

A sketch: How do individual and collective cognition emerge when using AI; and in which ways create AI disruptions and expansions of social and cognitive/knowledge processes in multiple types of settings.

Statement

What is unique about humans and our societies is our capacity to construct skyscrapers, develop advanced socio-technical infrastructures, communicate in various modes, and survive in almost impossible conditions across diverse ecological niches on the planet.

As stated by Henrich and Muthukrishna, “Contrary to conventional wisdom, human brilliance emerges not from our innate brainpower or raw computational capacities, but from sharing information in communities and networks over generations” (Topics in Cognitive Science, 16, 322-342, 2024).

The implications of this statement and citation raise several issues and questions regarding the design, use, and analysis of AI in various settings.

Anchoring and relevance for TRUST (RA 8, RA 12, AC 11, and maybe AC14, and others?)

From RA 8: Interaction between humans and AI can generate dual emergent behavior in artificial and human cognition. AI-driven cognitive augmentation integrates technological artefacts into human cognitive processes, enhancing reasoning, sensemaking, and creativity. This can improve learning, productivity, health, and decision-making. However, the subsequent impact of AI on human cognition, and its feedback on retrained AI systems, remains unclear. TRUST will study the consequences of this circularity by modelling human-AI interactions using neuroscientific, computational, and visual methods, which will also offer new insights into human cognition and the subjective dimension of trust.

From RA 12: In organizations, AI’s content- and process-based qualities may challenge long-standing delineations between different forms of knowledge (factual/technical vs judgments/value-based) and expertise (technical vs domain-centric vs private consultants). We will investigate how AI engenders, supports, re-composes, and unsettles knowledge production and expertise deployment by studying some user partners with a qualitative practice-oriented research design.

Assumption: to describe, understand and explain the human-AI interaction is a foundational issue in Trust and in many research field in which Trust researchers will contribute.

Introduction/premises – hybrid intelligence/cognition

The aim of this project is to develop a position paper within the first six to twelve month that will serve as a point of departure for empirical studies on AI, as well as the design and analysis of human interactions with various AI tools and infrastructures. The concept of hybrid intelligence is employed in part of the research in educational psychology and subdomains that conceptualize learning, knowledge processes, and outcomes involving digital artifacts and AI tools. However, it is less clear what authors mean when they advocate for hybrid intelligence or hybrid cognition. A commonly used definition is:

Hybrid intelligence is about reinventing the way in which humans and technology coexist and what they can accomplish together (Dellermann et al., 2019).

Pragmatically, this definition makes sense in an everyday context, emphasizing the coordination and collaboration between humans and computational artifacts (AI tools). However, humans are not a monolithic entity in their capacities, nor are AI tools or infrastructures uniform. This implies the need to explore the human mind and actions both in itself and in interaction with other humans and different types of AI tools and infrastructures. The difference between emerging individual and collective cognition with AI-systems serve a more precise analytic starting point.

To describe, understand, and explain how and why AI functions in specific settings, such as selected fields in higher education and workplaces, we need to analyze at least three types of interaction structures: human-human interactions, human-AI interactions, and how multi-agent configurations interact with each other and with humans. Within these interaction structures, three key cognitive functions are foundational: attention, memory, and reasoning. These can be conceptualized as functions within a distributed cognitive system (Hutchins, 2020; 2025).

By AI tools and infrastructures, we refer to the computational artifacts through which human activities are performed (including design features), as well as the degrees of automation and augmentation related to processes and tasks, categories of knowledge, representations of knowledge, knowledge hierarchies, explicit and implicit ontologies, and epistemologies. The design and use of AI in these contexts will connect individual and collective aspects.

The principle to consider is: if knowledge is neither shared nor validated among humans, does it truly exist? Furthermore, does AI as an interactive configuration alter this principle?

Problem areas: multi-level problems

A foundational issue in studies of learning and knowledge development is known as the multi-level problem. This problem is connected to the unit of analysis and the methods or instruments one wishes to employ. To illustrate the implications, studying individual learning with specific instruments, such as knowledge tests, differs significantly from examining how a team of computer scientists agrees on code and procedures in the design of a new AI tool for a customer. Individual cognition can be studied as the development of executive functions without considering social norms or values.

In studies of groups and teams in work settings, researchers can analytically “zero out” the social aspects; however, one might also choose a unit of analysis that includes social and cultural layers as parts of a socio-technical-cognitive infrastructure. The assumption here is that studies of micro-interaction are too limited, and other layers need to be included.

When new AI tools and infrastructures are introduced, they can often disrupt learning and work processes within organizations, as new routines, norms, values, and divisions of labor need to be developed. This involves potential transformations of multiple interaction structures with varying degrees of automation and augmentation. In higher education, large language models (LLMs) create numerous disruptions related to learning and cognitive development, including cognitive efforts, self-regulation, motivation, norms and values, and grading. In work settings, LLMs can alter the validity of types of knowledge (numeric, qualitative, internal/external), even when supplemented with retrieval-augmented generation (RAG) solutions or domain-specific knowledge. Additionally, LLMs can change the status of different types of knowledge, the division of labor among professions, and the distribution of work within global networks. It is reasonable to contend that AI tools and infrastructures potentially disrupt work processes at multiple levels in organizations.

In both higher education and workplace settings, the problems and disruptions described can be articulated differently at a local level, yet they may still be related to underlying (causal) mechanisms within the organizations. This raises the question of how these mechanisms can account for contextual variation.

AI disruptions: Mechanisms and mechanical explanation

In many domains, including biology, neuroscience, cognitive science, and more recently, learning sciences, mechanisms and mechanical explanations are seen as significant advancements. This allows for a more coherent and cumulative path (Bechtel, 2009). In the human and social sciences, the term “mechanism” is most commonly used in the fields of cognitive science and economics. The argument here is that we can describe mechanisms and develop mechanical explanations for how AI tools and infrastructures become part of causal explanations. This could increase the precision and robustness of

theories in the social and human sciences (Caton, 2025). It's timely that we develop mechanical explanation in the interdisciplinary field of learning sciences.

Mechanical explanations should be treated as constitutive generalizations that encompass situated actions and cognition within specific contexts. One particular example of how AI tools can disrupt human activities is the sources used (knowledge resources). Sources can be structured in different ways—such as through LLMs or domain-specific systems—which can create tensions and contradictions in the flow of information and the potential for constructing shared knowledge.

In the selected areas of study, we need to define the unit of analysis and determine the level of generality at which we will describe mechanisms and causal explanations.

Modelling of learning trajectories and collective cognition in teams/organizations

Two key issues/concerns arise in the field of learning sciences. First, formal computational modeling is seldom used to test theories or develop new designs for learning, both for generic and domain-specific knowledge (Reimann and Baker, 2025). Second, modeling collective cognition or intelligence within organizations is also a rare phenomenon (Wolley, 2025).

With the introduction of new genAI tools and types of data, we should develop methodological approaches that can contribute to the formal modeling of individual learning trajectories. Additionally, we should create predictive models of elements and entities (mechanisms) that can facilitate productive collective intelligence and cognition in teams, larger groups, and organizations (Wolley et al., 2010). While some research exists on both issues mentioned, the advent of tailored generative AI allows us to take new steps in modeling specific phenomena.

Selection of paradigmatic cases

As mentioned, the first phase of this project will focus on articulating a set of key concepts and a conceptual structure for the multi-level analysis of the design and use of AI tools and infrastructures. The cases need to be based on both advanced AI design and actual usage in relevant settings.

In Higher Education:

This could involve small, controlled experiments in a few selected domains. The criteria are that modeling should create added value in understanding learning mechanisms across different levels and units of organization. Levels of analysis include knowledge structure and content (complexity within a domain), dyadic interaction with AI, and individual interaction with AI (such as through a tailored chatbot). The time frame for these experiments is 10 hours, distributed over time.

Design and testing of automatic feedback in multiple domains for students.

In Work Settings:

A small number of cases will be explored in different types of organizations, where parts of the organization implement new AI tools and work systematically to develop shared knowledge about how to use AI and why it can or will create added value for citizens or customers. In the first phase, ethnographic methods should be employed, and then, in agreement with the involved actors and organizations, specific interventions can be tested either in controlled settings or in real work processes.

In both settings, issues related to phenomena such as AI accuracy, bias, blind spots, complacency, opacity, probability, sourcing, reliability, and validity (both internal and external), as well as computational models and their correspondence to the real world, will need to be specified and studied within specific contexts. Many of these phenomena relate to questions about epistemology. The outcomes may provide generalizations at different levels.

High-Level Generic Questions

AI Behavioral Science: Jackson et al. (2025) argue in a recent paper based on workshops about AI in multiple fields of knowledge that we should consider whether we can conceptualize AI as a behavioral science. Alternatively, how can AI be integrated as a relevant unit of analysis across fields of knowledge? Below follows some questions inspired by Jackson et al and other sources in the sketch.

Relevant high-level generic questions could include:

Cluster 1: How can AI performance be assessed in different contexts? How can AI be designed to serve specific purposes for learning and knowledge development? Which types of normative models can be used to assess AI behavior?

Premise: Methods from social and human sciences.

Cluster 2: How can AI be used to gain improved insights into human behavior, such as executive cognitive function, shared memories, learning in specific domains, collective cognition in organizations, decision-making in dilemma-based situations, and decisions in complex systems with contradictory consequences (e.g., green transitions)?

Premise: Methods based on AI tools, such as formal modeling.

Cluster 3: How do students and AI interact when reasoning about complex problems? How do experts and AI interact when performing advanced tasks (e.g., diagnostic work, legal problems, estimation of efforts in complex tasks like software engineering)? How can coordination and communication in organizations be facilitated when AI influences multiple levels of interaction (individual, dyadic, community)?

Premise: New combined methods from behavioral, computational, and statistical sciences.

These questions imply forms of interdependencies that AI systems (of different types) can illuminate regarding our understanding of human individual and collective cognition (through modeling) and that AI systems can support human individual actions and collective efforts for the public good.

Interdisciplinarity – contribution from different fields of knowledge

Learning sciences (interdisciplinary in itself), educational psychology, sub-fields in psychology, cognitive anthropology, ethnography, sub-areas in computer sciences, philosophy of science and philosophy of mind.

Project leader and interested colleagues: Sten Ludvigsen, IPED, UiO; participants; Cato Bjøkli and Thomas Espeseth, PSI, Monika Nerland, Crina Damsa, Lene Pettersen, Gudrun Rudningen and Anders Mørch (all IPED), Omid Mirmotahari and Yngvar Berg IFI, UiO, Siv Gamlem University College Volda, Anders Malthe-Sørensen, Physics Dep. UiO, Malcolm Langford et al Law Faculty UiO. We will involve also other participants across field of knowledge and institutions based on interest. The issue of a position paper has been discussed in two meeting related to several RA and ACs, in November and 7.1 2026 (mostly related to the Law and Social Science part of Trust).

References (selective):

Bechtel, W. (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology*, 22, 543–564.

Carton, A.M. (2025) The Six Dimensions of Strong Theory. *Organization Science* 36(4):1242-1270

Dellermann, D., Ebel, P., Söllner, M. *et al.* (2019). Hybrid Intelligence. *Bus Inf Syst Eng* 61, 637–643. <https://doi.org/10.1007/s12599-019-00595-2>

Henrich J. & Muthukrishna, M (2024). Topic in Cognitive Science, 16, 322-342.

- E Hutchins (2020). [The distributed cognition perspective on human interaction](#). Roots of human sociality, 375-398. Chap. 14. Routledge.
- Hutchins. E. (2025). A New Cognitive Ethnography. Proceedings of the Paris Institute for Advanced Study, 2025, ff10.5281/zenodo.15828214ff. fffhal-05151677
- Jackson, M. et al (2025). AI Behavioral Science. <https://doi.org/10.48550/arXiv.2509.13323>
- Reimann, P. & Baker, M. (2025). Editorial notes: Mechanism as unifying construct for CSCL Research. (forthcoming, December 2025), IJCSCL
- Woolley, A.W. *et al.* (2010). Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science* **330**, 686-688 .DOI: [10.1126/science.1193147](https://doi.org/10.1126/science.1193147)
- Woolley, A.W. (2025). Generative AI and collaboration: opportunities for cultivating collective intelligence. *J Org Design*. <https://doi.org/10.1007/s41469-025-00199-z>