogical value of a Mindtool can now be delegated to the tool itself. The risk, therefore, is not that learners will fail to think with such systems, but that they will once again learn from them, reverting to the transmission model the framework was designed to displace, now in a more seductive and fluent form, with a corresponding risk that over-reliance on technology may undermine learners' autonomy (Ifenthaler & Schumacher, 2023).

Two implications follow. First, the foundational criterion endures, as a system enhances cognition only insofar as it prompts humans to process information more deeply, not less. Second, the machine's incapacity can no longer guarantee the allocation of cognitive responsibility, because that incapacity has largely vanished. It must instead be secured by design. This reconfigured, design-dependent division of cognitive labour provides the theoretical foundation for Hybrid Intelligence and motivates the development of a pedagogically meaningful multi-agentic architecture.

HYBRID INTELLIGENCE

What is the concept of Hybrid Intelligence? Hybrid Intelligence (HI) is an adaptive system that arises from the synergistic integration of human and artificial intelligence, enhancing cognitive abilities. Each element of this definition is deliberate and warrants elaboration, because it is the conjunction of these elements, rather than any one of them alone, that distinguishes HI from the more familiar notions with which it is frequently conflated, such as human-in-the-loop (Joksimovic et al., 2023; Siemens et al., 2022) or human + machine interaction (Daugherty & Wilson, 2018).

The term *system* indicates that the appropriate unit of analysis is neither the human nor the machine in isolation, but rather their combined configuration (Daugherty & Wilson, 2018). Accordingly, Engelbart (1962) argued that human performance is best improved by treating the whole arrangement of language, artefacts, methodology, and training as a single interacting system rather than by isolating any one component. Siemens et al. (2022) reach a parallel conclusion from a contemporary standpoint, treating human cognition and artificial cognition as two distinct systems, each possessing distinct strengths and capabilities. Thus, intelligence is not a property residing in either component but a phenomenon emerging from their coordination. Dellermann et al. (2019) emphasise this argument explicitly when they characterise hybrid intelligence systems as socio-technical ensembles of human and artificial agents that co-evolve and improve over time.

The term *synergistic* expresses the objective that the combined system should accomplish what neither component could accomplish alone. Engelbart (1962) explicitly preferred the language of synergism over that of amplification, locating augmentation in the way a system's many sub-processes reinforce one another. Hence, the goal of hybrid intelligence is to attain results superior to those any participating actor could have achieved independently, by combining the complementary strengths of heterogeneous forms of intelligence (Dellermann et al., 2019). Synergy therefore presupposes complementarity, i.e., human and artificial intelligence exhibit systematically different profiles of strength (Daugherty & Wilson, 2018). A broader perspective further distinguishes domains of cognitive activity (Siemens et al., 2022): sensory processes, general operations, and complex integrated activities such as reasoning, judgement, and sensemaking (Baddeley et al., 2001; Comes, 2026). While AI presently excels at the first two, particularly at detecting patterns across large datasets and generating predictions in near real time (Lodge & Corrin, 2017), humans retain decisive advantages in the complex, integrated activities that demand contextual judgement, the grounding of concepts in physical and social experience, and the determination of why a solution is needed (Siemens et al., 2022). Accordingly, the non-overlapping nature of these profiles enables synergy.

The term *adaptive* distinguishes Hybrid Intelligence from static configurations in which roles are fixed at the outset. A key requirement is continuous learning through which the socio-technical system improves both as a whole and through the mutual learning of its human and machine components (Dellermann et al., 2019). Understanding continuous learning as a developmental trajectory, three phases can be identified (Siemens et al., 2022): First, humans train the artificial system in a unidirectional relationship. Second, the trained system assists humans with data processing, analysis, and decision-making in a bi-directional relationship in which the human requests and the machine responds. Third, humans and machines engage in genuine collaboration, working as a team in which each is aware of the other's state, and outputs are passed back and forth for ongoing processing (Daugherty & Wilson, 2018). However, the emphasis on true collaboration depends on a form of shared cognition (Cannon-Bowers & Salas, 2001), in which both partners have insight into one another's knowledge and a common understanding of the task (Ifenthaler, 2014). A Hybrid Intelligence system is therefore one capable of operating in, and moving between, these modes as the demands of a problem evolve.

Finally, to *enhance cognitive abilities* identifies the system's purpose and supplies its normative criterion. Hybrid Intelligence is not pursued for efficiency alone. Dellermann et al. (2019) argue that the value of such systems extends beyond the mere efficiency of task completion to ensembles that also improve their human members, so that performance is measured not only by the superior output of the whole but also by the learning of the human. Accordingly, any approach to Hybrid Intelligence must extend beyond the efficiency and productivity of outputs to encompass considerations of bias, ethics, suitability, and long-term societal impact (Li et al., 2023; Yeaman et al., 1994). Particular caution is needed for biases embedded in artificial systems, as they can carry over into human ones (Baker & Hawn, 2021).

Accordingly, Hybrid Intelligence differs from human augmentation in the narrow technological sense of extending sensory or memory capacity through devices (Gonzalez, 2026). Hybrid Intelligence differs from collective intelligence, which aggregates the contributions of many human agents. It should not be conflated with artificial general intelligence, which keeps the human out of the loop (Dellermann et al., 2019). Finally, if AI systems now participate in both the internalising (*fin*) and externalising (*fout*) functions of knowledge representation, inserting an interaction layer that intersects the object, knowledge, and representation zones (Ifenthaler, 2025), then Hybrid Intelligence is the deliberate organisation of that shared participation. Hybrid Intelligence is an arrangement in which the machine's capacity to internalise vast corpora and externalise fluent re-representations is coupled to the human's capacity to frame problems, interpret meaning in context, and exercise accountable judgement.

TOWARD A MULTI-AGENTIC ARCHITECTURE OF HYBRID INTELLIGENCE

AI AGENTS AND AGENTIC AI

A terminological clarification is required regarding AI agents and agentic AI. Some educational literature treats them as interchangeable, using both to denote automated decision-making systems (Lodge & Loble, 2026). An *AI agent*, is a single-entity system that performs goal-directed tasks by utilising external tools, applying sequential reasoning, and integrating real-time information to complete a well-defined function (Sapkota et al., 2026). Unlike a classical generative model confined to producing text or images from training data, an AI agent can access external information through interfaces and operate tools, extending its reach into the world. Its scope of action nonetheless remains relatively bounded, and its degree of agency is limited (Hahn et al., 2026). *Agentic AI*, by contrast, denotes a multi-

agent configuration in which several specialised agents coordinate, communicate, and dynamically allocate sub-tasks within a broader workflow to pursue a common goal (Sapkota et al., 2026). Agentic AI represents a qualitative leap in artificial intelligence, defined by its capacity to set complex goals in changing and uncontrolled situations and pursue them by autonomously managing resources (Acharya et al., 2025).

Accordingly, three properties distinguish agentic AI from a single AI agent (Ashwani, 2025; Hahn et al., 2026; Nama, 2026): (a) Increased complexity, as the system is no longer confined to a single task. (b) Enhanced adaptability, as a network of agents adjusts to novel contexts more effectively than any single AI agent could. (c) Heightened agency, as the user delegates a goal to the system, which independently selects the appropriate tools and executes actions with minimal to no human intervention.

LIMITATIONS OF AI AGENTS AND AGENTIC AI

Clearly, contemporary machine-learning systems, such as AI agents and agentic AI, struggle to adapt to dynamic environments, sense contextual cues, self-adjust, lack what humans recognise as common sense, remain vulnerable to adversarial manipulation, and exhibit limited interpretability (Dellermann et al., 2019). Further, these systems remain largely confined to sensory processes and general operations, and they caution that biases embedded in artificial systems are likely to propagate into the human systems with which they interact (Siemens et al., 2022). The generative AI literature compounds these concerns, noting these systems' propensity for fluent yet unreliable output and confident fabrication (Lodge & Loble, 2026), as well as their propensity for fluent error and confident hallucination (Yan et al., 2025).

Yet, agentic AI configurations do not resolve these limitations, rather, they amplify them. Three properties that define agentic AI also intensify its risks (Hahn et al., 2026). First, the broad assistance that makes agentic AI useful also increases its capacity to cause harm, whether through unintended bias, user over-reliance, or deliberate malicious infiltration. Second, the complexity and interconnection of multi-agent systems erode transparency. Hence, when a task is decomposed across several specialised agents operating in sequence, it becomes difficult to trace which agent contributed what, where particular information originated, and, in the event of error, where the fault lies. Third, the heightened autonomy of agentic AI diffuses responsibility. While these systems are designed to operate with minimal human intervention, accountability can become indeterminate, creating responsibility gaps in the event of harm (Hahn et al., 2026).

As a consequence, these limitations threaten human oversight, raising with fresh urgency the question of whether a human remains in, on, or out of the loop (Nama, 2026). Where a single agent's error is contained, an orchestrated suite can propagate and compound errors across its constituent agents, and the autonomy that characterises agentic AI removes the very junctures where human judgement might otherwise intervene. This concern is reinforced by evidence that human–AI interaction can drift over repeated exchanges, with feedback loops amplifying biases rather than correcting them (Glickman & Sharot, 2025).

Thus, genuine human–machine collaboration has not yet been realised, and the principles for integrating machine outputs into human knowledge systems remain underspecified. Agentic AI scales machine capability without providing the contextual reasoning, accountability, or value-grounding that complex problem-solving demands.

A LAYERED MODEL OF HYBRID INTELLIGENCE

Hybrid Intelligence transcends the isolated limitations of AI agents and agentic AI by embedding them within a collaborative human layer, creating an adaptive, self-regulating system in which human contextual reasoning and machine computational power mutually

reinforce each other. Figure 1 shows the conceptual model. The layered model of HI comprises the following elements:

- The World (W). Consistent with the information-processing framing, this represents the object zone, i.e., the physical and social reality about which problems are posed and within which solutions must act.
- Contextual Spheres (CSW1…CSWx). Between the world and the collaborating system lie the contextual spheres that situate any problem, including the domain, stakeholders, institutional setting, values, etc. This layer formalises why a solution is needed as well as the irreducibly contextual character of complex problems.
- The Collaboration layer (C). This is the locus of shared cognition, a precondition of genuine collaboration. The interface at which human and machine partners exchange representations, calibrate trust, and negotiate the allocation of cognitive responsibility.
- The Hybrid Intelligence layer (HI1…HIx). Within the collaboration layer sits Hybrid Intelligence proper: the adaptive coupling in which the complementary strengths of human and artificial intelligence are combined to achieve results that neither could attain in isolation.
- The Multi-Agent Suites (MAS1…MASx) and Agents (A1…Ax). The agentic configurations and their constituent agents, including the machine computational power and relevant data that the human layer directs, interprets, and holds to account.

Figure 1. *A layered model of Hybrid Intelligence*

The defining feature of the layered model is that the machine layers are enclosed by, rather than merely connected to, the human and contextual layers. This containment directly addresses the aforementioned limitations, such as opacity, diffusion of responsibility, and erosion of oversight that accompany agentic AI, and is countered not by technical means alone but by embedding the machine layers within a human layer that provides framing, judgement, and accountability. AI agents provide scale, speed, and pattern detection, while the collaboration and contextual layers provide the reasoning that AI agents and agentic AI demonstrably

lack. The architecture is adaptive because responsibility can shift across layers as a problem evolves. Accordingly, the collaboration layer continuously monitors and recalibrates that allocation, similar to the concept of mental models (Ifenthaler & Seel, 2013; Johnson-Laird, 1996). In this way, the architecture operationalises the normative criterion, i.e., it is designed so that human–AI collaboration enhances, rather than supplants, human cognitive agency.

Equation (1) formalises Hybrid Intelligence (*HI*) as the output of a collaboration-layer function f_C that synergistically integrates (⊗) two contributions within a given world context *W*. The first, *H(CSW)*, is the human contribution operating through the contextual spheres, which includes framing, valuation, and accountable judgement. The second is the summed output of the multi-agent suites, each *MAS_i* orchestrating its constituent agents ($A_1 \ldots A_x$) for machine computation. The operator ⊗ signals that integration is non-additive, capturing the claim that the whole exceeds the sum of its parts (Dellermann et al., 2019); f_C denotes the system's adaptive self-regulation.

$$HI = f_C\left(H(CSW) \otimes \sum_{i=1}^{n} MAS_i(A_1, \ldots, A_x) \,\middle|\, W \right)$$

Equation (1). *A functional representation of the layered Hybrid Intelligence model.*

A two-layer illustration clarifies the architecture (see Figure 1). Consider adaptive mathematics tutoring (Ashwani, 2025). At the agent layer, individual agents each perform a discrete function: one diagnoses the learner's errors, another retrieves worked examples, and a third generates practice items, orchestrated as a multi-agent suite. At the enclosing human layer, the teacher operating through the contextual spheres frames the pedagogical goal, judges whether the agent's proposed learning path suits the learner, and retains accountability for the intervention. The machine layer supplies scale and adaptivity, while the human layer supplies framing, valuation, and ethical oversight. Neither, in isolation, constitutes Hybrid Intelligence.

HYBRID INTELLIGENCE PROBLEM SOLVING

THE NATURE OF COMPLEX PROBLEM SOLVING

Problem solving can be understood as constructing and applying mental representations of a problem to move from a given state to a desired goal state (Seel et al., 2009). Two features of this account are consequential for Hybrid Intelligence: First, problem solving is fundamentally *representational*, i.e., its quality depends less on the mechanical execution of a solution procedure than on the problem solver's active, mindful construction of the problem space, understood as the internal model of the situation, its entities, and their structural relationships (Dörner & Wearing, 1995). Second, problems differ in kind. Well-structured problems present known goal states and constrained solution paths, whereas ill-structured problems present unknown, contested, or negotiable goals and constraints (Jonassen, 1997). The features that render a problem complex have been identified as (a) the sheer number of variables in a system, (b) the connectivity among them, which

multiplies the consequences of any single intervention, (c) the dynamics of the situation, including its capacity to evolve independently of the problem-solver's actions, (d) the intransparency of variables and goals, which obliges the solver to acquire information actively, and (e) polytely, the presence of multiple and often antagonistic goals that must be reconciled through compromise and prioritisation (Funke, 2012). Accordingly, a system confronting a dynamic, intransparent, polytelic problem cannot simply compute an optimal solution, because the goal state itself is underspecified and value-laden.

THE NEED FOR COGNITIVE SYSTEMS IN COMPLEX PROBLEM SOLVING

Complex problem solving is no longer plausibly theorised as an exclusively individual, intracranial activity. Hence, scientific inquiry into problem solving has turned from the cognitive processes of isolated individuals toward cognitive systems composed of people, machines, and networked information working together (Gibson et al., 2022). On this view, humans remain the source of questions, the motivation for inquiry, and the ultimate arbiters of the value of any solution, while machines contribute the capacity to process large volumes of data at high velocity, to detect patterns that would otherwise go unnoticed, and to build and test computational models on the fly. However, human-centred cognitive computing depends on keeping humans continuously in the loop with machines and networked information (Lee et al., 2025).

In the layered model, the machine layers address computationally tractable features of complexity, such as the number of variables, the modelling of connectivity, and the tracking of dynamic change. The human layer addresses features that are not tractable, such as framing intransparent goals, reconciling polytelic value conflicts, and accountable judgement of solution quality. HI problem solving is therefore the deliberate organisation of a cognitive system in which these contributions are coupled rather than merely juxtaposed.

MAPPING THE LAYERED MODEL ONTO PROBLEM SOLVING

The problem-solving process is commonly synthesised into three broad phases: constructing a representation of the problem, searching for solutions, and implementing and monitoring those solutions (Jonassen, 2010).

- In the *representation* phase, the contextual spheres (CSW) and the human layer are decisive. Because the deep structure of a complex problem is visible only to those with the requisite domain knowledge, the framing of the problem, i.e., the determination of what it *is* and *why* a solution is needed, cannot be delegated to the agentic AI layers. Here the machine layers serve a supporting role, gathering and structuring information to inform, but not to replace, the human construction of the problem space.
- In the *search* phase, the multi-agent suites (MAS) come to the fore. Detecting patterns across large datasets, generating possible solutions, and simulating dynamic system behaviour are activities at which agentic AI excels and that map onto the sensory and general-operational domains of machine cognition. Yet, error propagation, opacity, and the amplification of bias are acute precisely here, because the search space of a complex problem is vast and the machine's proposals are difficult to audit. The collaboration layer (C) must therefore maintain continuous oversight, calibrating trust in machine-generated options rather than accepting them uncritically.
- In the *implementation and monitoring* phase, the dynamic and intransparent character of complex problems makes self-regulation indispensable. Because such problems evolve over time and interventions can produce unintended consequences, the allocation of responsibility between the human and machine layers cannot be fixed in advance. It must be renegotiated as the situation unfolds. This is the sense in which

HI problem-solving is genuinely adaptive, i.e., the system monitors not only the problem but also its own division of cognitive labour.

This mapping shows that the layered model is not an abstract architecture but a functional response to the specific demands of complex problem-solving. It places the human contribution where the features of complexity most resist computation, including framing, valuation, and accountable judgement. Thus, it harnesses agentic AI where scale and speed are decisive. In doing so, it operationalises the principle that human–AI collaboration should enhance, rather than erode, the human cognitive agency on which solving consequential problems ultimately depends.

IMPLICATIONS AND A RESEARCH AGENDA

Sustainable Hybrid Intelligence requires enduring theoretical foundations that outlast individual model releases, ensuring that human–AI collaboration genuinely enhances rather than inadvertently constrains human cognitive agency.

The first implication is that theory, not technical capability, should anchor the field. The dominant empirical literature is preoccupied with the performance of specific systems, an orientation criticised as an effect in search of a cause, in which the medium is repeatedly conflated with the pedagogical method (Weidlich et al., 2025). HI offers foundations deliberately independent of any particular model generation. This durability matters because the central normative criterion that a system must enhance, not supplant, the human partner's capacity to reason and self-regulate, cannot be evaluated against benchmarks that measure short-term task performance alone.

A second implication is preserving cognitive agency. A growing body of evidence documents a performance paradox in which AI improves immediate output while eroding durable learning, along with a corresponding tendency toward metacognitive laziness and harmful cognitive offloading (Lodge & Loble, 2026). These findings sharpen the tension that the very fluency that makes generative systems useful can create an illusion of competence that discourages the effortful processing on which understanding depends. For HI, this establishes a design imperative. Because the machine layers can now perform the externalising functions once reserved for humans, the division of cognitive labour must be engineered to keep humans in a genuinely generative role rather than a merely confirmatory one. The equity dimension is acute here, since humans least equipped with prior knowledge and self-regulatory skills are most vulnerable to harmful offloading (Dang et al., 2026; Lodge & Loble, 2026).

A third implication concerns governance. The autonomy that characterises agentic AI diffuses responsibility and erodes transparency (Hahn et al., 2026). The shift from evaluating outputs to delegating authority repositions the human as a principal rather than a mere assessor (Nama, 2026). Sustainable HI therefore requires that the collaboration layer operationalise meaningful human control and accountability by design, not as an afterthought. This is a socio-technical rather than a purely technical requirement, encompassing the value-sensitive determination of which tasks may appropriately be delegated.

From a critical perspective, the field currently lacks validated instruments to assess whether a hybrid configuration enhances cognitive agency, and process-oriented measures of retention and transfer are needed in place of performance-only outcomes. Furthermore, the integration problem remains largely unresolved, particularly regarding the principles by which outputs from one cognitive system are meaningfully re-integrated into the other. Longitudinal, rigorous research is clearly required to trace the cumulative effects of delegation on human capability, given that the erosion of agency is likely to proceed gradually. Lastly, the architecture itself necessitates empirical validation across the

problem-solving phases, including its behaviour under dynamic, opaque, and polytelic conditions that characterise genuinely complex problems. Pursuing this research agenda is a prerequisite for human intelligence that is not only powerful but also sustainable.

CONCLUSION

Hybrid Intelligence represents an adaptive, self-regulating system that emerges from the synergistic integration of human and artificial intelligence to augment cognitive capabilities. The layered model operationalises this concept by embedding AI agents and agentic AI within contextual and collaborative human layers, ensuring that machine computational power and human contextual reasoning mutually reinforce rather than simply coexist. When applied to complex problem-solving, the architecture places the human contribution in aspects where complexity most impedes computation, namely framing, valuation, and accountable judgment, while leveraging agentic scale in areas where it proves decisive. Accordingly, sustainable Hybrid Intelligence requires enduring theoretical foundations that extend beyond individual model releases, ensuring that human–AI collaboration genuinely enhances, rather than unintentionally restricts, human cognitive agency. Achieving this vision demands not only advanced technical capabilities but also principled, design-centred stewardship of the division of cognitive labour.

REFERENCES

Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey. *IEEE Access*, *13*, 18912–18936. https://doi.org/10.1109/ACCESS.2025.3532853

Ashwani, S. (2025). Agentic AI in education: Transforming pedagogy and personalized learning. In C. K. K. Reddy, S. Joseph, H. Joshi, S. Doss, & M. Ouaissa (Eds.), *The power of Agentic AI: Redefining human life and decision-making. Sustainable artificial intelligence-powered applications* (pp. 321–332). Springer. https://doi.org/10.1007/978-3-031-89424-4_21

Baddeley, A. D. (1986). *Working memory*. Clarendon Press.

Baddeley, A. D., Chincotta, D., & Adlam, A. (2001). Working memory and the control of action: evidence from task switching. *Journal of Experimental Psychology General*, *130*, 641–657.

Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Cannon-Bowers, J. A., & Salas, E. (2001). Reflections on shared cognition. *Journal of Organizational Behavior*, *22*(2), 195–202.

Clark, R. E. (1994). Media will never influence learning. *Educational Technology Research and Development*, *42*(2), 21–29.

Comes, T. (2026). Sensemaking AI: Introducing a research and design agenda for human–AI networks. *EPJ Data Science*, *15*, 42. https://doi.org/10.1140/epjds/s13688-026-00634-5

Dang, A. H., Huynh, V. S., Nguyen, D. D., & Baek, Y. K. (2026). Beyond automation: Cultivating self-regulated learning through AI and reflective practice. *Technology, Knowledge and Learning*. https://doi.org/10.1007/s10758-026-09970-x

Daugherty, P. R., & Wilson, H. J. (2018). *Human + machine: Reimagining work in the age of AI*. Harvard Business Review Press.

De Laat, M., Joksimovic, S., & Ifenthaler, D. (2020). Artificial intelligence, real-time feedback and workplace learning analytics to support in situ complex problem-solving: a commentary. *International Journal of Information and Learning Technology*, *37*(5), 267–277. https://doi.org/10.1108/IJILT-03-2020-0026

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, *61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

Dörner, D. (1987). On the difficulties people have in dealing with complexity. In J. Rasmussen, K. Duncker, & J. Leplat (Eds.), *New technology and human error* (pp. 97-109). John Wiley & Sons.

Dörner, D., & Wearing, A. (1995). Complex problem solving: Toward a (computersimulated) theory. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 65-99). Lawrence Erlbaum.

Engelbart, D. C. (1962). *Augmenting human intellect: a conceptual framework*. Defense Technical Information Center. https://doi.org/10.21236/AD0289565

Funke, J. (2012). Complex problem solving. In N. M. Seel (Ed.), *The encyclopedia of the sciences of learning* (Vol. 3, pp. 682–685). Springer. https://doi.org/10.1007/978-1-4419-1428-6_685

Gibson, D. C., Ifenthaler, D., & Greiff, S. (2022). Teaching of general psychology: Problem solving. In J. Zumbach, D. Bernstein, S. Narciss, & G. Marsico (Eds.), *International handbook of psychology learning and teaching* (pp. 1–19). Springer. https://doi.org/10.1007/978-3-030-26248-8_8-1

Glickman, M., & Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, *9*, 345–359. https://doi.org/10.1038/s41562-024-02077-2

Gonzalez, C. (2026). Reply to 'Human–AI complementarity needs augmentation, not emulation'. *Nature Reviews Psychology*, *5*, 230–231. https://doi.org/10.1038/s44159-026-00537-2

Hahn, M., Tretter, M., & Dabrock, P. (2026). Ethical perspectives on AI Agents and Agentic AI. *AI and Ethics*, *6*, 218. https://doi.org/10.1007/s43681-026-01027-0

Ifenthaler, D. (2010). Scope of graphical indices in educational diagnostics. In D. Ifenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 213–234). Springer. https://doi.org/10.1007/978-1-4419-5662-0_12

Ifenthaler, D. (2014). Toward automated computer-based visualization and assessment of team-based performance. *Journal of Educational Psychology*, *106*(3), 651-665. https://doi.org/10.1037/a0035505

Ifenthaler, D. (2025). Information processing re-visited: How artificial intelligence supports human-centered systems. *AI-Enhanced Learning*, *1*(1), 229–243. https://doi.org/10.70725/465418kgasdx

Ifenthaler, D., & Schumacher, C. (2016). Connectivism. In S. Danver (Ed.), *The SAGE encyclopedia of online education* (pp. 409–411). Sage.

Ifenthaler, D., & Schumacher, C. (2023). Reciprocal issues of artificial and human intelligence in education. *Journal of Research on Technology in Education*, *55*(1), 1–6. https://doi.org/10.1080/15391523.2022.2154511

Ifenthaler, D., & Seel, N. M. (2013). Model-based reasoning. *Computers & Education*, *64*, 131–142. https://doi.org/10.1016/j.compedu.2012.11.014

Johnson-Laird, P. N. (1996). Mental models, deductive reasoning, and the brain. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 999–1008). MIT Press.

Joksimovic, S., Ifenthaler, D., De Laat, M., Siemens, G., & Marronne, R. (2023). Opportunities of artificial intelligence for supporting complex problem-solving: findings from a scoping review *Computers & Education: Artificial Intelligence*, *4*, 100138. https://doi.org/10.1016/j.caeai.2023.100138

Jonassen, D. H. (1995). Computers as cognitive tools: Learning with technology, not from technology. *Journal of Computing in Higher Education*, *6*(2), 40–73.

https://doi.org/10.1007/BF02941038
Jonassen, D. H. (1997). Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology Research and Development*, *45*(1), 65-94. https://doi.org/10.1007/BF02299613
Jonassen, D. H. (1999). *Computers as mindtools for schools: Engaging critical thinking* (2 ed.). Prentice Hall.
Jonassen, D. H. (2009). Externally modeling mental models. In L. Moller, J. B. Huett, & D. Harvey (Eds.), *Learning and instructional technologies for the 21st century. Visions of the future* (pp. 49-74). Springer.
Jonassen, D. H. (2010). *Learning to solve problems. A handbook for designing problem-solving learning environments*. Routledge. https://doi.org/10.4324/9780203847527
Jonassen, D. H., Beissner, K., & Yacci, M. (1993). *Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge*. Lawrence Erlbaum.
Jonassen, D. H., & Cho, Y. H. (2008). Externalizing mental models with mindtools. In D. Ifenthaler, P. Pirnay-Dummer, & J. M. Spector (Eds.), *Understanding models for learning and instruction. Essays in honor of Norbert M. Seel* (pp. 145-160). Springer.
Kirschner, F., Kester, L., & Corbalan, G. (2011). Cognitive load theory and multimedia learning, task characteristics and learning engagement: The Current State of the Art. *Computers in Human Behavior*, *27*(1), 1–4. https://doi.org/https://doi.org/10.1016/j.chb.2010.05.003
Kozma, R. B. (1991). Learning with media. *Review of Educational Research*, *61*(2), 179-211.
Kramer, O. (2026). *Artificial intelligence essentials*. Springer. https://doi.org/10.1007/978-3-032-06637-4
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539
Lee, S. M., Li, H., Zhang, S., Zhong, Z., Lee, J. E., & Botelho, A. F. (2025). So what? Unpacking the complexities in collaborative problem solving with AI-augmented sense-making. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial Intelligence in Education. AIED 2025* (Vol. 15881, pp. 431–438). Springer. https://doi.org/10.1007/978-3-031-98462-4_54
Li, F., Ruijs, R., & Lu, Y. (2023). Ethics & AI: a systematic review on ethical concerns and related strategies for designing with ai in healthcare. *AI*, *4*(1), 28–53. https://doi.org/10.3390/ai4010003
Lodge, J. M., & Corrin, L. (2017). What data and analytics can and do say about effective learning. *npj Science of Learning*, *2*(1), 5. https://doi.org/10.1038/s41539-017-0006-5
Lodge, J. M., & Loble, L. (2026). *Artificial intelligence, cognitive offloading and implications for education*. University of Technology Sydney. https://doi.org/10.71741/4pyxmbnjaq.31302475
Meier, C., Seufert, S., Guggemos, J., & Spirgi, J. (2021). Learning organizations in the age of smart machines. Fusion skills, augmentation strategies, and the role of HRD professionals. In D. Ifenthaler, S. Hofhues, M. Egloffstein, & C. Helbig (Eds.), *Digital transformation of learning organizations* (pp. 77–94). Springer.
Nama, R. (2026). From evaluator to principal: the agentic AI literacy framework (AALF) for delegated autonomy. *AI and Ethics*, *6*, 299. https://doi.org/10.1007/s43681-026-01167-3
Richardson, C. M. (2026). Human-Centric automation to intelligent information integration: A mixed-methods framework for industry 5.0 manufacturing. *Journal of Industrial Information Integration*, *51*, 101094. https://doi.org/10.1016/j.jii.2026.101094
Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2026). AI Agents vs. Agentic AI: A

Conceptual taxonomy, applications and challenges. *Information Fusion*, *126*, 103599. https://doi.org/10.1016/j.inffus.2025.103599

Seel, N. M. (1991). *Weltwissen und mentale Modelle*. Hogrefe.

Seel, N. M., Ifenthaler, D., & Pirnay-Dummer, P. (2009). Mental models and problem solving: Technological solutions for measurement and assessment of the development of expertise. In P. Blumschein, W. Hung, D. H. Jonassen, & J. Strobel (Eds.), *Model-based approaches to learning: Using systems models and simulations to improve understanding and problem solving in complex domains* (pp. 17-40). Sense Publishers.

Shavelson, R. J. (1972). Some aspects of the correspondence between content structure and cognitive structure in Physics education. *Journal of Educational Psychology*, *63*(3), 225–234. https://doi.org/10.1037/h0032652

Siemens, G., Marmolejo-Ramos, F., Gabriel, F., Medeiros, K., Marrone, R., Joksimovic, S., & de Laat, M. (2022). Human and artificial cognition. *Computers and Education: Artificial Intelligence*, *3*, 100107. https://doi.org/10.1016/j.caeai.2022.100107

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. https://doi.org/10.2307/1884852

Weidlich, J., Gašević, D., Drachsler, H., & Kirschner, P. A. (2025). ChatGPT in education: An effect in search of a cause. *Journal of Computer Assisted Learning*, *41*(5), e70105. https://doi.org/10.1111/jcal.70105

Yan, W., Nakajima, T., & Sawada, R. (2025). Beyond tool use: Tracking the evolution of generative AI literacy among university students through a process-oriented investigation. *Computers and Education: Artificial Intelligence*, *9*, 100465. https://doi.org//10.1016/j.caeai.2025.100465

Yeaman, A. R. J., Koetting, J. R., & Nichols, R. G. (1994). Critical theory, cultural analysis, and the ethics of educational technology as social responsibility. *Educational Technology*, *34*(2), 5–13.