Report on selected AI research areas and use cases

AIOLIA DELIVERABLE 2.1

Horizon Europe Grant Agreement N° 101187937
AIOLIA PUBLIC

Funded by the European Union

Project acronym	AIOLIA
Deliverable number	D2.1
Description	Report on selected AI research areas and use cases
Lead beneficiary	CEA
Lead authors	Alexei Grinbaum (CEA), Laurynas Adomaitis (RISE)
Contractual delivery date:	31/05/2025
Actual delivery date:	31/05/2025
Sensitivity	PUBLIC

Document History

Name	Organisation	Role	Action	Date
Alexei Grinbaum	CEA	Task 2.1 lead	Early draft and bibliographic reseach	March-April 2025
Laurynas Adomaitis	RISE	Co-author	Early draft and bibliographic research	March-April 2025
	All partners	Online contributors	Use case descriptions	March-May 2025
Laurynas Adomaitis	RISE	Co-author	Complete literature review	09/05/2025
Alexei Grinbaum	CEA	Task 2.1 lead	Complete draft	11/05/2025
Elena Lazutkaite	EUREC	Reviewer	Review	12/05/2025
Laurynas Adomaitis	RISE	Co-author	Revision following review	13/05/2025
Alexei Grinbaum	CEA	Project coordinator	Draft for EC, WG, and SAB consultations	13/05/2025
	European Commission	Consultative	Review	14/05/2025
	SAB members	Advisory	Review	15/05/2025
Laurynas Adomaitis, Alexei Grinbaum	RISE, CEA	Co-authors	Implementation of suggested changes	22-23/05/2025
Efthemios Tambouris, Christina Tikva	CERTH	Reviewers	Review of complete draft	29/05/2025
Laurynas Adomaitis	RISE	Co-author	Final revision	30/05/2025
Alexei Grinbaum	CEA	Project coordinator	Final revision and validation	31/05/2025

Dissemination level	
PU	Public

How to cite

Grinbaum, A., Adomaitis, L., and AIOLIA consortium (May 2025). AIOLIA D2.1: Report on selected AI research areas and use cases.

Funded by
the European Union

Acknowledgements

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein. Reproduction is authorised provided the source is acknowledged.

During preparation of this report the authors have used OpenAI Deep Research and Gemini Deep Research LLM agents for editorial and bibliographic research purposes.

SUMMARY

AIOLIA deliverable 2.1 contains a selection of AI use cases both in EU and internationally, for co-creation of operational AI ethics guidelines in work package 3. The deliverable begins with a literature review on the design of artificial intelligence systems in relation to human cognition and behavior, and on human cognition and behavior as they evolve via interaction with AI systems. This review was informed via interaction with a working group of 5 leading scientists with a multidisciplinary background. The deliverable then continues with a presentation of the 10 use cases categorized under two major themes: change in professional and private behavior. Each use case provides an insight on the change in human cognition and behavior due to the use of AI systems, providing an empirical foundation for developing pedagogical materials, conducting training sessions, and informing policy measures later in the AIOLIA project.

Funded by
the European Union

CONTENTS

1. Task Description

This AIOLIA deliverable D2.1 presents the results of task 2.1, which is defined in the Grant Agreement as follows:

> *This task will review scientific landscape on human behaviour and cognition to support the development of, and impacted by, AI systems, via literature review and by convening a group of leading researchers who work at the frontier between AI and neuroscience, in order to identify AI research areas that stand out in three aspects: 1°, have the most profound impact on human cognition and behaviour; 2°, show immediate potential for scientific development; 3°, represent the most significant ethical challenges. In consultation with SAB and the networks, this task will select AI research areas for operationalization and choose practical use cases for the co-creation of guidelines in WP3. Global partners will choose one or more use cases most relevant in their national context.*

The methodology for selecting AI research areas and use cases is defined in the Grant Agreement in the following way:

> *AI research areas can be broadly divided into basic (exploring new scientific ideas and types of models) and applied (by type of data or functionality). The applied research areas are directly connected with practical use cases. Instead of spreading its resources across all AI areas, AIOLIA will select three that are most impactful for human cognition and behaviour, which will also become best practices for other AI fields. To do so, early in the project, we will convene a working group of leading computer scientists, ethicists, cognitive scientists, neuroscientists, linguists working on AI and the human brain and cognition to review the way the human brain learns and transfers knowledge, as well as the impact of AI on human cognition and behaviour. This multidisciplinary working group will review research beginning with scientific AI research areas, to evaluate which applied AI research areas are subject to the most impactful scientific advances, highest potential for quick scientific development, highest impact on human cognition and behaviour, and the most significant ethical challenges. The SAB will be consulted on the selection.*
>
> *To each selected research area AIOLIA will associate concrete use cases and match competent academic with industrial partners specialized in applying and expanding the selected research areas. The overall number of use cases will be three or four, each informing more than one research area. The participation of the industrial partner will ensure the bottom-up operability, while academic partners will offer specialist and research-based guidance. The selection of use cases in AIOLIA will be in line with the Greek concept of parrhesia – instead of shying away from difficult use cases that pose the most significant ethical challenges, AIOLIA will tackle them head on.*

2. Introduction

The impact of technology on human cognition is scientifically well-established, e.g. learning to read changes the structure of the visual cortex in children[1]. Studying this impact is a classic topic in the history of philosophy, dating as far back as Plato. In the *Phaedrus*, Plato famously puts into balance the invention of writing and the ensuing loss of the human capacity of memorization.[2] While Plato is critical of the writing's impact, the ensuing technological revolution brought about the externalisation of memory and paved the way for the development of new skills, creating massive societal change through formal education of large groups of humans.

Similarly, the invention of printing in the 15th century brought about another big change in human cognition. Information contained in books became more accessible to even broader groups of people, while learning to read printed text produced physiological changes in the cortex of entire populations. Culturally, a much larger corpus of texts became available for study by a higher number of students. Younger generations could surpass their parents in learning for the first time in human history. Self-education became possible. Marshall McLuhan argues that the widespread use of printing has "shaped human cognition into a visual-linear mode of thinking, externalizing thought into a visible organized structure ordered by grammatical rules and conventions, thus making thought and cognition more objectively critiqueable, calculable, and logically linear."[3] In essence, writing first, then the printing press have laid the groundwork for the enlightened individual of the 18th century: a rational citizen. These technologies can be seen as enablers of modern political thought and modern society.

In AIOLIA we believe that over time artificial intelligence will enact a comparable change. Generative AI brings personalized knowledge and tools instantly to every human, rather than relying on schools or universities. In the short term, AI renders obsolete certain cognitive skills and requires new ones at the level of the individual. In the longer term, the very way in which the human brain learns will evolve, resulting in subtle modifications in the organization of the cortex. This cognitive change will be accompanied by a change in individual behavior and social bond.

The mission of AIOLIA is to study emergent ethical concerns related to the change introduced by AI in human cognition and behavior, implying the type of change that has materialized or is turning real within the time framework of the project. For AIOLIA, this implies a focus on the AI research bringing visible change already in the short term, rather than hypothetical or yet uncertain longer-term shifts in cognition. The short-term cognitive change to be studied in AIOLIA is brought about by AI technologies that are put to market and reach large numbers of ordinary users in 2025-2027. This also implies that the consortium will avoid focusing on the futurist projections with regard to AI, e.g., "Artificial General Intelligence" (AGI) or "conscious AI systems". However likely or speculative one may take them to be, these AI visions are not part and parcel of everyday life in 2025. Even if these technologies do materialize one day, their large-scale market deployment will take months or years.

[1] Maurer et al., 2006; Brem et al., 2009; Ben-Shachar et al., 2011; Dehaene-Lambertz et al., 2018. References from https://pmc.ncbi.nlm.nih.gov/articles/PMC8721231/#:~:text=Thus%2C%20when%20a%20child%20learns,Dehaene%2DLambertz%20et%20al.%2C

[2] Plato. 1972. *Plato: Phaedrus*. Cambridge University Press.

[3] McLuhan, Marshall. 1962. *The Gutenberg Galaxy: The Making of Typographic Man*. University of Toronto Press.

Consequently, it is premature to argue about any real change in human cognition and behavior in connection to these futurist AI visions.

Contrary to such visions, the purpose of this review is to highlight *current* scientific developments that will assist the AIOLIA consortium in addressing ethical challenges that AI technologies pose in real-life R&D projects during the period of project implementation. This will allow the AIOLIA consortium to identify AI research areas of scientific and ethical relevance that: 1) have the most profound influence on human cognition and behaviour; 2) demonstrate strong immediate potential for scientific advancement; and 3) pose the most critical ethical challenges. The review offers an overview of AI and human cognition in terms of their core concepts, their historical co-development, their current state of mutual influence, and anticipated future trajectories.

Cognitive processes in the brain are shaped by major technological revolutions, and the advent of artificial intelligence is exerting an influence on the human condition. Historically, the speeding up and the simplification of knowledge access are the two factors that have created most important shifts in human cognition and behaviour, making generalized education possible, promoting critical thinking, and ultimately preparing the advent of the individualistic paradigm in the 18th-century Western world. While cognitive change in individuals occurs in a matter of generations, the ensuing large-scale social change may take centuries to materialize. However, the history of technology in the 21st century seems to shorten these time scales. The acceleration paradigm claims that the speed of change has so increased that it has become a central problem in the relationship between technology and society.[4]

Even before the current wave of generative artificial intelligence, digital technologies have been the principal source of this acceleration. The invention of mobile phones has enabled offloading the average human ability to memorize phone numbers, or numbers in general to their devices. The broader effects of social networks likely include reduced human attention spans. The advent of GPS systems led to a decline in the practice and thus the proficiency of mapping and navigation skills. Broader societal effects followed: with GPS, more people began to take previously quiet country paths; at the same time, the planning of family, professional, or social gatherings has vastly improved. Similarly, with social media, political polarization came to flourish; at the same time, many humans get more personalized information suited to their needs and desires. Arguably, the speed gap between technological and societal change has further accelerated with the propagation of artificial intelligence, leading to increased social tension.

Societal effects of AI become visible in a matter of years rather than centuries. In this sense, the speed of change itself constitutes a political and anthropological challenge.[5] AI ethics should raise the question of adaptability of human cognition to new technologies. If the speed of technological change cannot be matched by the speed of adaptive capabilities in human brain, the result might be that many more humans will have the feeling of "falling behind", while vast numbers of professionals in the labour market will urgently need a new type of professional expertise and training in mid-career. This creates cognitive inequality, which is already visible in how different generations of workers on the labour market include AI tools in their professional practice.

[4] Rosa, Hartmut. 2015. *Social Acceleration: A New Theory of Modernity*. Translated by Jonathan Trejo-Mathys. Reprint edition. New York: Columbia University Press.

[5] Jonas, Hans. 1985. *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*. University of Chicago Press.

3. Methodology

Human cognition, encompassing processes such as perception, learning, reasoning, and memory, serves as inspiration for AI research. Conversely, AI, as it develops and becomes increasingly integrated into human life, is actively shaping and reshaping human cognitive processes. This review explores both of these critical directions of influence: how our understanding of human cognition shapes the creation of AI, and how, in turn, AI technologies are impacting and transforming human cognition and behavior.

To understand ethical issues related to AI and human cognition and behaviour, one needs to understand the bi-directional influences between human cognition and the development of artificial intelligence.

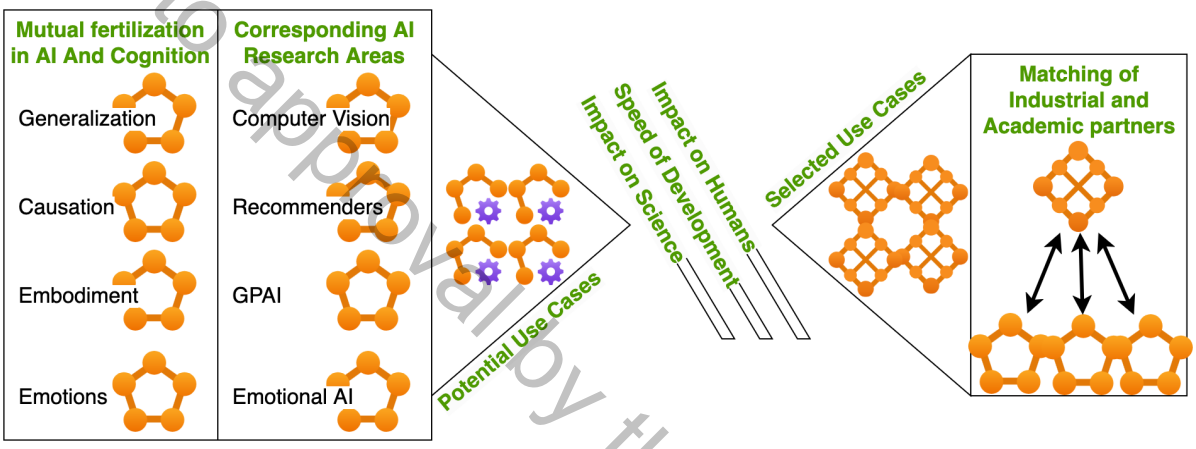

Figure 1. AIOLIA methodology for selecting use cases. There are several layers of selection, starting from left to right. First, during the literature review phase, and based on the discussions with the WG and SAB, we select four axes of mutual influence between AI and Human Cognition. Second, for each axis, there is a corresponding AI Research Area that best captures the bi-directional influences. Third, potential use cases are weighed and evaluated as representing the AI Research Areas in concrete operational environments. All three selection layers are considered in three cross-cutting criteria: impact on human cognition, speed and potential of scientific development, and impact on humans through ethical issues. Ranking all the criteria, particular use cases are selected. Lastly, academic and industrial partners in the project are matched based on their expertise and suitability for the co-creation exercise.

The analysis is grounded in an expert-based review of the existing literature by the Working Group created in AIOLIA task 2.1, composed of 5 leading experts in scientific domains related to human cognition and behaviour. The experts are not involved in the consortium. They provided advice on a *pro bono* basis.

The review methodology is based on expertise, rather than quantitative bibliometry. This is often called a "narrative review" methodology,[6] which is the "traditional" way of reviewing the existent literature and is skewed towards a qualitative interpretation of prior knowledge. Although this review

[6] Sylvester, Allan, Tate ,Mary, and David and Johnstone. 2013. "Beyond Synthesis: Re-Presenting Heterogeneous Research Literature." *Behaviour & Information Technology* 32 (12): 1199–1215. https://doi.org/10.1080/0144929X.2011.624633.

methodology lacks formal systematicity, it provides the reader with a carefully selected background for understanding current knowledge and highlighting the significance of new research. Its main advantages lie in synthesizing knowledge, rather than just providing results. Qualitative literature review recognizes that data needs a broader context for interpretation, and expert researchers need to consider how this context shapes the meaning of results being studied. This is achieved via engaging with the WG members collective and one-by-one. Each research field is shaped by its history, and in this narrative review, we choose to prioritize sensitivity to how it has developed.[7] The goal of the narration is to offer the reader a concise but precise overview providing an understanding of the bidirectional influence between AI and human cognition and behavior.

The selection of AI research areas in AIOLIA is based on the following criteria: the most impactful scientific advances, the highest potential for quick scientific development, the highest impact on human cognition and behaviour, and the most significant ethical challenges. These criteria are not rigorous or mathematical. Rather, we use expert consultation with the WG members and the narrative method to gain a contextualized understanding of each criterion and collect insights from the experts.

For sure, AI research fields are very broad and keep including more knowledge ever since the advent of generative AI in the previous decade. It is impossible to capture all of this science in one review. Here, we focus on what is most relevant for AIOLIA, i.e., the scientific developments related to human cognition and behavior that pose significant ethical challenges. While we do aim at a sufficiently exhaustive review, no such review can be absolutely complete in the very fast-changing domain of AI research. To make sure that we did not miss any significant content, we have consulted with AIOLIA's Stakeholder Advisory Board on May 14, 2025 and during the following week. Members of the SAB who participated in the discussion of this deliverable on a *pro bono* basis include: Brando Benifei (represented by staff), Maciej Chajnowski, Chiara Giovannini, Vinciane Gaillard, Dov Greenbaum, Natasha Govender, Maura Hiney, Theo Karapiperis, Tom Lindemann, Olga Megorskaya, Murielle Popa-Fabre, Pil Maria Saugmann, Natalia Shagarina, Maya Sherman, Simona Tiribelli, and Anupama Vijayakumar. Several insights from the SAB are shown in the text.

WG member	Affiliation	Expertise
Florent Meyniel	CEA/Neurospin	Neuroscience
Danko Nikolic	Frankfurt Institute for Advanced Studies	AI and cognitive science
Christiaan Vinckers	AUMC	Psychiatry
Karina Vold	University of Toronto	AI and neuroethics
Raja Chatila	Sorbonne University	AI, robotics, and AI ethics

[7] Kalpokas, Neringa, and Ivana Radivojevic. 2021. "Adapting Practices from Qualitative Research to Tell a Compelling Story: A Practical Framework for Conducting a Literature Review." *The Qualitative Report* 26 (5): 1546–66.

4. Artificial Intelligence: an overview

The field of artificial intelligence (AI) emerged as a distinct area of computer science in the 1950s, drawing inspiration and methodologies from diverse disciplines including human psychology, philosophy, and mathematics. Its development is directly inspired by, or attempts to model, aspects of human cognitive functions.

Much of AI development concentrates on specific applications, such as object recognition via the interpretation of human speech and visual data, autonomous navigation, or anomaly detection. AI technologies are now widely applied in transportation, education, healthcare, and communication, increasingly impacting how humans interact with technology and with each other. Benchmarking indicates that Gen AI systems are now matching or even exceeding human performance on standardized cognitive tests. However, they still lag in tasks requiring working memory.[8]

AI already has significant social, economic, and ethical impacts, directly and indirectly affecting human societies and individual experiences. These impacts include potential shifts in economic structures, changes in employment landscapes, transformations in the nature of work itself, and significant ethical considerations concerning autonomy, fairness, and the potential for misuse.

The development trajectory of AI has been marked by periods of rapid advancement and periods of stagnation, reflecting the importance of speed in innovation. Early AI researchers, drawing on digital computing and breakthroughs in information theory and logic, believed that achieving human-level AI was just a matter of decades. They proposed that intelligent computer systems could serve as models of human thought processes, indicating a strong link between AI development and cognitive science from the outset.[9]

The pursuit of "strong AI," aiming for genuine replication of human cognitive capabilities, has a history rooted in the formalization of thought by Thomas Hobbes, G. W. Leibniz, and George Boole. The field of AI emerged around 70 years ago, with Alan Turing's proposal of the Turing Test in 1950 and the foundational work of Warren McCulloch and Walter Pitts on artificial neurons in 1943, followed by Donald Hebb's learning rule in 1949. The pivotal Dartmouth Conference in 1956, attended by John McCarthy, Marvin Minsky, Allen Newell, and Herbert Simon, is considered the formal “birth” of AI. The progress in the field experienced a “quiet period” or an “AI winter” before the “AI Renaissance”, which is the current era marked by big data, machine learning, and ultimately transformers, like GPT. The notion of “strong AI” includes a capacity to perform any intellectual task a human being can, and hypothetically, to possess cognitive states comparable to human experience, raising fundamental questions about consciousness, intelligence, and the human mind.[10]

While many researchers pursued the “strong AI” vision, others questioned its feasibility or conceptual interest, particularly whether human cognition could truly be simulated using computational methods

[8] Galatzer-Levy, Isaac R., David Munday, Jed McGiffin, Xin Liu, Danny Karmon, Ilia Labzovsky, Rivka Moroshko, Amir Zait, and Daniel McDuff. 2024. “The Cognitive Capabilities of Generative AI: A Comparative Analysis with Human Benchmarks.” https://doi.org/10.48550/ARXIV.2410.07391.

[9] Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. 2018. “When Will AI Exceed Human Performance? Evidence from AI Experts.” arXiv. https://doi.org/10.48550/arXiv.1705.08807.

[10] Institute Problems of Mathematical Machines and Systems of the NAS of Ukraine, Ukraine, and Yashchenko V. 2024. “From McCulloch to GPT - 4: Stages of Development of Artificial Intelligence.” *Artificial Intelligence* 29 (AI.2024.29(1)): 31–44. https://doi.org/10.15407/jai2024.01.031.

alone,[11] leading to a divergence in AI research visions.[12] This latter group focused on developing AI systems capable of performing intelligent tasks within specific, limited domains, rather than general human-level intelligence, marking a shift towards more targeted applications of AI. The latter approach became known as “weak AI”[13].

“Weak AI” has become the dominant paradigm in contemporary AI research and development. It prevails in the research included in AIOLIA. This approach emphasises interdisciplinarity and integrating AI methods and techniques into existing fields, such as robotics, computer networking, and human-computer interaction. However, more speculative analyses also outline the possible risks and scenarios associated with “strong AI”, also called artificial general intelligence (AGI).[14]

There is another, more common notion of AI that is assumed in natural language, which means the user-facing aspects of computer applications. When non-specialists say they “use AI” or “interact with AI”, they mean they interact or use a particular application on a device, which consists of a graphical user interface (GUI), all kinds of alignment and safety filters, internal routing, and so on. Therefore, the common conception of AI is shaped by the models as much as the GUI. This means that the mode of interaction with AI makes a difference in how people understand AI, especially in folk narratives that form around it.

The key theories in human machine interaction (HMI) focus on the user's mental processes. Users develop internal representations of how a system works. The design of the AI's interface heavily influences the formation of these models, which may or may not accurately reflect the underlying AI's complexity or limitations.[15] Also, HMI is not just between a user and a machine, but as part of a purposeful "activity" situated within a context, so people often describe as “working together” with AI, or chatting with an algorithm as “looking for scenic spots”.

[11] Searle, John R. 1980. “Minds, Brains, and Programs.” *Behavioral and Brain Sciences* 3 (3): 417–24.
[12] Fjelland, Ragnar. 2020. “Why General Artificial Intelligence Will Not Be Realized.” *Humanities and Social Sciences Communications* 7 (1): 10. https://doi.org/10.1057/s41599-020-0494-4.
[13] Flowers, Johnathan Charles. 2019. “Strong and Weak AI: Deweyan Considerations.” In *AAAI Spring Symposium: Towards Conscious AI Systems*. Vol. 2287. https://ceur-ws.org/Vol-2287/paper34.pdf.
[14] Bostrom, Nick. 2016. *Superintelligence: Paths, Dangers, Strategies*. Reprint edition. Oxford: Oxford University Press.
[15] Norman, Donald A. 2008. *The Design of Everyday Things*. First Basic paperback, [Nachdr.]. New York: Basic Books.

5. Human cognition: an overview

Theories of human cognition that are relevant to AI development vary widely, often using either computers or the human brain as primary analogies. The computer analogy views cognition as information processing, whereas the brain analogy emphasizes connected, situated, and dynamic processes. Furthermore, the computer analogy is focused on tasks, their inputs and outputs, while the brain analogy aims for structural similarity to neurobiological systems, e.g. neuron networks. The computer approach is functional and specific, and the brain approach is structural and general.[16] Some hybrid models have been developed that combine elements from both symbolic AI (computational), and neural networks (brain-based). This is usually done to enhance the logical reasoning of AI systems with rule-based mechanisms.[17]

A core distinction between these analogies is that the computer model uses abstract symbols, while the brain model is grounded in sensorimotor experience. This brain-based view aligns with embodied cognition, recognizing that cognition extends beyond the brain to include external objects and artifacts.[18]

The computational model faces the symbol grounding problem: how do symbols acquire meaning?[19] The computational model struggles with symbol grounding because it generally relates abstract symbols to other abstract symbols, resulting in the lack of external meaning.[20]

Harnad suggested that sensory projections and conceptual categorization could solve symbol grounding.[21] Barsalou formalized this solution in what is known as the perceptual symbol system.[22] Drawing on cognitive neuroscience, Barsalou argued that cognition isn't separate from brain systems for action, perception, and introspection. He proposed that cognition arises from a combination of multimodal interactions with the world.[23] Predictive processing architectures in neuromorphic Gen AI are being explored as a future direction to address these grounding issues, potentially mimicking cortical hierarchy dynamics to better process sensory experiences.[24]

> **Working Group Insight**
>
> Currently, there is no scientific evidence that unsupervised or self-supervised machine learning could reveal radically new pathways or knowledge about the human brain, enabling completely new ways to manipulate brain data.

[16] Bermúdez, J. L. (2020). Cognitive Science: An Introduction to the Science of the Mind (3rd ed.). Cambridge University Press; Lieto, Antonio. 2021. *Cognitive Design for Artificial Minds*. London: Routledge. https://doi.org/10.4324/9781315460536.

[17] Marcus, Gary, and Ernest Davis. 2019. *Rebooting AI: Building Artificial Intelligence We Can Trust*. Vintage.

[18] Clark, Andy, and David Chalmers. 1998. "The Extended Mind." *Analysis* 58 (1): 7–19.

[19] Harnad, Stevan. 1990. "The Symbol Grounding Problem." *Physica D: Nonlinear Phenomena* 42 (1–3): 335–46.

[20] Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–24.

[21] Harnad, Stevan. 1990. "The Symbol Grounding Problem." *Physica D: Nonlinear Phenomena* 42 (1–3): 335–46.

[22] Barsalou, Lawrence W. 1999. "Perceptual Symbol Systems." *Behavioral and Brain Sciences* 22 (4): 577–660.

[23] Barsalou, Lawrence W. 2008. "Grounded Cognition." *Annual Review of Psychology* 59 (1): 617–45. https://doi.org/10.1146/annurev.psych.59.103006.093639.

[24] Keller, Georg B., and Thomas D. Mrsic-Flogel. 2018. "Predictive Processing: A Canonical Cortical Computation." *Neuron* 100 (2): 424–35.

Despite these approach differences, network frameworks now dominate cognitive models,[25] shifting the focus from localized brain regions to the dynamic organization of the brain as a network. Cognitive architectures (CAs) are models that can instantiate both computer and brain analogies. Adaptive Control of Thought—Rational (ACT-R) model and Soar CA are prominent CAs, but many others exist.[26] Symbolically-inspired CAs differ from biologically-inspired ones by using direct data rather than perceptual inputs. In contrast, biologically-inspired CAs utilize sensory and motor inputs through physical and simulated sensors. Biologically-inspired CAs are promising as they resemble human brain feature integration.[27]

CAs suggest human-like artificial cognition needs to: i) build causal world models, ii) ground learned concepts in environments,[28] and iii) generalize.[29] Emotion processing is also considered a crucial element, suggesting AI agents should iv) meaningfully produce and understand emotions. Neuroscience[30] and experiments[31] show that there are sensory-motor links in emotion processing. Therefore, emotion processing in AI is argued to be necessary for approximating social cognition.[32] Recent theoretical frameworks integrate generative models with CAs.[33]

[25] Barbey, Aron K. 2018. "Network Neuroscience Theory of Human Intelligence." *Trends in Cognitive Sciences* 22 (1): 8–20.

[26] Kotseruba, Iuliia, and John K. Tsotsos. 2020. "40 Years of Cognitive Architectures: Core Cognitive Abilities and Practical Applications." *Artificial Intelligence Review* 53 (1): 17–94. https://doi.org/10.1007/s10462-018-9646-y.

[27] Zmigrod, Sharon, and Bernhard Hommel. 2013. "Feature Integration across Multimodal Perception and Action: A Review." *Multisensory Research* 26 (1–2): 143–57. https://doi.org/10.1163/22134808-00002390.

[28] Marmolejo-Ramos, Fernando, Omid Khatin-Zadeh, Babak Yazdani-Fazlabadi, Carlos Tirado, and Eyal Sagi. 2017. "Embodied Concept Mapping: Blending Structure-Mapping and Embodiment Theories." *Pragmatics & Cognition* 24 (2): 164–85. https://doi.org/10.1075/pc.17013.mar.

[29] Lake, Brenden M., and Marco Baroni. 2023. "Human-like Systematic Generalization through a Meta-Learning Neural Network." *Nature* 623 (7985): 115–21. https://doi.org/10.1038/s41586-023-06668-3; Harnad, Stevan. 1990. "The Symbol Grounding Problem." *Physica D: Nonlinear Phenomena* 42 (1–3): 335–46.

[30] Holstege, Gert. 2014. "Chapter 20 - The Periaqueductal Gray Controls Brainstem Emotional Motor Systems Including Respiration." In *Progress in Brain Research*, edited by Gert Holstege, Caroline M. Beers, and Hari H. Subramanian, 209:379–405. The Central Nervous System Control of Respiration. Elsevier. https://doi.org/10.1016/B978-0-444-63274-6.00020-5.

[31] Marmolejo-Ramos, Fernando, Aiko Murata, Kyoshiro Sasaki, Yuki Yamada, Ayumi Ikeda, José A. Hinojosa, Katsumi Watanabe, Michal Parzuchowski, Carlos Tirado, and Raydonal Ospina. 2020. "Your Face and Moves Seem Happier When I Smile." *Experimental Psychology* 67 (1): 14–22. https://doi.org/10.1027/1618-3169/a000470.

[32] Pessoa, Luiz. 2017. "A Network Model of the Emotional Brain." *Trends in Cognitive Sciences* 21 (5): 357–71. https://doi.org/10.1016/j.tics.2017.03.002; Ziemke, Tom, and Robert Lowe. 2009. "On the Role of Emotion in Embodied Cognitive Architectures: From Organisms to Robots." *Cognitive Computation* 1 (1): 104–17. https://doi.org/10.1007/s12559-009-9012-0.

[33] Taniguchi, Tadahiro, Hiroshi Yamakawa, Takayuki Nagai, Kenji Doya, Masamichi Sakagami, Masahiro Suzuki, Tomoaki Nakamura, and Akira Taniguchi. 2022. "A Whole Brain Probabilistic Generative Model: Toward Realizing Cognitive Architectures for Developmental Robots." *Neural Networks* 150 (June):293–312. https://doi.org/10.1016/j.neunet.2022.02.026.

Learning impacts (Promises / Challenges / Needs)	Key components	Responsible stakeholders: Learners	Educators	Researchers	Policymakers	Technologists
Learning support	Scale personalised and adaptive support	●	●	●		●
	Learn within Zone of Proximal Development	●	●	●		
Learning resource	Efficient learning content creation		●	●		●
	Diverse multimedia learning resources		●	●		●
Learning feedback	Timely, specific, and constructive feedback		●	●		●
	Multimodal feedback delivery		●	●		●
Learning assessment	Adaptive and authentic assessments		●	●		●
	Enhanced real-world simulation		●	●		●
GenAI's imperfections	Hallucinations and inaccuracies			●		●
	Instability and biases			●		●
Ethical dilemma	Transparency and privacy concerns			●	●	●
	Equality and beneficence				●	●
Disruption of assessment	Human-AI hybrid cognition assessment		●	●		
	AI-induced performance illusion	●	●	●		
AI literacy	Basic understanding of AI systems	●	●	●	●	
	Critically evaluate AI-generated content	●	●	●	●	
Evidence-based decision making	Robust learning impacts evidence			●		
	Multi-party collaborative efforts	●	●	●	●	●
Methodological rigour	Evidence quality appraisal instruments			●		
	Long-term learning impact evaluations			●		

Figure 2. Overview of the impacts of generative artificial intelligence on human learning, categorised into promises (green), challenges (red), and needs (blue). From Yan, Lixiang, Samuel Greiff, Ziwen Teuber, and Dragan Gašević. 2024. "Promises and Challenges of Generative Artificial Intelligence for Human Learning." Nature Human Behaviour 8 (10): 1839–50. https://doi.org/10.1038/s41562-024-02004-5.

Regardless of the chosen paradigm, AI that mimics human cognition focuses on models that learn autonomously. However, while autonomous or unsupervised learning is the state of the art, it must be shaped by deliberate ethical considerations integrated directly into the design of such systems (ethics by design). This is often delayed to post-training alignment and fine-tuning procedures. Furthermore, ongoing human oversight and ethical reflection are crucial to ensure AI's responsible operation in real-world contexts. Developing and implementing ethical guidelines still remains a human endeavour that should be imposed on the self-learning systems.

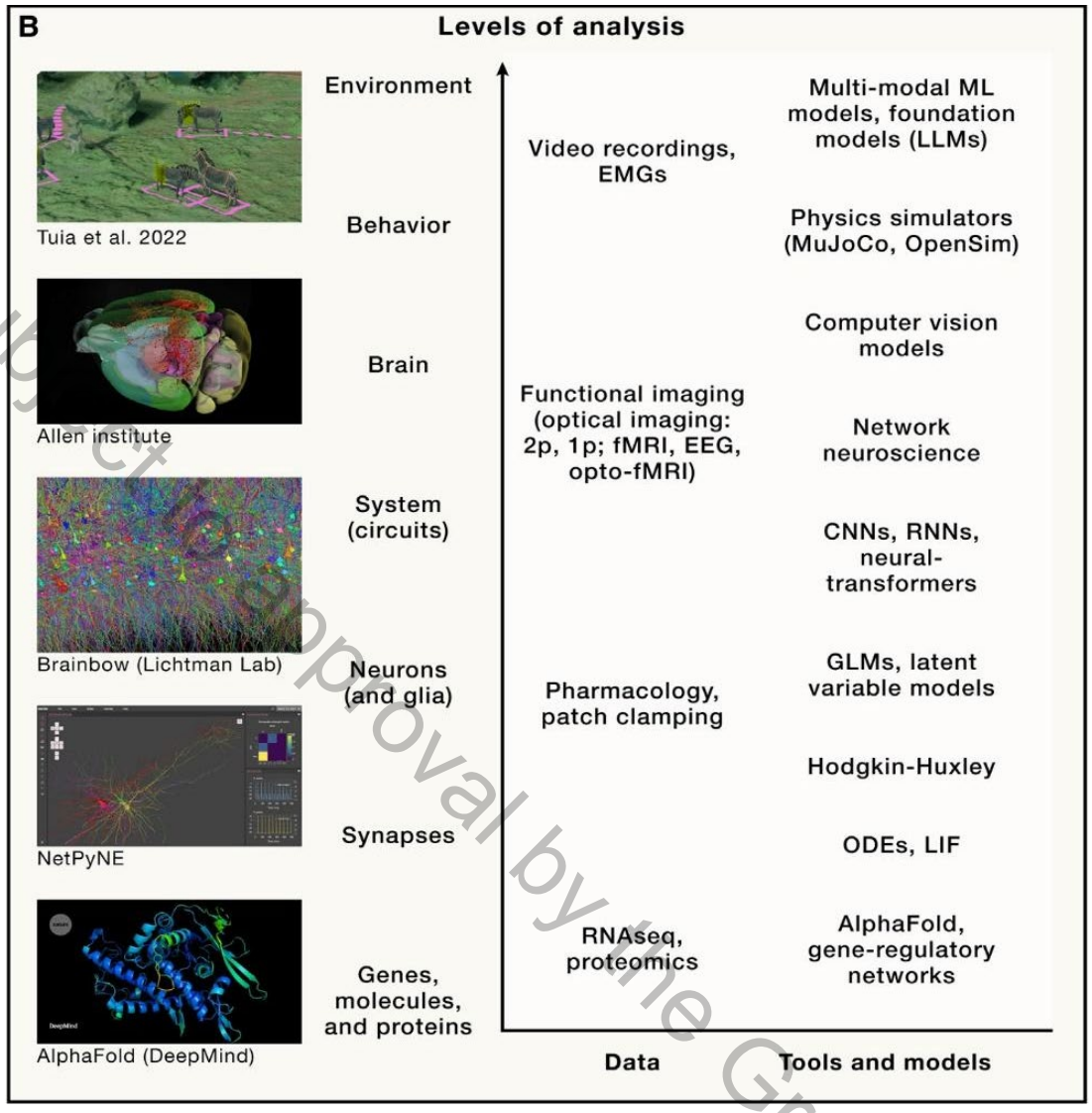

Figure 3. Human cognition and AI by level. From Mathis, Mackenzie Weygandt, Adriana Perez Rotondo, Edward F. Chang, Andreas S. Tolias, and Alexander Mathis. 2024. "Decoding the Brain: From Neural Representations to Mechanistic Models." Cell 187 (21): 5814–32.

6. Mutual influence between cognitive science and AI

Based on the dominating model of human cognition, we will further consider the bi-directional influence along the four axes and identify a use case that could potentially represent these issues. Using AIOLIA's experts, we then rank these research areas that: 1) have the most profound influence on human cognition and behaviour; 2) demonstrate strong immediate potential for scientific advancement; and 3) pose the most critical ethical challenges.

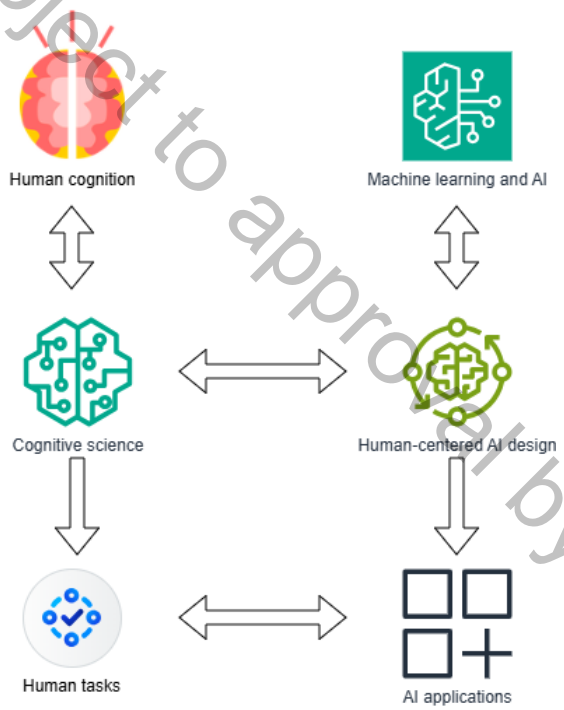

Figure 4. Mutual influence between cognitive science and AI.

Today, AI systems mediate many of the ways people perceive, decide, and behave individually and in society. For example, decision support systems on social media determine what information the users see online, thereby shaping opinions, beliefs, and even social dynamics. Generative AI has emerged as a cognitive aid across the entire palette of tasks from writing and coding to brainstorming or tutoring. This mediation is consequential: the choices made by the designers of AI systems, their algorithms and interfaces, as well as by the AI systems themselves during operation have an influence on human behaviour. In many cases this influence can be quickly judged as positive, e.g., helping students learn complex topics with targeted explanations, or negative, e.g., weakening critical thinking skills or leading to over-reliance on automated decision-making. The largest set of use cases lies in the middle: AI systems acting as mediators shape human behaviour in ways that are not evidently good or bad, however their influence changes the ways people think. One example is nudging, i.e., often implicit incitation to "do good things", like physical exercise.[34] Nudging in its various forms is widely spread and has a significant influence of human conduct individually and *en masse*, yet ethical judgement about nudging cannot be rushed.

Another example is cognitive offloading. People routinely use smartphones as external memory for notes or reminders, effectively enacting distributed cognition into their environment.[35] Offloading certain tasks can indeed save cognitive resources and improve efficiency. At the same time, while offloading improves immediate human performance, it may produce detrimental long-term effects on cognitive abilities.

[34] Thaler, Richard H., and Cass R. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press.

[35] Grinschgl, Sandra, and Aljoscha C. Neubauer. 2022. "Supporting Cognition With Modern Technology: Distributed Cognition Today and in an AI-Enhanced Future." *Frontiers in Artificial Intelligence* 5 (July):908261. https://doi.org/10.3389/frai.2022.908261.

Some experiments have found that people become complacent when an AI advisor is present, defaulting to the machine-made suggestion even when it is suboptimal, whereas others find that AI errors or hallucinations can trigger long-term aversion to AI, both of which alter normal decision-making patterns in human behaviour. However, during the AIOLIA Working Group Expert discussion, a clear red line was proposed - there is no evidence to say that the use of LLMs lead to illiteracy. It seems, that offloading tasks to LLMs still constitutes a relation to language, similar to reading or writing.

> Working Group Insight
>
> The use of LLMs or offloading tasks to them does not lead to illiteracy. Using LLMs constitutes a relation to language, similar to reading or writing.

AI systems that act as mediators in most types of human behaviour also shape the way humans spend and organize the time. For the younger generations born into the world of AI, this influence now occurs at the scale of their entire lifetimes. For example, recommendation systems that strive for attention may keep the user glued to a video game or a desirable product advertisement for a prolonged period of time. In growing brains, this creates a cognitive habit that shapes human cognition in a persistent and often irreversible fashion. Similarly, automated selection of food advice may influence well-being with consequences lasting through the person's entire life. Beyond the question of manipulation, the concern here belongs with irreversibility of the AI influence. The topics that humans find interesting, and busy their brains with, are increasingly suggested by AI systems to the extent that it would seem illusory to wish for a "world free of AI". The real question is how ethical judgement will evolve in this new world of accelerating and irreversible change, where humans live with, and are shaped by, AI systems.

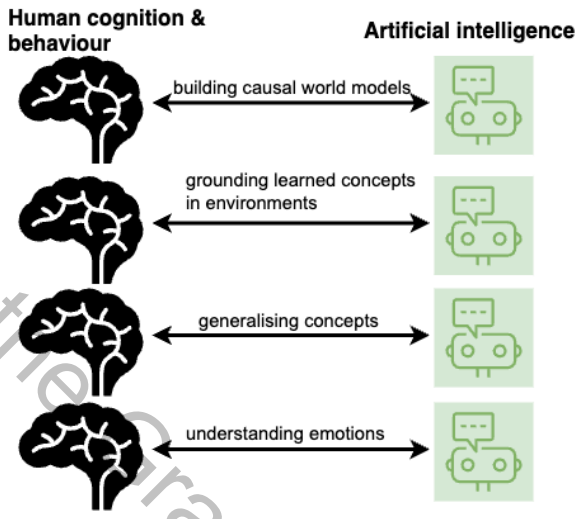

Figure 5. Four main axes of influence between AI and human cognition.

The four main axes of the influence between AI and human cognition and behaviour are i) building causal world models, ii) grounding learned concepts in environments, iii) generalising, and iv) understanding emotions. Each of those axes will have a bi-directional component. It will have an influence of human cognition into the design of AI, and AI's influence on human cognition and behaviour. This means that each axis will have two directions that need to be considered, producing eight vectors in total (Figure 5).

6.1 Generalization (in connection with Image recognition)

Almost any two things, events, or situations in the real world are unique. However, people learn to associate the properties of these unique encounters with other stimuli (Shepard, 1987). Generalization is the ability to apply a property from one stimulus to another stimulus when the two stimuli are close in an appropriate psychological space. It allows to apply learned knowledge and skills to novel situations and contexts that are different from the training data. Generalization differs from memorization and involves associations.

Humans constantly generalize based on prior experiences to deal with novel situations. Rather than treating each of these encounters as completely foreign, humans use prior experiences to expect certain things from the new situations. For example, we expect that we may exchange names when meeting a new person or that we will need to use a key to start the rental car.[36] Furthermore, we sometimes need to generalize over people, for example a hiring manager is trying to evaluate how a person would fit within the organization. This quality inevitably requires generalizing qualities and perceived competences over all the candidates.

Humans and image recognition AI systems share this ability to generalize. For instance, research in contrastive self-supervised learning has demonstrated how AI models can build representations by learning to encode what makes two visual inputs similar or different.[37] Similar to how humans can recognize objects from different angles or in different lighting conditions, these AI systems learn to extract invariant features from images. This approach has led to significant breakthroughs in self-supervised methods.[38] The process mirrors human visual cognition, where we naturally identify patterns and generalize across visual experiences, focusing on the underlying structure rather than superficial variations. For example, convolutional neural networks (CNNs) trained on ImageNet learn hierarchical features from simple edges to complex objects, which can then be transferred to specialized domains like medical imaging for tumor detection or industrial applications for defect identification. This transfer learning capability demonstrates how image recognition models can use knowledge gained from one visual domain to perform tasks in other contexts.

Nowadays, models trained using meta-learning approaches can identify patterns that generalize across domains by extracting domain-invariant features. For example, when trained on photos, art, and cartoons, these systems can accurately classify sketches despite never having encountered them during training, mirroring how humans use visual experiences to interpret new representations.[39] This scientific approach aligns with human cognitive processes, where knowledge acquired from multiple situations enables quick adaptation to new visual contexts. The PACS dataset experiment demonstrates this ability, where models learn to extract domain-invariant patterns essential for classifying images across all domains, including previously unseen ones.[40]

However, generalization remains a significant challenge for low-level vision models (typically trained on synthetic data), which often struggle with unseen degradations in real-world scenarios despite their success in controlled benchmarks. This generalization issue is a failure of existing training strategies, which lead networks to overfit specific degradation patterns. Studies show that guiding networks to focus on learning the underlying image content, rather than the degradation patterns, is key to

[36] Munakata, Yuko, and Randall C. O'Reilly. 2003. "Developmental and Computational Neuroscience Approaches to Cognition: The Case of Generalization." Cognitive Studies 10 (1): 76–92. https://doi.org/10.11225/jcss.10.76.

[37] He, Kaiming, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. "Momentum Contrast for Unsupervised Visual Representation Learning." arXiv. https://doi.org/10.48550/arXiv.1911.05722.

[38] https://doi.org/10.48550/arXiv.2209.06235

[39] Khoee, Arsham Gholamzadeh, Yinan Yu, and Robert Feldt. 2024. "Domain Generalization through Meta-Learning: A Survey." *Artificial Intelligence Review* 57 (10): 285. https://doi.org/10.1007/s10462-024-10922-z.

[40] Li, Da, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. 2017. "Deeper, Broader and Artier Domain Generalization." arXiv.Org. October 9, 2017. https://arxiv.org/abs/1710.03077v1.

improving generalization.[41] These findings suggest that successful generalization in image recognition requires a balance between learning the content distribution and handling various degradation types, much like how humans can distinguish between an object's intrinsic properties and circumstantial visual conditions.

6.2 Building causal models (in connection with Decision support)

This axis refers to the capacity of a cognitive system, whether human or artificial, to understand and represent the causal relationships. This capacity involves moving beyond simple pattern recognition to finding reasons for why things happen the way they happen. Essential difference between a correlation model and learning a causal model is that, once learned, the result is not probabilistic. Learning requires fewer steps in humans, who make a “leap” and “understand” causality, while AI systems generalize much later, if at all.

For a common example, think of a decision support system learning about user preferences. It differs from just observing that users who bought item A also bought item B. To have a causal understanding of user preferences, the decision support system will construct a causal model that distinguishes between which items users are exposed to and which items they genuinely prefer. Causal understanding allows prediction of what users will like in novel situations, not just based on historical patterns but on underlying preference mechanisms. This causation-based decision support system was introduced to provide fair and unbiased recommendations.[42]

Traditional decision support systems face several critical challenges due to their reliance on correlational patterns. These systems often suffer from popularity bias, where already-popular items receive disproportionate exposure, creating a self-reinforcing loop that narrows user experiences. Additionally, as demonstrated in studies across multiple datasets and metrics, the performance of decision support algorithms is highly dependent on dataset characteristics and evaluation metrics, with no single algorithm consistently outperforming others across different contexts.[43] This sensitivity highlights the limitation of purely correlation-based approaches. The real world is driven by causality, not just correlation. When decision-support systems naively extract patterns from observational data without accounting for the causal mechanisms that generated that data, they risk reinforcing existing biases, recommending items users would have discovered anyway, or failing to adapt to changing user preferences.[44] The potential outcome (PO) framework and structural causal model (SCM) represent two dominant paradigms for introducing causality into recommendation.

[41] Hu, Jinfan, Zhiyuan You, Jinjin Gu, Kaiwen Zhu, Tianfan Xue, and Chao Dong. 2025. “Revisiting the Generalization Problem of Low-Level Vision Models Through the Lens of Image Deraining.” arXiv. https://doi.org/10.48550/arXiv.2502.12600.

[42] Liang, Dawen, Laurent Charlin, and David M Blei. n.d. “Causal Inference for Recommendation.”

[43] McElfresh, Duncan C., Sujay Khandagale, Jonathan Valverde, John P. Dickerson, and Colin White. 2022. “On the Generalizability and Predictability of Recommender Systems.” In . https://openreview.net/forum?id=wO53HILzu65.

[44] Gao, Chen, Yu Zheng, Wenjie Wang, Fuli Feng, Xiangnan He, and Yong Li. 2023. “Causal Inference in Recommender Systems: A Survey and Future Directions.” arXiv. https://doi.org/10.48550/arXiv.2208.12397.

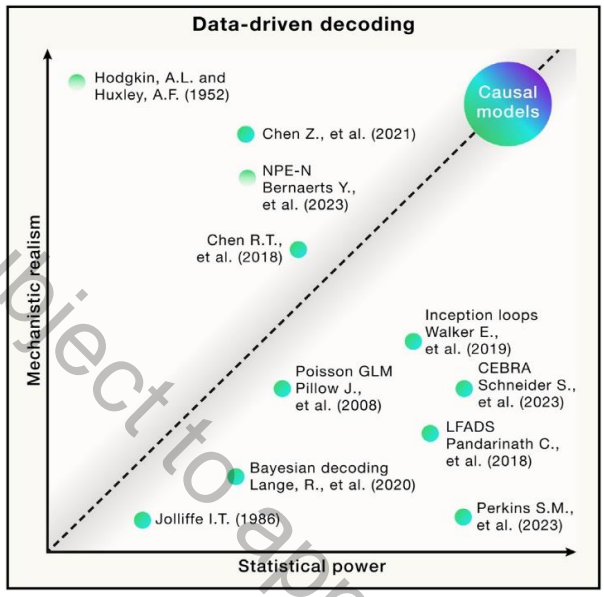

Figure 6. *Data-driven models: statistical power vs mechanistic realism. One wishes to map mechanism to computation in order to have causal, testable models. On one side are mechanistic models, and on the other statistical models that aim to best encode neural dynamics, but there is a large gap between them. From Mathis, Mackenzie Weygandt, Adriana Perez Rotondo, Edward F. Chang, Andreas S. Tolias, and Alexander Mathis. 2024. "Decoding the Brain: From Neural Representations to Mechanistic Models." Cell 187 (21): 5814–32.*

6.3 Grounding learned concepts in embodied learning

This axis addresses the symbol grounding problem discussed above. It emphasizes that meaningful concepts are rooted in sensorimotor experience and interaction with the physical and social environment.

In a famous thought experiment in cognitive science, Mary is a neuroscientist who studies colour but, in her own life, she has only been exposed to shades of black and white. She has all the scientific knowledge about colour, yet no qualitative experience of it. One could argue that the first time Mary actually sees colours, it will enhance her knowledge.[45] This additional cognitive value of experience is called "qualia". Similarly, our concept of an "apple" is not just a dictionary definition. Rather, it is grounded in the sensory experiences of seeing red, green, and yellow apples, tasting their sweetness and crispness, feeling their round shape, and performing motor actions like picking, biting, and chewing an apple. This embodied experience of "apple-ness" captures something about the world that a formal definition cannot grasp.

Qualia are a significant topic in the discussions of 3-dimensional systems endowed with AI, i.e., "intelligent" robots. For example, Rodney Brooks has argued that action or behavior are the appropriate sources of intelligence to be used in robotics, not brute computational solutions.[46] Oudeyer argues that robots should have an intrinsic motivation to explore and discover the world without extrinsic design.[47] Similarly, socially assistive robots,[48] and social robots in general need to be

[45] Jackson, Frank. 1986. "What Mary Didn't Know." *The Journal of Philosophy* 83 (5): 291–95. https://doi.org/10.2307/2026143.

[46] Brooks, Rodney. 2003. *Flesh and Machines: How Robots Will Change Us*. Reprint edition. New York: Knopf Doubleday Publishing Group.

[47] Oudeyer, Pierre-Yves, Frdric Kaplan, and Verena V. Hafner. 2007. "Intrinsic Motivation Systems for Autonomous Mental Development." *IEEE Transactions on Evolutionary Computation* 11 (2): 265–86.

[48] Feil-Seifer, David, and Maja J. Matarić. 2011. "Socially Assistive Robotics." *IEEE Robotics & Automation Magazine* 18 (1): 24–31. https://doi.org/10.1109/MRA.2010.940150.

immersed in social and group interactions to learn social behavior. Human behavior cannot be codified, thus it cannot be replicated in a rule-based system. The specificity of social robots is that most socially interactive robots do not need to physically interact with their environments in order to perform their tasks.[49] This phenomenon is mostly visible in current LLM-based chatbots, which are bots designed for social interaction, be it assistive, romantic, companionship, or other.

There is an interesting dynamic aspect of human-machine interaction and automation of tasks via the use of General Purpose AI (GPAI), especially Generative AI. As Bainbridge argues, automating tasks (even if the human handles exceptions), deprives the user of the routine opportunities to practice their practical judgement and exercise their cognitive skills needed for the task.[50] In a way, it takes away from their ability to encounter the operational environment of their field of expertise. A study by Lee et al. shows that experts use significantly less effort when performing their tasks with Gen AI.[51] Whether this can lead to the loss of critical thinking and competence is a very important open research question. Moreover, meta-analytic evidence indicates that human-AI collaborations perform worse than the best of humans or AI alone, especially in decision-making tasks, potentially due to conflicting reasoning approaches.[52]

Current GPAI systems, including advanced language models, fundamentally lack grounding in physical experience. They operate in a purely symbolic domain disconnected from the sensorimotor interactions that shape human cognition. This disconnection represents a significant limitation in their ability to develop true understanding of the world they attempt to model. In humans, sensorimotor experiences create a foundation for higher cognitive processes. Research by Beilock et al. on hockey experts' comprehension of action language revealed that experts' motor experience significantly enhanced their understanding of hockey-related language.[53] Activity in brain regions associated with high-level action planning correlated positively with both hockey language comprehension and hockey experience.

> **Working Group Insight**
>
> There is a difference between “hosted” and “integrated” embodiment. Often AI interface is simply put into a physical device but it does not lead to integrating it into the environment. Only AI that has a meaningful outlet to the environment can be considered integrated.

[49] Deng, Eric, Bilge Mutlu, and Maja J. Mataric. 2019. “Embodiment in Socially Interactive Robots.” *Foundations and Trends® in Robotics* 7 (4): 251–356. https://doi.org/10.1561/2300000056.
[50] Bainbridge, Lisanne. 1983. “Ironies of Automation.” *Automatica* 19 (6): 775–79. https://doi.org/10.1016/0005-1098(83)90046-8.
[51] Lee, Hao-Ping Hank, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. “The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers.” https://hankhplee.com/papers/genai_critical_thinking.pdf.
[52] Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone. 2024. “When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis.” *Nature Human Behaviour* 8 (12): 2293–2303. https://doi.org/10.1038/s41562-024-02024-1.
[53] Beilock, S.L., and S. Goldin-Meadow. 2010. “Gesture Changes Thought by Grounding It in Action.” *Psychological Science* 21 (11): 1605–10. https://doi.org/10.1177/0956797610385353.

True GPAI systems may require not only computational power but also interaction with the physical world to develop genuine understanding of concepts. Cognition arises from sensory and motor experiences rather than existing as an abstract entity divorced from sensorimotor experience. This is an open research area with a lot of potential for scientific development.

6.4 Cognition and emotions

Emotions are not separate from human cognition but are deeply intertwined with and influence cognitive processes. Understanding and processing emotions, both internally and socially, is a critical aspect of intelligence, empathy, care, but also the negative aspects of addiction, manipulation, and loss of trust.

For a human-centric example, hostage negotiators use tactical empathy to build rapport and trust with the other party. This involves actively listening to and acknowledging the other side's emotions and perspective, even if the negotiator does not agree with them. Chris Voss, a former lead FBI hostage negotiator, tells the story of how working at a suicide hotline has taught him the lessons needed for his later career.[54] In fact, it is a recommended practice for officers involved in crisis negotiation to volunteer in a suicide call center.[55] It teaches active listening, identifying and labeling emotions, mirroring, and having empathy.

Affective computing has evolved as a multidisciplinary field dedicated to understanding, modeling, and responding to human emotion through computational systems.[56] Early definitions emphasized the integral role of emotions in cognitive processes and highlighted the complexity of incorporating affect into technology. This complexity arises from the need to interpret various physiological, behavioral, and contextual signals accurately, as emotional expressions can manifest through facial cues, speech, textual data, physiological changes, and other subtle indicators. The field thus seeks to capture a more holistic picture of affect by integrating multimodal information.

Methodological developments in affective computing rely on data-driven analyses of multiple channels such as facial cues, speech patterns, text inputs, and physiological measurements that vary over time and context.[57] This multimodal perspective is also enabling emotion recognition in more naturalistic settings, such as conversational environments. Current AI systems, while not having emotions in a subjective way, can imitate them in their responses.

For example, a companion chatbot directly engages with human emotions.[58] Although it is not extrinsically given a way of doing it, it learns it in operation through feedback. The most common feedback metric for such chatbots is engagement - the longer the person chats with them, the more positive feedback it gets. Short or non-recurrent conversations mean negative feedback. This teaches

[54] Voss, Chris, and Tahl Raz. 2016. *Never Split the Difference: Negotiating as If Your Life Depended on It*. Random House.
[55] Soskis, David A., and Clinton R. Van Zandt. 1986. "Hostage Negotiation: Law Enforcement's Most Effective Nonlethal Weapon." *Behavioral Sciences & the Law* 4 (4): 423–35. https://doi.org/10.1002/bsl.2370040406.
[56] Picard, Rosalind W. 2000. *Affective Computing*. MIT press.
[57] Lian, Zheng, Bin Liu, and Jianhua Tao. 2021. "CTNet: Conversational Transformer Network for Emotion Recognition." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29:985–1000. https://doi.org/10.1109/TASLP.2021.3049898.
[58] Grinbaum, Alexei, and Laurynas Adomaitis. 2022. "Moral Equivalence in the Metaverse." *NanoEthics* 16 (3): 257–70. https://doi.org/10.1007/s11569-022-00426-x.

the chatbot patterns that it should follow to keep the user maximally engaged. This often involves playing into the user's guilt (e.g., "Why did you leave me alone for so long?") or other emotion mechanisms. This can lead to psychological dependence through loneliness, trust, and chatbot personification,[59] which are significant ethical issues in the field.

6.5 Conclusion

This section, including the subchapters, has delineated the bi-directional influence between artificial intelligence and human cognition. This interplay was examined along four principal axes: generalization, causal model building, embodied grounding, and emotional cognition. Each axis reveals distinct mechanisms though which AI systems shape human cognitive processes and behaviour, and conversely, human cognitive understanding informs AI development. The analysis underscores significant cognitive impacts and promising avenues for scientific advancement. Finally, it points to critical ethical challenges inherent in this evolving symbiosis.

Research areas	Cognitive impact	Scientific potential	Ethical concerns
Generalization	Alters human learning patterns and application of knowledge. Risk of over-reliance on AI's (potentially flawed) generalizations.	Advances in self-supervised learning and domain generalization. Creating AI that mimics human-like flexible generalization.	Propagation and amplification of biases. Ensuring reliability and fairness in AI systems across diverse contexts.
Building Causal Models	Potential for improved human decision-making through AI insights into causality. Risk of narrowed experiences if AI causality is flawed or biased.	Developing AI that uses causal understanding. Building fairer and more transparent decision support systems.	Ensuring AI identifies true causal links, not spurious correlations. Preventing manipulation through causal claims. Transparency of AI's causal reasoning.
Grounding Learned Concepts	Risk of human skill degradation and reduced critical thinking due to cognitive offloading. Irreversible changes in cognitive development.	Creating integrated AI systems with sensorimotor interaction. Investigating cognition as an embodied phenomenon.	Deskilling of individuals and workforce. Cognitive dependency on AI. Accountability for actions of AI systems lacking grounded understanding.
Cognition and Emotions	AI influencing human emotional states and social interactions. Potential for altered empathy, trust dynamics, and psychological dependence.	Developing sophisticated multimodal emotion recognition. Creating AI for empathetic and supportive human-AI interaction.	Risks of psychological dependence and emotional manipulation. Privacy of sensitive emotional data. Defining boundaries for AI's affective influence on humans.

[59] Xie, Tianling, Iryna Pentina, and Tyler Hancock. 2023. "Friend, Mentor, Lover: Does Chatbot Engagement Lead to Psychological Dependence?" *Journal of Service Management* 34 (4): 806–28. https://doi.org/10.1108/JOSM-02-2022-0072.

7. Human interaction modalities

7.1 Cognition and language

Large Language Models (LLMs), such as those in the GPT series, demonstrate remarkable capabilities in generating human-like text, often achieving a level of fluency that is difficult to distinguish from human writing.[60] These models excel at capturing statistical regularities, syntactic structures, and semantic associations present in vast text corpora. However, this proficiency in imitation does not necessarily equate to genuine understanding in the human sense. A fundamental divergence lies in the input/output pathways and the grounding of concepts. Human language comprehension and production are deeply embedded within a rich network of sensorimotor experiences, emotional states, social context, and world knowledge.[61] LLMs, trained primarily on text, lack this embodied grounding.

Research comparing the internal representations of LLMs and human brain activity during language processing suggests some shared computational principles, particularly regarding next-word prediction.[62] Studies indicate that the internal, layer-by-layer representations learned by these models predict human brain activity during natural language processing,[63] with performance often improving as model size increases (scaling effect).[64] However, this alignment does not imply identical processing. LLMs can process contexts far exceeding human working memory capacity, and limiting this context access can sometimes improve their simulation of human comprehension. Furthermore, fine-tuning techniques like instruction following, while improving alignment with user preferences, may not necessarily mimic natural human language comprehension processes. The semantic networks derived from LLMs may differ structurally from human ones, which integrate experiential information beyond statistical co-occurrence, as suggested by studies comparing GPT-3's network organization to human semantic memory.[65]

Human language processing is also constrained by cognitive architecture and biological predispositions. This has been the standard justification for universal grammatical principles (like hierarchical structure over linear order) and limiting the types of languages humans can naturally acquire.[66] Moro argues that the existence of "impossible languages"—rule systems the human brain

[60] Cohn, Neil, and Joost Schilperoord. 2024. A Multimodal Language Faculty: A Cognitive Framework for Human Communication. London ; New York: Bloomsbury Academic.
[61] Barsalou, Lawrence W. 1999. “Perceptual Symbol Systems.” Behavioral and Brain Sciences 22 (4): 577–660; and Harnad, Stevan. 1990. “The Symbol Grounding Problem.” Physica D: Nonlinear Phenomena 42 (1–3): 335–46.
[62] Goldstein, Ariel, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, et al. 2022. “Shared Computational Principles for Language Processing in Humans and Deep Language Models.” Nature Neuroscience 25 (3): 369–80. https://doi.org/10.1038/s41593-022-01026-4.
[63] Schrimpf, Martin, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. “The Neural Architecture of Language: Integrative Modeling Converges on Predictive Processing.” Proceedings of the National Academy of Sciences 118 (45): e2105646118. https://doi.org/10.1073/pnas.2105646118.
[64] Antonello, Richard, Aditya Vaidya, and Alexander Huth. 2023. “Scaling Laws for Language Encoding Models in fMRI.” Advances in Neural Information Processing Systems 36 (December):21895–907.
[65] Goldstein, Josh A., Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. “Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.” arXiv. https://doi.org/10.48550/arXiv.2301.04246.
[66] Moro, Andrea. 2016. Impossible Languages. 1st edition. Cambridge, MA: The MIT Press.

cannot process naturally—points to these innate biological constraints, suggesting language rules are not merely arbitrary cultural conventions. Experiments comparing brain activity when processing possible versus impossible linguistic rules show differential activation, supporting a biological basis for these constraints. LLMs, operating primarily on statistical pattern matching from their training data, are not inherently bound by these specific biological constraints, although their training data reflects human language patterns.

Users interacting with highly fluent LLMs may be prone to anthropomorphism, attributing human-like qualities such as understanding, intentionality, moral judgment, or emotional awareness where it does not exist in a comparable form.[67] Such misattributions can lead to misplaced trust, oversharing of private information, vulnerability to manipulation, or unrealistic expectations about the reliability and capabilities of these systems.[68] The lack of grounding means LLMs can generate outputs that are syntactically correct and statistically plausible yet semantically nonsensical, factually incorrect ("hallucinations"), or pragmatically inappropriate. Their inability to evaluate the real-world implications of the language they produce necessitates robust human oversight and critical evaluation, particularly in high-stakes domains. This challenges the notion of fully autonomous AI communication partners and underscores the need for users to maintain critical judgment when interacting with LLMs.[69]

7.2 Cognition and sensory data

AI systems process sensory data—visual, auditory, olfactory, tactile, and motor information—using computational techniques that often differ from biological sensory processing. While AI has achieved remarkable success in specific sensory tasks, these differences impact robustness, adaptability, and the potential for AI to replicate the richness of human perceptual experience.

Computer vision has made significant strides, particularly in object recognition, where deep learning models can achieve or surpass human-level performance on benchmark datasets. However, the underlying mechanisms often diverge. The traditional understanding of the human vision employs parallel processing streams: a ventral pathway primarily for object recognition ('what'/'perception') and a dorsal pathway for spatial processing and action guidance ('where'/'how' or 'vision for action').[70] This segregation allows for specialized processing, with the ventral stream associated with conscious perception and longer memory, and the dorsal stream with faster, potentially unconscious processing for guiding actions, possibly with shorter memory requirements.[71] In contrast, many standard computer vision systems rely on a single, predominantly feedforward pathway optimized for pattern

[67] Grinbaum, Alexei. 2023. Parole de machines. HUMENSCIENCES; and Shanahan, Murray. 2024, "Simulacra as Conscious Exotica." Inquiry 0 (0): 1–29. https://doi.org/10.1080/0020174X.2024.2434860..
[68] Epley, Nicholas, Adam Waytz, and John T. Cacioppo. 2007. "On Seeing Human: A Three-Factor Theory of Anthropomorphism." Psychological Review 114 (4): 864–86. https://doi.org/10.1037/0033-295X.114.4.864; and Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, and Atoosa Kasirzadeh. 2021. "Ethical and Social Risks of Harm from Language Models." arXiv Preprint arXiv:2112.04359.
[69] Placani, Adriana. 2024. "Anthropomorphism in AI: Hype and Fallacy." AI and Ethics 4 (3): 691–98. https://doi.org/10.1007/s43681-024-00419-4.
[70] Goodale, Melvyn A., and A. David Milner. 1992. "Separate Visual Pathways for Perception and Action." Trends in Neurosciences 15 (1): 20–25. https://doi.org/10.1016/0166-2236(92)90344-8.
[71] Milner, A. D., and M. A. Goodale. 2008. "Two Visual Systems Re-Viewed." Neuropsychologia, Consciousness and Perception: Insights and Hindsights, 46 (3): 774–85. https://doi.org/10.1016/j.neuropsychologia.2007.10.005.

recognition. This architectural difference may contribute to the brittleness of AI vision systems, rendering them susceptible to adversarial attacks (imperceptible input perturbations causing misclassification) or performance degradation when encountering inputs outside their training distribution—situations humans often handle more robustly. Such limitations have critical safety implications in applications like autonomous driving. Recognizing these limitations, research is increasingly exploring brain-inspired architectures. For example, dual-stream vision models explicitly mimicking the human ventral/dorsal pathways, potentially with different inputs and learning objectives, aim to enhance robustness, adaptability, and efficiency in AI vision systems.[72] Such models show promise in aligning with functional segregation observed in human brain activity. The move towards brain-inspired designs, incorporating principles like multimodal integration, robust learning frameworks, attention mechanisms, and efficiency constraints, represents a trend towards capturing functional principles of biological vision, such as resilience and adaptability, rather than just optimizing for narrow task performance.

AI has demonstrated significant capabilities in processing auditory information, especially speech. Speech recognition systems are ubiquitous, and recent advancements in speech synthesis, such as neural codec language models like VALL-E, can generate highly personalized speech from very short audio prompts, achieving high speaker similarity.[73] While impressive, challenges remain in achieving full prosodic naturalness and consistently conveying the subtle emotional and pragmatic cues inherent in human speech, highlighting the complexity of human vocal communication which integrates acoustics with context. Concurrently, research is exploring the decoding of speech perception directly from non-invasive brain recordings (MEG/EEG). Using techniques like contrastive learning and leveraging pretrained deep speech representations, researchers have shown it's possible to identify listened speech segments from brain activity with accuracy significantly above chance. This progress delineates a promising path towards non-invasive brain-computer interfaces (BCIs) for communication.[74] However, the ability to synthesize realistic voices facilitates malicious uses like generating deepfakes for fraud, disinformation, or creating false narratives, undermining trust and potentially causing significant harm. Furthermore, decoding neural signals associated with speech, or even thoughts inferred from neural activity, touches upon fundamental issues of mental privacy (neuroprivacy) and security (neurosecurity). The potential for "brain-hacking" (illicitly accessing or manipulating neural information via BCIs or other neural devices) demands careful consideration of consent, security, potential misuse, and the definition of new normative rules or "neurorights" to protect mental privacy, identity, and autonomy.[75]

[72] Lee, Jiyoung, Seungho Kim, Seunghyun Won, Joonseok Lee, Marzyeh Ghassemi, James Thorne, Jaeseok Choi, O.-Kil Kwon, and Edward Choi. 2023. "VisAlign: Dataset for Measuring the Alignment between AI and Humans in Visual Perception." Advances in Neural Information Processing Systems 36 (December):77119–48.
[73] Wang, Chengyi, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, et al. 2023. "Neural Codec Language Models Are Zero-Shot Text to Speech Synthesizers." arXiv. https://doi.org/10.48550/arXiv.2301.02111.
[74] Défossez, Alexandre, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. 2023. "Decoding Speech Perception from Non-Invasive Brain Recordings." Nature Machine Intelligence 5 (10): 1097–1107. https://doi.org/10.1038/s42256-023-00714-5.
[75] Ienca, Marcello. 2015. "Neuroprivacy, Neurosecurity and Brain-Hacking: Emerging Issues in Neural Engineering." Bioethica Forum Volume 8 (January):51–53. https://doi.org/10.24894/BF.2015.08015; and Burr, Christopher, and Nello Cristianini. 2019. "Can Machines Read Our Minds?" Minds and Machines 29 (3): 461–94. https://doi.org/10.1007/s11023-019-09497-4.

Compared to vision and audition, AI processing of olfactory information is significantly less developed (Bahremad et al., 2022). This lag stems from fundamental challenges: the high dimensionality of smell, significant inter-individual variability in perception, the difficulty in standardizing linguistic descriptions of smells, and the scarcity of large-scale, standardized digital datasets linking molecular structures to perceptual qualities.[76] Representing smells digitally remains difficult, and developing reliable 'digital noses' or integrating olfactory displays effectively faces hurdles. While machine learning models have shown some success in specific contexts (e.g., the DREAM Olfaction Prediction Challenge), a general, robust AI for olfaction remains elusive. Some argue that progress requires grounding models in the causal mechanisms of the biological olfactory system.[77] Integrating olfaction with other modalities like vision in robotics represents another avenue, though still facing challenges.

AI, particularly using techniques like reinforcement learning and self-supervised learning, is driving progress in robotic motor control and learning complex manipulation skills. Crucial to robust and adaptive manipulation is tactile sensing. Advanced vision-based tactile sensors, such as GelSight and its derivatives (e.g., GelSlim), provide high-resolution information about contact geometry and forces.[78] Integrating such sensors allows for more dexterous grasping, enabling robots to adjust grasps based on tactile feedback and perceive object properties. Learning policies directly from raw visuo-tactile data is an active area, with self-supervision used to learn compact multimodal representations to improve sample efficiency. This integration of action, perception, and touch allows robots to ground their learning in physical interaction.

7.3 Conclusion

The following table shows an evaluation of each interaction modality based on its cognitive impact on humans, its potential for scientific development or integration with AI systems, and the associated ethical risks.

[76] Keller, Andreas, Richard C. Gerkin, Yuanfang Guan, Amit Dhurandhar, Gabor Turu, Bence Szalai, Joel D. Mainland, et al. 2017. “Predicting Human Olfactory Perception from Chemical Features of Odor Molecules.” Science 355 (6327): 820–26. https://doi.org/10.1126/science.aal2014.

[77] Barwich, Ann-Sophie. 2021. “Imaging the Living Brain: An Argument for Ruthless Reductionism from Olfactory Neurobiology.” Journal of Theoretical Biology 512 (March):110560. https://doi.org/10.1016/j.jtbi.2020.110560.

[78] Donlon, Elliott, Siyuan Dong, Melody Liu, Jianhua Li, Edward Adelson, and Alberto Rodriguez. 2018. “GelSlim: A High-Resolution, Compact, Robust, and Calibrated Tactile-Sensing Finger.” In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 1927–34. https://doi.org/10.1109/IROS.2018.8593661.

Interaction modality	Cognitive impact	Scientific/AI potential	Ethical concerns
Language (via LLMs)	Anthropomorphism leading to misplaced trust and altered judgment. Potential for user manipulation and erosion of critical thinking.	Advanced text generation capabilities. Insights into statistical language processing. Large context processing.	Generation of misinformation ("hallucinations"). Susceptibility to manipulation and misuse. Privacy violations. Over-reliance by users.
Visual (AI Vision Systems)	Altered reliance on visual assessment. Potential for misjudgment due to AI's divergence from human visual processing mechanisms.	Near or super-human object recognition. Development of brain-inspired architectures for enhanced robustness and adaptability.	Vulnerability to adversarial attacks. Critical safety failures (e.g., in autonomous systems). Risk of misinterpretation of AI visual outputs.
Auditory (AI Speech & Brain Interfaces)	Changed human-technology interaction via speech. Potential for augmented communication (BCIs). Challenges to perceived authenticity.	Advanced speech synthesis and recognition. Direct decoding of speech perception from brain activity for BCIs.	Malicious use of voice deepfakes (disinformation, fraud). Neuroprivacy and neurosecurity violations. Potential for "brain-hacking". Need for neurorights.
Olfactory (AI Olfaction Systems)	Currently minimal direct impact. Potential future alteration of digital experiences or assistive olfactory technologies.	Nascent field with significant challenges (data, standardization). Potential in bio-inspired models and multimodal robotics.	Currently low. Future risks could involve manipulation via scent or privacy issues if technology becomes highly advanced and personalized.
Tactile/Motor (Robotics with Tactile Sensing)	Altered interaction with physical tasks and environments. Changed paradigms for human-robot collaboration and skill augmentation.	Development of dexterous robotic manipulation via advanced tactile sensing. Embodied AI learning through physical interaction.	Physical safety concerns in human-robot interaction. Potential for job displacement. Accountability for robot-induced harm. Tactile data privacy.

8. Cognitive and behavioral domains

Beyond processing modalities, AI exerts influence across various domains of human cognition and behavior. AI systems act as cognitive tools, partners, and environmental factors, shaping how individuals make decisions, create, interact socially, acquire and utilize skills, and experience autonomy. This section explores these impacts, highlighting the dual potential for augmentation and detriment, and the associated ethical considerations based on peer-reviewed literature.

8.1 Decision-making

AI is increasingly deployed as a decision support tool across diverse sectors, with fields such as "Behavioral AI" aiming to leverage data for improved decision-making (Tabrizi, 2025). AI systems can analyze vast datasets, identify complex patterns, and potentially help mitigate certain human cognitive biases.[79] However, their integration poses significant challenges. AI algorithms can inherit and amplify biases present in their training data, leading to discriminatory or unfair outcomes.[80] The lack of transparency and explainability in many complex AI models (the "black box" problem) hinders users' ability to understand, trust, and critically evaluate AI recommendations.[81] This opacity contributes to automation bias, the tendency for humans to over-rely on automated systems, accepting their outputs without sufficient scrutiny, even when incorrect.[82] This can lead to operators devolving responsibility and failing to detect errors, a key concern in the "control problem" of human-AI interaction.[83] The effectiveness of AI decision support thus depends critically on human users maintaining vigilance, critical judgment, and appropriate integration of AI inputs, necessitating AI literacy and critical assessment skills.[84] Shared responsibility in human-AI decision-making also creates accountability gaps when adverse outcomes occur.

8.2 Human creativity

Generative AI (GenAI) tools are impacting human creative practices. GenAI can enhance individual creativity by serving as a source of inspiration, helping overcome blocks, generating variations, and automating parts of the creative process.[85] However, emerging evidence indicates that widespread use might reduce the *collective* diversity of creative outputs, potentially leading to homogenization as

[79] Kahneman, Daniel. 2011. Thinking, Fast And Slow. http://archive.org/details/DanielKahnemanThinkingFastAndSlow.
[80] Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. "A Survey on Bias and Fairness in Machine Learning." ACM Computing Surveys 54 (6): 1–35. https://doi.org/10.1145/3457607.
[81] Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." Big Data & Society 3 (1): 2053951715622512. https://doi.org/10.1177/2053951715622512.
[82] Parasuraman, Raja, and Dietrich H. Manzey. 2010. "Complacency and Bias in Human Use of Automation: An Attentional Integration." Human Factors: The Journal of the Human Factors and Ergonomics Society 52 (3): 381–410. https://doi.org/10.1177/0018720810376055.
[83] Ienca, Marcello, and Roberto Andorno. 2017. "Towards New Human Rights in the Age of Neuroscience and Neurotechnology." Life Sciences, Society and Policy 13 (1): 5. https://doi.org/10.1186/s40504-017-0050-1.
[84] Lee, Hao-Ping Hank, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. "The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers." https://hankhplee.com/papers/genai_critical_thinking.pdf.
[85] Zhou, Eric, and Dokyun Lee. 2024. "Generative Artificial Intelligence, Human Creativity, and Art." PNAS Nexus 3 (3): pgae052. https://doi.org/10.1093/pnasnexus/pgae052.

models converge on statistically probable solutions.[86] Theoretical frameworks help contextualize AI's role; Margaret Boden distinguishes between combinational, exploratory, and transformational creativity.[87]

Combinational creativity involves generating fresh ideas by making unusual connections or combinations between existing concepts, like creating metaphors or artistic collages. Exploratory creativity operates within established rules or styles, systematically investigating the potential and boundaries of that conceptual space, much like a jazz musician improvising within a scale or a scientist working within a known paradigm. Transformational creativity, the most radical form, involves fundamentally altering the conceptual space itself by changing or rejecting its core rules, leading to paradigm shifts like those seen in major scientific revolutions or artistic movements. Current GenAI excels primarily at the first two, acting as a tool or collaborator. Whether AI can achieve transformational creativity or possess genuine "insight" remains debated.[88] The integration of AI challenges traditional notions of authorship, originality, and artistic merit, necessitating a re-evaluation of human creativity.

8.3 Social and personal relations

AI is increasingly participating in human social life, from recommendation algorithms in social networks to AI-powered companion chatbots designed for social interaction and emotional support.[89] These companions can offer perceived benefits like alleviating loneliness but raise ethical concerns. Humans tend towards anthropomorphism, attributing human-like qualities to socially interactive AI.[90] Epley et al. proposed a three-factor theory suggesting that anthropomorphism is influenced by cognitive and motivational drivers: elicited agent knowledge (the accessibility and applicability of human concepts in our minds), effectance motivation (the drive to understand, predict, and control one's environment), and sociality motivation (the fundamental need for social connection). When interacting with AI companions designed to mimic social cues, these factors can be strongly triggered, making users more inclined to perceive the AI not just as a tool, but as a social agent with genuine understanding or feelings.

AI companions use these mechanisms, simulating empathy to foster user trust and emotional connection.[91] This creates risks of users forming strong one-sided parasocial relationships, leading to unhealthy dependence, emotional manipulation, unrealistic expectations of human interactions, and

[86] Doshi, Anil R., and Oliver P. Hauser. 2024. "Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content." Science Advances 10 (28): eadn5290. https://doi.org/10.1126/sciadv.adn5290.
[87] Boden, Margaret A. 2005. The Creative Mind: Myths and Mechanisms. 2. ed., Reprint. London: Routledge.
[88] Halina, Marta. 2021. "Insightful Artificial Intelligence." Mind & Language 36 (2): 315–29. https://doi.org/10.1111/mila.12321.
[89] Chaturvedi, Rijul, Sanjeev Verma, Ronnie Das, and Yogesh K. Dwivedi. 2023. "Social Companionship with Artificial Intelligence: Recent Trends and Future Avenues." Technological Forecasting and Social Change 193 (August):122634. https://doi.org/10.1016/j.techfore.2023.122634.
[90] Grinbaum, Alexei, and Laurynas Adomaitis. 2022. "Moral Equivalence in the Metaverse." NanoEthics 16 (3): 257–70. https://doi.org/10.1007/s11569-022-00426-x; and Epley, Nicholas, Adam Waytz, and John T. Cacioppo. 2007. "On Seeing Human: A Three-Factor Theory of Anthropomorphism." Psychological Review 114 (4): 864–86. https://doi.org/10.1037/0033-295X.114.4.864.
[91] Deshpande, Ameet, Tanmay Rajpurohit, Karthik Narasimhan, and Ashwin Kalyan. 2023. "Anthropomorphization of AI: Opportunities and Risks." arXiv.Org. May 24, 2023. https://arxiv.org/abs/2305.14784v1.

exploitation of vulnerability.[92] Privacy is a major concern too, as intimate conversations reveal sensitive personal data, which users might mistakenly assume is confidential but could be used for model training or other purposes, eroding trust. Widespread adoption might influence societal attitudes towards privacy and reshape social skills, potentially reducing tolerance for the complexities of genuine human relationships.

8.4 Education

The applications of GenAI in education are beginning to be widely applied in creating personalized learning pathways and adaptive content, as well as the automation of administrative tasks like grading and the generation of diverse instructional materials in text, images, and video.[93] The perceived benefits are better student engagement, increased efficiency for teachers, and tailored educational experiences.

An important concern is the accuracy and reliability of AI-generated information, and knowledge conflict detection. Baidoo-Anu and Owusu Ansah highlight the risk of plausible yet incorrect or even fabricated content, including non-existent references, which students and educators might not critically evaluate.[94] This issue is made worse by the fact that these models are often trained on data only up to a certain point (e.g., 2021), limiting their knowledge of more recent developments and potentially leading to the dissemination of outdated information. Furthermore, biases inherent in the datasets used for training GenAI can be supported in the training materials. This can lead to inequitable educational outcomes. Generating content that reflects societal stereotypes or grading essays unfairly can lead to discrimination. The data collection required for personalized learning systems also raises privacy issues about student data security, storage, and usage.

Currently, the use of GenAI in academic contexts is known but not systematically addressed. It can be a threat to academic integrity and the development of cognitive skills. Noroozi et al. present a review showing that while GenAI can support tasks like qualitative data coding and lesson plan design, there are significant concerns about its impact on critical thinking and academic honesty.[95] The ease with which students can generate essays, summaries, or even research outputs using these tools makes it difficult to distinguish between student-authored and AI-generated work. This can lead to an over-reliance on AI, potentially curtailing the development of students' own critical thinking, argumentation, and creative capacities. The cultivation of AI literacy among both students and educators is necessary to encourage responsible use of GenAI in educational practices.

[92] McTear, Michael. 2021. Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots. Synthesis Lectures on Human Language Technologies, #48. San Rafael, California: Morgan & Claypool Publishers.

[93] Mittal, Uday, Siva Sai, Vinay Chamola, and Devika Sangwan. 2024. "A Comprehensive Review on Generative AI for Education." IEEE Access 12:142733–59. https://doi.org/10.1109/ACCESS.2024.3468368.

[94] Baidoo-anu, David, and Leticia Owusu Ansah. 2023. "Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning." Journal of AI 7 (1): 52–62. https://doi.org/10.61969/jai.1337500

[95] Noroozi, Omid, Saba Soleimani, Mohammadreza Farrokhnia, and Seyyed Kazem Banihashem. 2024. "Generative AI in Education: Pedagogical, Theoretical, and Methodological Perspectives." International Journal of Technology in Education 7 (3): 373–85.

8.5 Cognitive offloading and skill loss

Interaction with AI tools significantly impacts human skills through mechanisms like cognitive offloading—delegating tasks such as memorization, calculation, or writing to AI.[96] While potentially freeing resources, this carries the risk of skill atrophy due to diminished practice ("use it or lose it"). Powerful GenAI tools exacerbate these concerns across skills such as writing and critical thinking. Over-reliance on AI can weaken users' ability to evaluate information, detect bias, or reason independently, as they may uncritically accept AI outputs. This aligns with Bainbridge's (1983) "ironies of automation,"[97] where automating tasks can reduce operator expertise and capability to handle exceptions or failures. Bainbridge observed that attempts to automate human tasks often create unforeseen paradoxes: while automation handles routine operations effectively, it can leave human operators deskilled and less prepared to manage complex situations or system failures, precisely when their intervention is most needed. The core irony lies in automation removing the opportunities for regular practice that maintain crucial operator expertise and situational awareness.

Reduced direct engagement with tasks due to AI assistance can diminish practical judgment and tacit knowledge built through experience. Over-dependence on GPS navigation, for instance, has been linked to reduced spatial orientation skills.[98] Promoting AI as a *tool* augmenting skills versus a *substitute* bypassing skill development is crucial. Widespread skill degradation raises concerns about societal resilience, adaptability, and vulnerability to misinformation or AI failures.

The offloading and deskilling processes raises the questions about which foundational cognitive skills are non-negotiable for humans to retain, perhaps basic arithmetic, logical deduction, or text and visual composition. Addressing this proactively can mean a deliberate focus on 'mental fitness,' analogous to the physical fitness culture that emerged in response to increasingly sedentary lifestyles at the end of 20th Century. Just as gyms and structured exercise compensated for reduced physical demands in daily life, we might need to normalize practices aimed at preserving cognitive abilities. There are instances of such activities undertaken by older adults to maintain brain health, such as puzzles, learning new skills, or engaging in complex reasoning. However, they are not mainstream, lifelong habits.

8.6 Nudging and manipulation

AI systems, particularly persuasive technologies like recommenders and chatbots, are often designed to influence user preferences and behaviors, sometimes subtly nudging individuals towards choices serving external interests.[99] Burr and Cristianini argue that persuasive technologies, such as recommendation algorithms and social bots, are often designed not merely to assist but to actively

[96] Risko, Evan F., and Sam J. Gilbert. 2016. "Cognitive Offloading." Trends in Cognitive Sciences 20 (9): 676–88. https://doi.org/10.1016/j.tics.2016.07.002.
[97] Bainbridge, Lisanne. 1983. "Ironies of Automation." Automatica 19 (6): 775–79. https://doi.org/10.1016/0005-1098(83)90046-8.
[98] Ishikawa, Toru, Hiromichi Fujiwara, Osamu Imai, and Atsuyuki Okabe. 2008. "Wayfinding with a GPS-Based Mobile Navigation System: A Comparison with Maps and Direct Experience." Journal of Environmental Psychology 28 (1): 74–82. https://doi.org/10.1016/j.jenvp.2007.09.002.
[99] Burr, Christopher, and Nello Cristianini. 2019. "Can Machines Read Our Minds?" Minds and Machines 29 (3): 461–94. https://doi.org/10.1007/s11023-019-09497-4.

shape user preferences and behaviours, potentially steering individuals towards choices aligned with platform or commercial interests rather than their own self-determined goals.

This shaping of the choice environment can constrain the human experience.[100] Concerns exist about AI fostering addictive engagement patterns or exploiting psychological vulnerabilities. Designing AI systems that respect and support, rather than undermine, human autonomy is a critical ethical imperative. Furthermore, the increasing potential for AI to interface with neural data via BCIs introduces profound concerns about neuroprivacy and neurosecurity. The possibility of AI accessing, decoding, or influencing neural processes related to thought and intention challenges traditional notions of mental privacy and self-determination. This necessitates proactive ethical frameworks, potentially including "neurorights," to safeguard cognitive liberty and mental integrity against manipulation or unauthorized access. Ienca argues that traditional human rights may be insufficient to guard against novel threats like AI-driven neuro-manipulation or unauthorized mental surveillance via brain-computer interfaces.[101] He proposes establishing rights such as cognitive liberty, mental privacy, and mental integrity to proactively safeguard individuals' control over their own minds and cognitive processes, ensuring that technological advancements do not erode the very foundations of human self-determination and mental sovereignty.

8.7 Conclusion

This section has examined AI's influence across critical domains of human cognition and behaviour. From decision-making and creative endeavours to social interactions, education, skill retention, and personal autonomy, AI presents a dual capacity. It offers potential for significant augmentation, streamlining processes and offering novel support. However, this is counterbalanced by risks of bias amplification, skill erosion, diminished critical thought, and manipulative influence. A unifying thread across these domains is the need for AI literacy, ensuring transparency and accountability, and proactively developing frameworks to safeguard human autonomy and well-being.

[100] Valenzuela, Ana, Stefano Puntoni, Donna Hoffman, Noah Castelo, Julian De Freitas, Berkeley Dietvorst, Christian Hildebrand, et al. 2024. “How Artificial Intelligence Constrains the Human Experience.” Journal of the Association for Consumer Research 9 (3): 241–56. https://doi.org/10.1086/730709.

[101] Ienca, Marcello. 2015. “Neuroprivacy, Neurosecurity and Brain-Hacking: Emerging Issues in Neural Engineering.” Bioethica Forum Volume 8 (January):51–53. https://doi.org/10.24894/BF.2015.08015.

Cognitive and behavioral domains	Cognitive influence	Scientific/AI potential	Ethical concerns
Decision-making	Automation bias, reduced critical scrutiny; potential for human bias mitigation.	Advanced data analysis for improved decisions; development of transparent and explainable AI.	Amplification of data biases leading to unfair outcomes; opacity hindering trust and accountability; automation bias causing errors.
Human creativity	Augmentation of individual creativity; risk of reduced collective diversity and homogenization of outputs.	Generative AI as a tool for creative tasks; research into AI's capacity for higher-order creativity.	Homogenization of creative content; challenges to authorship and originality; potential devaluation of human creative skills.
Social & personal relations	Tendency towards anthropomorphism; formation of parasocial relationships; potential alteration of social skills.	Development of AI companions for social support and interaction; AI systems simulating empathy.	Emotional manipulation and unhealthy dependence on AI; privacy violations from sensitive data; erosion of authentic human connection.
Education	Personalized learning potential; risk to critical thinking development and academic integrity from over-reliance.	AI for adaptive content generation and administrative automation; tools for promoting AI literacy.	Dissemination of inaccurate or biased information; student data privacy concerns; threats to academic honesty and skill development.
Cognitive offloading & skill loss	Delegation of cognitive tasks leading to skill atrophy; reduced critical evaluation and practical judgment.	Design of AI tools that augment rather than replace human skills; research into "mental fitness" practices.	Widespread skill degradation impacting societal resilience; increased vulnerability to misinformation and AI failures; deskilling.
Nudging & manipulation	Subtle shaping of preferences and behaviors; potential for addictive engagement; challenges to self-determination.	Development of persuasive technologies; research into AI-BCI interfaces; design of autonomy-preserving AI.	Undermining of human autonomy and cognitive liberty; exploitation of psychological vulnerabilities; neuroprivacy and neurosecurity threats.

9. Use case selection

9.1 ADDITIONAL METHODOLOGY

The analysis of AI research areas regarding their impact on human cognition and behavior in the previous section has provided helpful insights for the selection of the AIOLIA industrial use cases. The use cases will be analyzed to co-create operational AI ethics guidelines at both a technical and a non-technical level. The internal objective is to use these guidelines in building AIOLIA pedagogical materials and training modules later in the project. The external objective is to use the operational guidelines in policy discussions on AI regulation in Europe and internationally.

While long-term changes in the structure of the cortex due to human-AI interaction appear almost certain, these changes will require time that extends well beyond the timeframe of the AIOLIA project. This project comes too early to collect empirical evidence about any physiological change in human brain. Current R&D projects do not directly address it yet. More tangible short-term changes concern human behavior, although epistemic effects linked to cognition should not be ruled out. The latter will most likely become visible in the education sector in the medium term. While AIOLIA will itself organize training sessions and educational activities for various adult target audiences (in work packages 4 and 5), this consortium does not specifically address the education of children, particularly of young children whose brain is undergoing significant restructuring. This is not to diminish the importance of such studies that remain to be conducted outside the scope of this project. For example, China has recently decided to strengthen AI education in primary and secondary school[102]; the Council of Europe is actively working on the use of AI in education.[103] If practiced at scale, AI education will surely enact a measurable change in human cognition. While AIOLIA includes a Chinese partner, we do not study early-stage education as a research domain and do not include partners with relevant scientific expertise. However, members of the Working Group in task 2.1 have provided important insights on this question, which are included in this deliverable, thereby concluding their WG mission. This topic is also covered is section 7.6 and will be addressed in work package 4 during the design of pedagogical materials, based on the experience of partner CERTH.

To address the impact of AI on human behavior, AIOLIA divides it into two questions bearing, respectively, **on professional and private behavior**. While the former covers a wide range of the types of professional conduct, the latter can be subdivided into individual effects and family-circle effects. The study of professional behavior will let us focus on the change in human expertise as part of human cognition oriented towards labor. The study of individual effects in the private sphere will also include studying cognitive change in human persons due to the use of AI systems.

The introduction of AI tools in human labor enacts the most swift and visible change. Industrial R&D projects of this kind are numerous. They immediately raise ethical questions that require expert evaluation and ethics-by-design measures. Addressing such issues is a feasible task within the timeframe of the AIOLIA project. We study this change in the professional behavior of human workers on a selection of use cases covering different industrial domains, involving several AI research areas,

[102] https://www.globaltimes.cn/page/202412/1324230.shtml

[103] Havinga, Beth, Wayne Holmes, and Jen Persson. 2024. "Preparatory Study for the Development of a Legal Instrument on Regulating the Use of Artificial Intelligence Systems in Education." Council of Europe. https://discovery.ucl.ac.uk/id/eprint/10201666/1/1680af118c.pdf.

and leading to different significant ethics issues that cover the entire range of fundamental ethical principles and values (see below and Deliverable 2.2).

The effects of AI in the private sphere can be subdivided into effects on the individual and the impact on the person's close circle (e.g., family). Here, behavioral and cognitive change are intricately linked, especially in the case of young children and elderly people. In AIOLIA we have the capacity to conduct several detailed studies of the use of AI systems as personal assistants to the elderly. We also study the emerging use of AI systems as family-level assistants, which assist in and reconfigure family relations. Finally, we address the now-classic issue of AI systems as individual virtual assistants on two examples: a general-purpose assistant and a griefbot. This large palette of the studies of AI systems in private behavior and cognition allows AIOLIA to build a complete picture of ethical issues that will be addressed in work package 3.

9.2 CHANGE IN HUMAN EXPERTISE AND PROFESSIONAL BEHAVIOUR

Use case 1: Medical doctors (radiologists, surgeons) using AI tools in diagnostics and treatment

AI research areas	Academic partners	Industrial partners
Decision-support systems Image recognition	AUMC	Oxipit Afliant

Description

Technology-wise Oxipit's suite can be categorised as primarily a Computer Vision algorithm and an AI decision support tool. It alters how clinical decisions are made and raises important questions about implementation of human-in-the-loop principle and responsibility and accountability (say, how and of what should the user be informed of). Oxipit's chest x-ray suite looks for 75 most common radiological findings and, in addition, identifies high confidence normal chest x-rays. This allows several applications. First, the product is CE certified (per MDR - EU Medical Device Regulation) for autonomous reporting of high-confidence normals. This means that radiologists are not required per MDR for up to 40% of chest x-ray reports, as they can be reported by the system automatically - raising the question of appropriate human oversight. Second, it has CAD (Computer Assisted Diagnosis) functionality - it identifies findings and provides initial indications along with heatmaps (overlay on chest x-ray image) to locate the suggested findings. This directly affects the radiological workflow, as a radiologist can consult AI findings prior to reading the image on his/her own. Additionally, this allows triage of chest x-rays based on the severity of AI findings. Third, it screens the final radiological reports for potential misalignment with AI image interpretation (a safety net). This allows to catch potentially missed findings where radiologists do not use CAD functionality.

Afliant is working on the integration of AI decision support systems for clinicians. One specific case tackled relates to use of an AI agent to support the treatment of abdominal aortic aneurysms (AAAs) at three key decision points: pre-operative planning, intra-operative execution, and post-operative follow-up. The agent's scope includes automatically extracting morphological features from CT scans to populate a planning worksheet (e.g., vessel diameters and lengths, aneurysm characteristics),

suggesting procedural details such as the choice of prosthesis and surgical access site, and highlighting critical vessels in real-time during surgery (via integration in a simulator environment). Post-operatively, the AI agent aims to predict the likelihood of complications such as endoleaks, guiding clinicians in determining patient follow-up schedules.

This integrated decision-support system is designed to reduce planning time, enhance procedural accuracy, and rationalize patient monitoring. Rather than replacing surgeons, the AI tool offers real-time, context-aware insights that can help reduce cognitive load and improve surgical outcomes. The technology's synergy with a high-fidelity surgical simulator (ANGIO Mentor) further enables training benefits, allowing trainees and experienced clinicians alike to evaluate how the AI's recommendations compare to their own assessments, in a setting that captures metrics such as speed, accuracy, and avoidance of adverse events.

Insights regarding change in human cognition and behavior

AI tools developed by Oxipit and Afliant re-draw the boundary between routine professional judgment and advanced expertise. The clinicians' and health professionals' behavior shifts from high-volume interpretation toward quality assurance, arbitration of edge cases, management and supervision of AI agents, and stewardship of population-level imaging strategy. Professional expertise therefore expands to include algorithmic literacy—understanding confidence metrics, heat-map visualizations, and failure modes—so that clinicians can decide when to trust the model, when to override it, and how to explain AI-mediated findings to patients and colleagues. Medical doctors must now document how they interacted with the AI system and why a particular AI suggestion was accepted or rejected, turning oversight itself into a recognized clinical competency.

In vascular surgery, surgeons who once honed intuition through manual planning now confront "second opinions" generated in milliseconds. Consequently, their expertise evolves toward integrative decision-making, scenario simulation, and dynamic risk management. Real-time AI assistants can reduce cognitive load, but only if surgeons cultivate the situational awareness to reconcile AI suggestions with tactile, visual, and hemodynamic cues in the operative field. Professional expertise, therefore, includes learning to interrogate the model's reasoning, manage uncertainty, and decide based on probabilistic forecasts.

Across both medical domains, clinicians are shifting from sole diagnosticians and procedural decision-makers to *epistemic stewards* who critically appraise, contextualize, and justify machine-generated recommendations. Radiologists and surgeons now share a new professional mandate: to exercise calibrated human-in-the-loop oversight.

Ethically speaking, this requires reinforced transparency, explaining uncertainty, and documenting the rationale for either accepting or overriding AI output, relating to ethical principles of accountability, trustworthiness, and human-centred approach.

Use case 2: Safety engineers using AI tools to speed up software release approvals

AI research areas	Academic partners	Industrial partners
Decision-support systems GPAI	CEA	NIT

Description

NIT is conducting a series of consultancy projects for the automotive industry, in which they are tasked to conduct safety analyses for the automotive products. Many automotive products include hardware and algorithms for autonomous and assisted driving. Analysis can be quite cumbersome and formal – NIT needs to follow the strict guidelines from standards such as ISO 26262 and ISO 21448. Especially for ISO 21448 – SOTIF (Safety of the Intended Functionality), potential hazards and risks are assessed using the System Theoretic Process Analysis (STPA). STPA has the goal to analyze control actions between the system components, identify unsafe control actions and whether they can lead to loss scenarios (induce harm to road users). This analysis is very important since its accuracy influences the safety-related architecture and safety measures which are needed to make vehicles safe for use.

SAB insight

Since AI safety is a universal concern beyond the EU, this use case may be particularly relevant for international cooperation and global discussions about AI ethics.

Automation of safety analyses is one of the most recent improvements in this area. The situation in the automotive market, in which the pace of the race quickened due to the aggressive approaches of "disrupting" companies in the electric vehicles segment (such as BYD and Tesla), is concerning. Cutting corners to release fast and frequently, with frequent software patches and upgrades, is becoming a growing practice. Automation of safety analyses would help approve the release updates faster, since the stage gates related to safety in the current DevOps (CI/CD) cycle include manual checks which may last for weeks. Automated tools include the addition of templates, databases of scenarios and failure modes, and more recently, AI suggestions.

Companies are adopting various strategies for the accelerated software updates for the vehicles in the field. The automation existing for general software development in CI/CD pipelines is now starting to become employed for other connected tasks, such as safety analyses and rechecks. Automating safety analyses by means of AI suggestions and prefills is particularly dangerous, since this analysis introduces safety goals which needs to be fulfilled for the safety to be preserved. NIT needs to understand whether the use case of AI-backed safety analysis leads to the unsafe releases and if it introduces complacency and negligence among the safety engineering population.

NIT is currently developing a tool in which AI agents are used to speed up and facilitate the STPA analysis. This tool would help the user by giving initial suggestions on how to conduct the analysis. It would also kick-off the analysis based on the general description of the system architecture, by providing the initial list of control actions and failure modes. It would also help assess the analysis which is nearly completed, to identify the remaining gaps and suggest fixes. This tool should speed up STPA significantly. NIT is beginning to use it in preliminary studies with an independent parallel manual process. As the tool should become a part of each STPA analysis, it is crucial that the end quality and safety at the output are preserved.

Insights regarding change in human cognition and behavior

The introduction of AI tools recasts the safety engineer's professional expertise from exhaustive manual decomposition toward epistemic and meta-analytic stewardship of machine-generated hazard assessments. Where expertise once centred on the enumeration of control actions and loss scenarios, it now expands to curating domain-specific templates, validating probabilistic AI suggestions, and documenting why a generated control-action taxonomy is accepted, refined, or

rejected. Safety engineers must therefore learn to cultivate algorithmic literacy and epistemically as well as pragmatically limited trust, understanding model provenance, scenario-coverage limits, and failure modes.

Use case 3: Recruiters using AI tools in hiring processes

AI research areas	Academic partners	Industrial partners
Decision-support systems GPAI	CENTRIC	Eticas + Barcelona Activa

Description

Eticas has worked on auditing AI-driven hiring systems, focusing on bias detection and fairness in algorithmic recruitment. One of its key projects involved Barcelona Activa's hiring process, where the goal was to analyze how AI systems screened and selected candidates. AI is used in recruitment to automate and enhance various stages of the hiring process, including candidate screening, interview selection, and final hiring decisions. In the Barcelona Activa use case, AI systems are designed to process large volumes of candidate data, extracting and analyzing structured information from CVs, application forms, and interview records to assess qualifications, skills, and work experience. This involves parsing unstructured text, matching candidate profiles to job descriptions, and scoring applicants based on criteria like educational background, work history, and specific skills. Moreover, Eticas has emphasized the importance of identifying and mitigating proxy features, such as names, educational institutions, or geographic locations, which can act as indirect indicators of protected attributes like gender or ethnicity.

At the core of this work is a commitment to fairness and non-discrimination. Eticas has translated these principles into operational steps by setting measures to evaluate impact, auditing AI models for biases in training and hiring decisions and ensuring that adjustments can be made to improve equal opportunities. Transparency and accountability have been central to the methodology, leading to the development of system cards and system maps that document how the AI functions, processes data, and makes decisions. By auditing systems at every stage (pre-processing, in-processing, and post-processing) Eticas has ensured that hiring algorithms remain explainable and accessible to non-technical stakeholders.

Another crucial aspect has been human oversight and interpretability. Rather than relying solely on algorithmic decisions, the approach emphasizes human intervention where necessary, improving explainability and ensuring that candidates have mechanisms to request feedback or challenge hiring decisions. Legal and ethical compliance is also a fundamental consideration, with audits aligning AI hiring tools with existing regulations and continuously monitoring their performance to detect any discriminatory outcomes. In sum, the fairness metrics include effectiveness, sensitivity, informative value, actionability, comprehensibility and compliance.

Bias mitigation strategies have played a key role in refining hiring algorithms. This has included testing systems with synthetic data to identify hidden biases, reweighting training data to improve representational fairness, and applying post-processing fairness adjustments when necessary. By integrating these operational steps, Eticas has worked to make AI-driven hiring systems not only

legally compliant but also fair, transparent, and ethically sound, reducing the risks of discrimination while promoting responsible AI use in recruitment.

Insights regarding change in human cognition and behavior

The AI hiring use case developed by Eticas is a clear example of how automated systems are deeply shaped by human cognition and behavior. Hiring decisions have always been influenced by subjective judgments, shortcuts, and biases. What Eticas shows through this project is that AI systems, far from being neutral, often absorb and reproduce these same biases when they are trained on imperfect or incomplete data. It stresses that algorithms do not operate in isolation: they reflect the assumptions, behaviors, and inequalities present in society. By focusing on bias auditing and fairness assessments, Eticas links the technical analysis of AI systems back to the very human behaviors they are meant to support or replace, demanding greater accountability in how these systems are built and used.

Use case 4: Security professionals using AI tools to detect hate speech

AI research areas	Academic partners	Industrial partners
Decision-support systems GPAI	CENTRIC	NEBRC, South Yorkshire Police

Description

As a method to combat rising rates of online hate speech, police forces are considering AI tools that scan social media content in real time to recognise and flag online hate speech which incites violence or hatred against vulnerable and minority communities. The AI capabilities scan public social media posts over multiple platforms (e.g. X, Instagram and TikTok) and flag posts that indicate online hate speech to law enforcement officials for further investigation. Flagged content is raised to a law enforcement officer who decides whether to pursue an investigation into the individual based on specific criterion (e.g. incitement to violence or harm, likelihood of harm, past content).

Such AI tools rely on natural language processing and machine learning algorithms to analyse large volumes of online text in posts and photos that could indicate hate speech. The AI model is trained on large public datasets where content was manually labelled as “hate” or “not hate”.

Such capabilities are intended to improve over time through continuous learning, feedback loops and fine-tuning the model. For example, if an officer deems content as hate and requiring investigation, the AI can use this decision to refine its algorithms.

SAB insight

This use case is particularly relevant for policy discussions at the EU level.

The use of this tool alters how police officers conduct investigations in terms of determining which individuals should or should not be investigated or considered as posing a threat. In addition, it raises several questions around responsibility, bias and privacy (whether individuals should be informed their social media is being surveilled).

Insights regarding change in human cognition and behavior

Deploying real-time GPAI and decision-support AI systems to identify online hate speech relocates investigative expertise from primary detection towards evaluation and gatekeeping. Police officers no longer begin unstructured data. The officers operate with a ranked queue of AI-flagged posts involving decision thresholds and contribute to feedback loops that they partially reinforce themselves. Professional expertise therefore widens to include critical competencies in bias auditing, evidential weighting, proportionality analysis, and avoiding overconfidence. Officers must discern when an AI "hit" satisfies the legal tests for incitement or threat, remaining alert to linguistic ambiguity, dialect variation, or adversarial manipulation. They must also provide new types of documentation, explaining how each investigative decision was shaped by algorithmic input and creating an auditable trail. Like in other use cases, police officers become somewhat like epistemic stewards who scrutinize, contextualize, and justify machine output.

Use case 7: Workplaces equipped with AI tools for behavioral analysis

AI research areas	Academic partners	Industrial partners
Emotional AI Image recognition Video analysis	Osaka University	NEC, Mitsubishi Electric, RiCOH

Description

Osaka University's industrial partners are developing workplace-video–analysis systems that couple fixed cameras and wearable sensors with machine-learning pipelines trained on annotated footage of real production lines. The core computer-vision model classifies each frame into fine-grained task steps, detects anomalies (e.g., slips, tool misuse), and recognises safety infringements such as workers entering exclusion zones or failing to wear helmets. Down-stream modules transform these detections into several application tiers:

- Productivity optimisation: footage is segmented into elemental tasks so that differences between novice and expert technique can be visualised; Mitsubishi Electric reports cycle-time reductions of up to 99 % after redistributing best practices.
- Automated reporting: NEC's platform converts daily camera streams into structured work reports, eliminating manual logs.
- Real-time safety: vision models trigger alarms for missing Personal Protection Equipment or boundary violations and can combine with facial recognition to identify the individual at risk.

- Well-being analytics: by fusing facial cues and heart-rate telemetry, an "emotion engine" flags stress or fatigue episodes that may degrade quality of work.

Insights regarding change in human cognition and behavior

Continuous AI surveillance in the workplace recasts factory labour from tacit, locally monitored activity into a digitally audited non-local performance space. Line workers must internalise camera-auditable standard micromovements, monitor dashboards, and anticipate that deviations will be reported to managers. Workers' behavior therefore includes basic AI literacy and feedback from human-machine interaction: understanding how one's motions are monitored, which thresholds trigger interventions, and how to contest false positives.

Supervisors' behavior evolves towards AI-facilitated orchestration of the workers. Their expertise shifts toward interpreting anomaly scores, curating training datasets, and evidencing that any disciplinary or safety decision satisfies transparency, fairness, and proportionality norms. In effect, both workers and supervisors become part of AI-mediated labor with a new expertise in explaining, challenging, and ethically contextualising machine judgments.

9.3 CHANGE IN HUMAN COGNITION AND PRIVATE BEHAVIOR

Use case 5: AI systems as personal and family virtual assistants

AI research areas	Academic partners	Industrial partners
GPAI Emotional AI	THWS	Immerstream Aurora First

Description

Within the AIOLIA project, THWS partners will propose two distinct personal assistant scenarios. The first involves a one-to-one virtual assistant geared toward human-AI role-play, where the primary user interacts privately with the system first describing the character they want to interact with and then communicating with this digital character emulated by a generative language model. The second focuses on a family-oriented assistant Aurora. Aurora's characteristics are pre-defined and fixed to optimise for safety and family-relevant context comprehension. The assistant engages with multiple individuals—such as parents and children—within a shared chat environment helping trace user preferences and account for them when planning joined activities, shopping or managing family calendars. Although both use cases necessitate the handling and storage of sensitive personal information (e.g., user preferences and family dynamics), the multi-user context in the family assistant poses additional complexity. The system must simultaneously manage and interpret data from several people while ensuring secure boundaries around each individual's personal details.

SAB insight

This use case, especially concerning family-level bots, is particularly relevant for policy discussions at the EU level.

Insights regarding change in human cognition and behavior

The introduction of always-on, memory-rich personal assistants recalibrates the very notion of "private life." In a one-to-one setting, individuals are nudged to externalize preferences, emotions, and daily routines so that the assistant can anticipate needs or role-play supportive conversations. Over time, this sustained self-disclosure invites new habits: users may curate their speech for algorithmic capture, lean on the assistant for emotional regulation instead of human confidants, or outsource planning and micro-decisions that once reinforced a sense of personal agency. The locus of privacy thus shifts from *whether* data are shared to *how* transparently the user can audit, redact, or reset the digital self that the model is constantly refining.

In a shared family assistant, the behavioral stakes are even more complex. Each utterance is simultaneously a disclosure of private thoughts or emotions to the machine and a potential reveal to other household members through AI system learning and inference. For example, the wishlist that a child wanted to keep secret may resurface in a group shopping reminder, or tension between parents may be amplified, rather than reduced, by the AI system's conflict-resolution strategy. Family members must negotiate new behavioral norms (e.g., "Can the bot mention my calendar to others?") and cultivate collective decision-making about how the AI assistant should arbitrate competing preferences. Consequently, private behavior is managed differently, as individuals learn to balance the convenience of a shared planning assistant with the imperative to preserve personal boundaries and authentic human communication.

Use case 6: Deepfake AI-based psychotherapy for processing trauma and grief

AI research areas	Academic partners	Industrial partners
GPAI (multimodal) Emotional AI	AUMC	ARQ Centrum '45

Description

Deepfakes can be categorized as a form of synthetic media created through advanced deep learning technology ('generative AI'). In the healthcare context, deepfake technology can be applied within psychotherapy, where AI-generated hyper-realistic video footage can be used to simulate conversations with, for instance, a realistic but fabricated representation of a deceased loved one (in grief counseling) or a perpetrator of trauma (e.g. in treatment of PTSD from sexual violence). The aim is to unlock closure through emotional processing or confrontation that would otherwise be impossible in traditional talk therapy, such as saying goodbye or confronting trauma. Other potential applications of deepfakes in healthcare are found in projects studying 'deepfaked' clinicians giving medical advice, eg to improve medication adherence. The first reported applications are exploratory and involve research groups in both the EU and US contexts. A research report from 2022 was the first

publication to describe the development of deepfake technology for psychotherapy, particularly PTSD-treatment through the confrontation with a deepfaked perpetrator of sexual violence.[104]

The implementation of a deepfake therapeutic protocol for such trauma treatment is currently being further studied in the Dutch evaluation study 'Blik op trauma' (EN: View on trauma)[105] and the AUMC partner is in contact with these researchers. Early reports suggest potential therapeutic value and positive experiences in specific cases, but deepfake technology also raises ethical questions about data protection and informed consent, as the process of creating the deepfake (which may require voice and visual data of a deceased person or a perpetrator), and it reconfigures how a therapeutic interaction is staged. As such, deepfake therapy risks impacting the boundaries between reality and simulation, and the psychological safety of the patient. In particular, it may introduce risks of or re-traumatization through the therapy itself or through the process of collecting data (images, videos, audio recordings) to create the deepfake. In the potential future application of deepfake technology for grief counseling, the patient may risk emotional overattachment to the deepfake.[106]

In the first article on ethics of deepfake therapy, Bak (AUMC) and colleagues argues that if the therapy is a last resort (i.e. when a direct conversation is not possible and other treatments are exhausted), clinically supervised, and the resulting media is handled securely and clearly marked as deepfake (e.g. by continuing to use the therapist's voice), then the interests of the patient may ethically and legally outweigh the rights of the person depicted, and deepfake therapy can be an acceptable new treatment strategy in mental healthcare.[107] The paper stated, however, that further guidance will be needed on the specific responsibilities of therapists and on how deepfake technology in mental healthcare can fulfil Trustworthy AI requirements, thus making it a good use case for interdisciplinary normative reflection in AIOLIA. Moreover, developments are fast and in the near future, chatbots might be involved in deepfake therapy; currently, automated (AI-generated) conversation scripts are being developed for use by chatbots during the intake before a deepfake therapy session.[108] The potential combination of AI-generated video with audio and chatbots, makes deepfake therapy a truly multimodal form of GPAI, which might exacerbate the risks previously identified and create new ethical concerns. Therefore, there is a need for anticipatory guidance and setting up safeguards before widespread clinical adoption.

Insights regarding change in human cognition and behavior

Deepfake therapy shifts the therapeutic encounter from an exclusively interpersonal dialogue to a hybrid interaction mediated by AI-simulated presence. This affects cognition and behavior of both patient and therapist. As described above, patients may experience powerful emotional reactions to a digital interaction that blurs the boundary between real and artificial. Therapists must now not only evaluate psychological readiness and vulnerability, but also judge whether the therapeutic fiction will facilitate or distort healing. Therapists are therefore evolving into facilitators of emotionally charged, AI-mediated interactions, potentially requiring new competencies such as algorithmic literacy,

[104] https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2022.882957/full
[105] https://www.sass.nl/projecten/blik-op-herstel and https://ntvp.nl/kennisdeling/nieuwsbrieven/nieuwsbrief-maart-2025/deepfake-technologie-de-behandeling-van-seksueel
[106] Kraaijeveld S, Ivanova D, Bak MAR. Het nieuwe rouwen? Over deepfakes en relaties met overleden dierbaren. [The new grieving? On deepfakes and relations with deceased loved ones]. Podium voor Bio-ethiek. 2024:31;4. https://nvbe.nl/wp-content/uploads/2025/03/20250305-Podium-24-4-digitaal.pdf
[107] https://jme.bmj.com/content/early/2024/09/30/jme-2024-109985.long
[108] https://ivi.uva.nl/content/news/2024/09/deepfakes-with-therapeutic-effect.html

informed risk assessment of psychological triggers, and the ability to debrief simulated experiences ethically. Professional responsibility expands to include not only managing therapeutic relationships but also overseeing the design, timing, limits, and impact of AI involvement in that relationship. Principles like trustworthiness, psychological integrity, and human-centeredness will be central.

Use case 8: AI systems for smart elderly care in Wuxi

AI research areas	Academic partners	Industrial partners
GPAI Emotional AI Image recognition Decision-support systems	CASTED	Industry: Jiangsu AIYU Wencheng Elderly Care Robot Co., Ltd. (highlighting ethical conflicts in product design). Academia: Renmin University, Guangxi University of Science and Technology (theoretical ethical frameworks). Government: Ministry of Civil Affairs, Wuxi Elderly Care Association (policy-practice alignment).

Description

The Da Tou A Liang (大头阿亮) elderly care robot, developed by Jiangsu Aiyu Wencheng Elderly Care Robot Co., Ltd., is an AI-powered companion designed to enhance the quality of life for seniors, particularly those living alone. Launched in 2025, this robot integrates artificial intelligence (AI), Internet of Things (IoT), big data analytics, and sensor technologies to provide comprehensive support in health monitoring, emotional companionship, and daily assistance.

Key Features:

- Health Monitoring: Equipped with sensors to track vital signs (heart rate, blood pressure, blood glucose) and detect emergencies like falls (triggering alerts within 3 seconds).
- Emotional Interaction: Uses natural language processing (NLP) and dual AI engines (iFlyTek and DeepSeek) to engage in conversations, play music, and adapt responses based on emotional cues (e.g., detecting loneliness or distress).
- Daily Assistance: Reminds users to take medication, controls smart home devices (lights, appliances), and schedules activities.
- Remote Connectivity: Allows family members to monitor seniors via video calls and receive real-time health updates through a mobile app.
- Autonomous Navigation: Uses computer vision and obstacle-avoidance algorithms to patrol homes and ensure safety.

The robot operates within Wuxi's smart elderly care ecosystem, which includes cloud platforms, IoT-enabled devices, and partnerships with local care institutions.

Insights regarding change in human cognition and behavior

Da Tou A Liang addresses cognitive and behavioral challenges faced by elderly users:

- Cognitive Support:
  - ◊ Memory Assistance: Automated reminders counteract age-related memory decline (e.g., medication schedules).
  - ◊ Information Accessibility: Simplified voice commands and touchscreen interfaces reduce cognitive load for users unfamiliar with technology.
- Behavioral Adaptation:
  - ◊ Routine Reinforcement: The robot's structured prompts (e.g., exercise reminders) encourage healthier habits.
  - ◊ Social Interaction: By simulating companionship, it mitigates loneliness—a critical factor in mental health decline among seniors.
- Safety Assurance: Real-time monitoring reduces anxiety for both seniors and caregivers, fostering trust in autonomous systems.

Use case 9: AI systems as personal companions assisting senior citizens

AI research areas	Academic partners	Industrial partners
GPAI Decision-support systems	STEPI	Hydol, Dasom, ROAIGEN (KIST)

Description

Senior care chatbots are multimodal companion robots aimed at home-based care for senior citizens. The platform delivers entertainment, companionship and ambient monitoring, while simultaneously collecting continuous streams of biometric, behavioural and emotional data. Each device couples far-field microphones and a 360-degree radar with an on-board AI stack that executes:

- Conversational intelligence: cloud speech-to-text / text-to-speech, intent recognition that draws on an ontology of daily-living concepts, and a short-term or long-term memory layer that stores user facts so the agent can maintain context for a certain period of time.
- Affective sensing: LLM APIs are used to classify sentiment in spoken utterances, while pose-recognition modules cross-check visual cues for immobility and motion.
- Context-aware dialogue and scheduling: the robot blends object recognition with knowledge of the user's calendar to suggest taking medication, physiotherapy or leisure activities in the user's own dialect and preferred speech style.
- Autonomous mobility and safety: in some versions, a mobile base patrols the home, issues emergency calls if falls are detected, and can locate household items on request.

Insights regarding change in human cognition and behavior

Elders may begin curating their speech, facial expressions or daily routines to elicit positive feedback from the device, internalising its nudges as normative expectations. Over time, decision-making about medication, exercise or social contact can migrate from reflective self-regulation to near automatic compliance with AI assistant suggestions, potentially altering the user's feeling of autonomy and agency.

Cognitively, the presence of a seemingly empathetic interlocutor encourages anthropomorphism and emotional dependency: personal memories, grievances and hopes are off-loaded to a non-human confidant. The user must therefore learn to monitor what the robot "knows" in order to preserve their sense of self as their cognitive capacities invariably undergo a robot-induced change.

Use case 10: Generative Ghosts and the Grieving Process

AI research areas	Academic partners	Industrial partners
GPAI Emotional AI	McGill University	Canadian Psychological Association, Association des psychologues du Québec, Canadian Counselling and Psychotherapy Association

Description

Griefbots (sometimes called deathbots, deadbots, thanabots, or generative ghosts) are AI chatbots designed to simulate the personality and speech of a deceased loved one. They range from basic chatbots interacting with users through text-based interfaces to more advanced virtual avatars or robots aiming to mimic how a person looked, sounded, and moved.

These chatbots are based on Large Language Models (LLMs). They are likely to become widely available as it becomes easier to develop personalized chatbots through RAG systems, which are relatively easy to deploy and cheap. Retrieval-augmented generation (RAG) systems allow conversational AI to deliver more precise and contextually relevant responses than generic LLMs by drawing on external information sources in addition to a basic model's internal knowledge. In other words, a user or a company could easily use a LLM to draw on the digital data of a recently deceased person – which can reflect a person's speech pattern, sense of humour, preferences, typical activities, etc. – to create a chatbot supposed to behave like they would. These griefbots challenge how we think about death and how we should protect a person's interests and rights over time, risk impacting the wellbeing of grievers, and affect how we approach the grieving process.

Insights regarding change in human cognition and behavior

Griefbots raise philosophical questions at the ontological, relational, and socio-political levels. At the ontological level, they notably raise the question of how to think of personal identity over time and the ontology of the griefbot itself: should it be conceived as an extension of the living person? That is, can we consider that, under some conditions, the deceased persists as a griefbot? Or should they be conceived in a way akin to pictures and videos representing the deceased? How we behave towards these systems and whether we should accord them some intrinsic value or authority – to settle or

inform disagreements on how to deal with the inheritance, for instance – depends heavily on how one responds to these questions.

Relationally, these bots are not only a way to mimic a deceased individual, but they are likely to be used mainly by grievers – a particularly vulnerable population – to assist them through the grieving process or to maintain a certain bond with the deceased. This raises the question of what it means to "relate" to virtual persons, if this type of relationship can have intrinsic or instrumental value, and how it challenges our conception of the grieving process. These chatbots are novel in that they are not simply pictures or videos of past memories, but they allow for new types of non-scripted interactions with mimics of the deceased. This is likely to have a significant impact on the wellbeing of the grievers, and change how we approach the grieving process; we need to tackle the question of how to deal with grievers using such chatbots in healthcare.

Socio-politically, it is also particularly interesting to consider the impact of such griefbots in a diverse and multicultural country such as Canada. These automated bots may be particularly useful for people living in isolated communities, where access to more traditional healthcare and psychological support is difficult, and they are likely to be perceived differently, depending on the socio-cultural background against which they are deployed.

9.4 CONCLUSION

This section presents the AIOLIA portfolio of 10 industrial and socio-technical use cases to anchor AIOLIA's co-creation of AI ethics guidelines in concrete ethics-by-design work. This ambitious selection is the result of extensive consultations within the consortium as well as with external experts (WG and SAB members). The selection of 10 use cases follows logically from these consultations and represents a challenge chosen by the consortium in comparison to the initial intention to select 3 or 4 use cases.

Guided by the narrative review and expert consultations, the consortium first separated impacts on professional expertise from those on private cognition and behaviour. Five use cases therefore analyse how AI reshapes human judgement at work—from radiology and vascular surgery to software safety, algorithmic recruitment and online hate-speech policing. The other five use cases probe AI's influence in the private sphere through individual and family-level assistants and chatbot companions, griefbots, and smart-elderly care. Together they span the project's selected AI research areas: decision-support, image and video recognition, conversational and multimodal GPAI, and emotional AI. The use cases selection in AIOLIA allows for a meaningful comparison of AI ethics guidelines across EU and non-EU partners and ensures: (i) coverage of important cognitive and behavioural domains, (ii) a spectrum of ethical challenges, (iii) practical policy relevance, (iv) relevance to use in designing pedagogical materials and training in AIOLIA, and (v) feasibility for co-creating operational guidelines in WP3.

Funded by
the European Union

10. Selection tables

10.1 AI research areas

AI research area	Impact on cognition and behavior	Potential for scientific development	Level of ethical challenges
Decision-support systems (knowledge representation and reasoning)	High	Moderate	High
Image recognition	Moderate	Limited	Moderate
Video analysis	Moderate	High	Moderate
General Purpose AI: conversational generative AI	Very high	High	Very high
General Purpose AI: multimodal generative AI	Very high	Very high	Very high
Emotional AI	Very high	High	Very high
Bayesian learning	Low	Limited	Limited (out of AI Act scope since February 2025)

This table shows the AIOLIA assessment of AI research areas according to three main selection criteria. Color codes: light orange – high relevance, included in the AIOLIA selection on multiple use cases; light purple – moderate relevance, included in the AIOLIA selection on a small number of use cases; light grey – low relevance, not included in the AIOLIA selection.

10.2 Use cases by EU partners

#	Use case description		AI research areas	Academic partners	Technological and industrial partners
1	Change in human expertise and professional behaviour	Medical doctors (radiologists, surgeons) using AI tools in diagnostics and treatment	Decision-support systems Image recognition	AUMC	Oxipit Afliant
2		Safety engineers using AI tools to speed up software release approvals	Decision-support systems GPAI	CEA (LIST)	NIT
3		Recruiters using AI tools in hiring processes	Decision-support systems GPAI	CENTRIC	Eticas + Barcelona Activa
4		Security professionals using AI tools to detect hate speech	Decision-support systems GPAI	CENTRIC	NEBRC, South Yorkshire Police
5	Change in human cognition and private behavior	AI systems as individual and family-level virtual assistants	Conversational GPAI Emotional AI	THWS	Immerstream Aurora First
6		Deepfake therapy for processing grief or trauma	Multimodal GPAI Emotional AI	AUMC	ARQ Centrum '45

This table summarizes AIOLIA use cases selected by EU partners. They are grouped into two main categories: change in human expertise and professional behavior (color code light blue) and change in human cognition and private behavior (color code light green). The selection within the EU contains 4 use cases of the first type and 2 use cases of the second type.

10.3 Use cases by non-EU partners

#	Use case description		AI research areas	Academic partner	Technological and policy partners
7	Change in human expertise and professional behaviour	Workplaces equipped with AI tools for behavioral analysis	Emotional AI Image recognition Video analysis	Osaka	NEC, Mitsubishi Electric, RiCOH
8	Change in human cognition and private behavior	AI systems for smart elderly care in Wuxi	GPAI Emotional AI Image recognition Decision-support systems	CASTED	Industry: Jiangsu AIYU Wencheng Elderly Care Robot Co., Ltd. (highlighting ethical conflicts in product design). Academia: Renmin University, Guangxi University of Science and Technology (theoretical ethical frameworks). Government: Ministry of Civil Affairs, Wuxi Elderly Care Association (policy-practice alignment)
9		AI systems as 'personal companions' to assist senior citizens	GPAI Decision-support systems	STEPI	Hydol, Dasom, ROAIGEN (KIST)
10		AI systems as grief-alleviating personal assistants (griefbots, chatbots designed to imitate a deceased loved one)	GPAI Emotional AI	McGill	Canadian Psychological Association, Association des psychologues du Québec, Canadian Counselling and Psychotherapy Association, MILA (Québec Artificial Intelligence Institute), EmoScienS.

This table summarizes AIOLIA use cases selected by non-EU partners. They are grouped into two main categories: change in human expertise and professional behavior (color code light blue) and change in human cognition and private behavior (color code light green). The selection outside the EU contains 1 use case of the first type and 3 use cases of the second type, providing a complementary perspective to the EU selection.

10.4 Relevance of use cases to cognitive and behavioral domains

#	Use case description		Decision making	Human creativity	Social and personal relations	Education and training	Cognitive offloading and skill loss	Nudging and manipulation
1	Change in human expertise and professional behaviour	Medical doctors (radiologists, surgeons) using AI tools in diagnostics and treatment	High	Medium	Low	Medium	High	Medium
2		Safety engineers using AI tools to speed up software release approvals	High	Low	Low	Medium	High	Medium
3		Recruiters using AI tools in hiring processes	High	High	Medium	Low	Medium	High
4		Security professionals using AI tools to detect hate speech	High	Medium	Low	High	High	Medium
7		Workplaces equipped with AI tools for behavioral analysis	Medium	Low	Medium	High	Medium	High
5	Change in human cognition and private behavior	AI systems as individual and family-level virtual assistants	Medium	High	High	Medium	High	High
6		Deepfake therapy for processing grief or trauma	Low	High	High	Low	Medium	High
8		AI systems for smart elderly care in Wuxi	Low	Low	High	Medium	High	High
9		AI systems as 'personal companions' to assist senior citizens	Low	Low	High	Medium	High	High
10		AI systems as grief-alleviating personal assistants	Low	High	High	Medium	High	High

This table illustrates the connection between the use cases and cognitive domains reviewed in this deliverable. The shading helps understand both relevance and complementarity of the factors used in AIOLIA selection. Color codes are the same as in previous tables in this section.

10.5 Conclusion

Tables presented in this section help to visualize the selection of AIOLIA use cases and AI research areas discussed and motivated in this deliverable. They also help to connect this selection with cognitive and behavioral domains reviewed earlier. The selection is based on three main criteria defined in the Grant Agreement and provides a rich and complementary perspective on the interaction between AI and human cognition and behavior. This selection forms a foundation for all future work in AIOLIA.