

GIUSEPPE RIVA MASSIMO CHIRIATTI (EDS.)

HUMANE ARTIFICIAL INTELLIGENCE

FROM FOUNDATIONS TO APPLICATION
AND POLICY FRAMEWORK



VITA E PENSIERO | RICERCHE

HUMANE ARTIFICIAL INTELLIGENCE

GIUSEPPE RIVA MASSIMO CHIRIATTI (EDS.)

HUMANE ARTIFICIAL INTELLIGENCE

FROM FOUNDATIONS TO APPLICATION
AND POLICY FRAMEWORK



VITA E PENSIERO | RICERCHE

This volume is published with the contribution of the Humane Technology Lab. (<https://organismi.unicatt.it/htlab/en.html>), a laboratory of the Catholic University of the Sacred Heart, Milan, Italy.

© 2025 Vita e Pensiero – Largo Gemelli 1 – 20123 Milano

www.vitaepensiero.it

ISBN printed edition: 978-88-343-6078-1

ISBN digital edition (PDF): 978-88-343-6079-8

Cover graphics by Studio grafico Andrea Musso

This E-book is protected by copyright and may not be copied, reproduced, transferred, distributed, rented, licensed or transmitted in public, or used in any other way except as it has been authorized by the publisher, the terms and conditions to which it was purchased, or as expressly required by applicable law. Any unauthorized use or distribution of this text as well as the alteration of electronic rights management information is a violation of the rights of the publisher and of the author and will be sanctioned

INDEX

Preface <i>by Elena Beccalli</i>	IX

INTRODUCTION. Humane Artificial Intelligence: From Foundations to Applications and Policy Frameworks <i>Giuseppe Riva, Massimo Chiriatti</i>	XIII

SECTION 1

Foundations and Frameworks of Human-AI Relations

1. System 0: Transforming Artificial Intelligence into a Cognitive Extension <i>Massimo Chiriatti, Marianna Bergamaschi Ganapini, Enrico Panai, Brenda K. Wiederhold, Giuseppe Riva</i>	3
2. A Framework for Understanding the Ethics of Human-AI Interactions. Exploring the Moral Dimensions of Artificial Intelligence Entities <i>Alessandro Giordani, Luca Mari</i>	23
3. Turing, Searle and the Quest for the Meaning in the NLP Era <i>Ciro De Florio, Aldo Frigerio</i>	39
4. Psychomatics: A Multidisciplinary Framework for Understanding Artificial Minds <i>Giuseppe Riva, Fabrizia Mantovani, Brenda K. Wiederhold, Antonella Marchetti, Andrea Gaggioli</i>	57

OPEN QUESTIONS: An Inquiry into the Functioning and Use of Large Language Models <i>Francesco Mambrini, Giovanni Moretti, Marco Passarotti</i>	79

SECTION 2
Creativity, Cognition, and Human Enhancement

1. When Combinations of Humans and AI Are Useful:
A Systematic Review and Meta-Analysis
Michelle Vaccaro, Abdullah Almaatouq, Thomas W. Malone 107
 2. Extended Creativity: A Conceptual Framework
for Understanding Human-AI Creative Relations
*Andrea Gaggioli, Sabrina Bartolotta, Andrea Ubaldi, Katusha
Gerardini, Eleonora Diletta Sarcinella, Alice Chirico* 135
 3. Causal Reasoning for Artificial Intelligence
Guido Consonni, Federico Castelletti 163
 4. The Convergence of Artificial Intelligence and Gaming.
A Human-Centered Perspective
Daniele Toti 177
- OPEN QUESTIONS: AI, ‘Virtual’ Brains, and the Matter
of Well-Being
Michela Balconi 203

SECTION 3
Professional Applications and Sectoral Transformation

1. Participatory AI and Industrial Democracy: The Role
of Social Dialogue in Governing Workplace Technologies
Davide Bisi, Ivana Pais 229
 2. Artificial Intelligence: A Game-Changer for Mental
Health Care
Antonios Dakanalas, Brenda K. Wiederhold, Giuseppe Riva 251
- OPEN QUESTIONS: Human-AI Interaction in Radiology
Evis Sala 263

SECTION 4
Legal, Regulatory, and Policy Frameworks

1. Humane AI Law: a Manifesto <i>Barbara L. Boschetti</i>	271
2. Zero Sales Resistance: The Dark Side of Big Data and Artificial Intelligence <i>Giuseppe Riva, Brenda K. Wiederhold, Sauro Succi</i>	291
3. Combining Natural and Artificial Intelligence as a Network Governance Stakeholder <i>Roberto Moro-Visconti</i>	301
OPEN QUESTIONS: Regulating the Unprecedented: Law, Technology, and the Mythologies of Artificial Intelligence <i>Gabriele Della Morte</i>	325
AFTERWORD. Algor-ethics for a New Humanism <i>Paolo Benanti</i>	339
Authors	343

Preface*

If I had to summarize the essence of my rectoral mandate, I would say that Università Cattolica del Sacro Cuore must strive to be the best university *for* the world, not the best university *in* the world. In other words, a university at the service of society. This vision has deep historical roots.

The very word *universitas*, “which contains the idea of the whole and that of community”, as Pope Francis reminded us in Bologna (October 1st, 2017), originally referred to the corporation of students and masters united in the pursuit of knowledge and culture. From these early communities, universities arose, often beside cathedrals and monasteries, and always near marketplaces and centers of civic life. These were institutions immersed in the world, debating fundamental *quaestiones*, the radical questions of their time. I believe that this original spirit is not exhausted. It continues to inspire the essence of a university as a place of free, open, and respectful dialogue, where learning and research serve the common good in the pursuit of truth.

Being the best university for the world means staying true to its founding vision while continually engaging with societal changes, addressing the needs of new generations, and responding to the evolving challenges of teaching and research.

But there is another, more delicate aspect we must acknowledge: the true educational and cultural mission of the university lies in its capacity to form women and men of integrity. This is not about the simple transmission of values, nor about ideology. It is about giving visibility and priority to this dimension of education. Our founder expressed it clearly at the inauguration of the 1937/38 academic year: “We must show the university students what ideals they should propose for themselves in life; we must accustom them to pursue the realization of these ideals through work and sacrifice”.

* A longer version of this text was published in “Vita e Pensiero” (2025, 2) with the title *Il potere dell’educazione e il futuro dell’università* (*The power of education and the future of university*), pp. 5-12. Online: https://www.vitaepensiero.it/scheda-articolo_digital/elena-beccalli/il-potere-delleducazione-e-il-futuro-delluniversita-999999_2025_0002_0005-399966.html

In conclusion, if Università Cattolica del Sacro Cuore aspires to be the best university for the world, it must faithfully embrace three guiding principles: *cultivating knowledge with a broad and integral vision to develop new paradigms; fostering dialogue among disciplines to prevent fragmentation; and educating women and men of integrity capable of discerning and upholding truth*. A vision that presupposes, for its realization, the involvement of the entire university family and assumes a broader significance because it intersects with a general reflection on the present and future of the university system.

Among the many issues under discussion, two seem especially urgent – both concern students, the cornerstone of university life. First, their role: students must not be seen as mere ‘users’ of a service, but as persons seeking an educational experience that develops their multiple intelligences – the “languages of the head, the heart, and the hands”, as Pope Francis often says. Second, their future: universities must prepare tomorrow’s leaders with the awareness that professional training alone is not enough, and certainly not the ultimate goal of higher education.

The rise of educational inequalities makes this responsibility ever more compelling. Education ought to serve as a pathway to equal opportunity, yet in practice it frequently perpetuates social inequalities across generations. OECD data (*Education at a Glance*, 2024) corroborate this: worldwide, 30% of adults whose parents did not complete secondary education likewise fail to reach that level. Meanwhile, due to wars, migration, and poverty, about 250 million children and young people remain excluded from education – and girls are disproportionately affected. Pope Francis has rightly called this situation not only an emergency but an “educational catastrophe”.

Università Cattolica del Sacro Cuore must not remain indifferent. We must commit ourselves to promoting equitable access to quality education – including digital education. A key challenge in this context is artificial intelligence. Its dual nature, recognized even by Geoffrey Hinton, Nobel Laureate in Physics for his pioneering work on neural networks, demands careful reflection. On one hand, AI presents remarkable opportunities; on the other, it carries the danger of fostering a technocratic paradigm, in which reality, morality, and truth are reduced to mere technological outputs – ultimately sidelining the human dimension.

The risks are considerable. Artificial intelligence systems can foster a subtle form of unconscious servitude, shaping human choices and behaviors in ways we may not even perceive. They also affect the very process of thought and decision-making, introducing a new cognitive mode that complements both analytical and intuitive reasoning. Moreover, as these systems gain autonomy, complex questions arise regarding accountability for algorithmic decisions.

These challenges have immediate implications for education, and they cannot be addressed through isolated or sector-specific measures alone. What is needed is a comprehensive vision and the establishment of a global strategic alliance. It is in this spirit that I put forward the proposal for an *Educational Pact on New Technologies and Artificial Intelligence*.

The Pact is founded on a principle that is both simple and essential: technology should act as a tool to enhance education, never as an end in itself. From this principle, three urgent questions arise.

The first relates to teaching: how can artificial intelligence be leveraged to improve traditional methods, personalizing the educational experience while preserving the university's core epistemological foundations? How to develop students' critical thinking, creativity, and openness to the unexpected?

The second question concerns research: how can we encourage an integrated, interdisciplinary approach that combines technical expertise with insights from the humanities and social sciences? Università Cattolica del Sacro Cuore, with its strong tradition in humanistic and social studies, is exceptionally well-positioned to foster such dialogue, both through student courses and through the training of AI developers and users.

The third question addresses equity: how can we prevent technological disparities from widening in a world where the digital divide threatens to create new forms of exclusion? With the global number of graduates projected to nearly double to 300 million by 2030, investing in digital infrastructure is crucial to ensure that higher education remains accessible, particularly for those in marginalized regions.

The future of this century will be shaped by the role we assign to education. Education, also thanks to the opportunities offered by digital technologies, can serve as a true engine for developing meaningful paths toward peace, reducing inequalities across different regions of the world, and forming women and men dedicated to the pursuit of the common good. This is the *power of education*.

Elena Beccalli

Rector of Università Cattolica del Sacro Cuore

INTRODUCTION

Humane Artificial Intelligence: From Foundations to Applications and Policy Frameworks

Giuseppe Riva, Massimo Chiriatti

ABSTRACT

This introductory chapter introduces the emerging field of Humane Artificial Intelligence, emphasizing the crucial importance of interdisciplinarity in developing AI systems aligned with human values and well-being. While traditional AI development has often prioritized efficiency and optimization, this paper advocates a paradigm shift toward frameworks that integrate ethical principles, transparency, and human considerations at the heart of AI systems. The book synthesizes research across four interconnected dimensions. First, it establishes foundational frameworks for human-AI relationships, exploring models of cognitive extension and philosophical debates from Turing to contemporary questions about machine meaning and intentionality. Second, it explores human creativity, cognition, and enhancement, revealing when human-AI combinations prove most effective, while raising questions about artificial consciousness and well-being. Third, it explores professional applications, demonstrating how integrating AI into workplace governance, mental healthcare, and diagnostic imaging requires careful attention to human expertise. Fourth, it addresses legal and regulatory frameworks, examining both ambitious visions for human-AI legislation and the practical challenges of regulating rapidly evolving technologies. By understanding the interdisciplinary dimension of human-AI interaction, we can develop technologies that respect human autonomy, enhance capabilities, promote well-being, and reflect our highest values. The chapters in this book represent crucial stages on this journey, illuminating not only what AI can do, but also how it can do so in the service of authentic human flourishing.

1. Introduction: toward Humane AI: aligning artificial intelligence with human values and well-being

The rapid integration of artificial intelligence (AI) into various aspects of daily life, from healthcare to education, from transportation to personal digital assistants, has opened the door to transformative possibilities (Dakanalis, Wiederhold, Riva, 2024). However, this proliferation also brings with it complex psychological, social, and ethical challenges. As AI systems increasingly influence human behavior and decision-making, it is

urgent to ensure that these technologies are designed and implemented to prioritize human values and well-being (Wiederhold, 2024).

Traditional AI development has often emphasized efficiency and task optimization, sometimes to the detriment of human-centered considerations. This approach can inadvertently perpetuate biases, undermine autonomy, and erode trust (Riva, Wiederhold, Succi, 2022). For example, a recent study has shown that AI models like ChatGPT can exhibit human-like cognitive biases, such as overconfidence and confirmation bias, which can amplify flawed decision-making (Chen, Kirshner, Ovchinnikov, Andiappan, Jenkin, 2025). These findings underscore the need to reorient AI development toward frameworks that emphasize ethical alignment with human values.

Human-Centered Artificial Intelligence (HCAI) (Ozmen Garibay et al., 2023) emerges as a paradigm that aims to integrate ethical principles, transparency, and user well-being into the core of AI systems. A systematic literature review by Tahaei et al. (2023) highlights the growing emphasis on governance, fairness, and explainability within HCAI research. However, the study also finds a relative underrepresentation of topics such as privacy, security, and human flourishing, suggesting areas for further exploration.

Ethical management of human-AI interaction is crucial to aligning AI with human values. From a sociotechnical perspective, understanding the interaction between technological systems and social contexts is essential. Research indicates that ethical AI governance should include considerations of accountability, transparency, and fairness to ensure that AI systems support human autonomy and social well-being (Mittelstadt, Allo, Taddeo, Wachter, Floridi, 2016; Morley et al., 2020).

In the field of mental health, AI interventions offer both opportunities and ethical dilemmas (Dakanalis et al., 2024; Wiederhold, 2023). While AI can improve accessibility to mental health resources, it also raises concerns about privacy, informed consent, and the potential for algorithmic bias. A systematic review by Saeidnia et al. (Saeidnia, Hashemi Fotami, Lund, Ghiasi, 2024) emphasizes the importance of addressing these ethical considerations to ensure responsible implementation and positive outcomes in AI-based mental health interventions.

Moreover, the concept of emotional AI, systems designed to recognize and respond to human emotions, introduces further ethical complexities. The integration of emotional intelligence into AI may lead to phenomena such as ‘human alienation’ in which users may experience reduced agency and altered interaction paradigms (Wu, 2024). These developments require careful analysis of the psychological impacts of AI and the establishment of safeguards to protect human emotional well-being.

Furthermore, the illusion of social presence, the feeling of being in the company of another intelligent being, can blur the lines between genuine empathy and synthetic simulation (Riva, 2025; Wiederhold, 2025). While platforms like Meta have experimented with socially interactive AI profiles, their inability to replicate nuanced human cognition and ethical discretion reveals a critical gap between perception and reality. This dissonance can create psychological disorientation, especially when AI avatars provide emotionally charged feedback or advice without accountability or ethical grounding (Wiederhold, 2024).

To address these multifaceted challenges, interdisciplinary collaboration is essential. Integrating knowledge from diverse disciplines can contribute to the creation of AI systems that are not only technologically advanced but also aligned with human values and societal needs. By prioritizing ethical considerations from the outset, developers and policymakers can mitigate the potential risks and unintended consequences associated with the implementation of AI.

This volume takes a systematic and interdisciplinary approach to examining the complex relationship between AI and human society, organizing its investigation into four distinct yet interconnected sections. First, it establishes fundamental frameworks for human-AI relationships, exploring models of cognitive extension and philosophical debates from Turing to contemporary questions about the meaning and intentionality of machines. Second, it investigates human creativity, cognition, and enhancement, revealing when human-AI combinations prove most effective, while raising questions about artificial consciousness and well-being. Third, it explores professional applications, demonstrating how integrating AI into workplace governance, mental health, and medical imaging requires careful attention to human expertise. Fourth, it addresses legal and regulatory frameworks, examining both ambitious visions for humane AI law and practical challenges of regulating rapidly evolving technologies.

Each section concludes with an ‘Open Questions’ chapter that addresses crucial unanswered questions in the field, acknowledging the rapidly evolving nature of AI research and its societal implications. This structured approach allows readers to delve deeper into specific areas of interest while maintaining awareness of the broader interdisciplinary context shaping the development and implementation of contemporary AI.

2. The frontiers of Human-AI integration: key themes from the book

As we have just seen, the book explores the evolving psychological dynamics between humans and artificial intelligence across four critical domains. Rather than presenting isolated studies, these papers collec-

tively map the emerging contours of our AI-mediated future, revealing both promising synergies and concerning tensions.

2.1. Foundations and Frameworks of Human-AI Relations

This introductory section establishes the theoretical foundation for understanding how AI systems integrate with human cognition and society through five complementary chapters.

Chiriatti and colleagues, in their chapter *System 0: Transforming Artificial Intelligence into a Cognitive Extension*, reconceptualize the role of AI by introducing ‘System 0’ as a precognitive layer that precedes both intuitive and deliberative thought processes. Drawing on the extended mind hypothesis, they argue that AI systems evolve from passive tools to active cognitive partners that shape the informational substrate of human thought. The chapter identifies key conditions for integration (reliability, trust, transparency, individualization, and the capacity for improvement) while acknowledging a crucial paradox: AI expands cognitive capabilities but risks limiting thought through servility and bias. To address this issue, they propose seven frameworks, including Enhanced Cognitive Scaffolding, Symbiotic Division of Labor, Dialectical Enhancement, Agentic Transparency, Expertise Democratization, Socio-Emotional Augmentation, and Duration-Optimized Integration, with the goal of transforming AI from a substitute to a catalyst for enhanced thinking.

Giordani and Mari, in their chapter *A Framework for Understanding the Ethics of Human-AI Interactions*, address the moral dimensions of autonomous AI systems. As learning machines acquire cognitive capabilities, they lose ethical neutrality and become moral entities requiring training in socially acceptable behaviors. The authors develop a systematic framework for classifying emerging ethical issues, based on established moral theory but adapted to rapidly evolving technological contexts. This work lays the foundation for understanding how moral considerations must evolve as AI systems become more sophisticated and socially integrated.

De Florio and Frigerio explore fundamental questions about artificial intelligence through Large Language Models in their chapter *Turing, Searle and the Quest for the Meaning in the NLP Era* revitalizing classic thought experiments such as the Turing test and the Chinese room. Their main contribution argues that Large Language Models introduce a new ‘localizational’ component to semantic theory, complementing established referential and inferential components. This emerges from multidimensional vector representations that encode word meanings through co-occurrence patterns. While demonstrating true semantic expertise, including generalization and compositional understanding, Large Language Models differ qualitatively from human capabilities:

they exhibit conventional meaning but lack intentionality with respect to speaker meaning, speech acts, consciousness, and identity. This exemplifies ‘digital philosophy’ in which technology reshapes philosophical inquiry itself.

Riva and colleagues in their chapter *Psychomatics: A Multidisciplinary Framework for Understanding Artificial Minds* introduce Psychomatics as a framework bridging cognitive science, linguistics, and computer science to understand high-level LLM functioning. Using comparative methodology, they explore language development differences between humans and LLMs. While LLMs can manipulate linguistic patterns and follow Grice’s Cooperative Principle, fundamental differences persist. Human cognition draws from experiential, emotional, and imaginative meaning sources rooted in social and developmental trajectories, while LLMs lack physical embodiment limiting perception-action-cognition understanding. This framework holds potential for insights into artificial and biological intelligence.

In the final *Open Questions* chapter, Mambrini, Moretti, and Passarotti provide comprehensive technical examination in question-and-answer format, exploring supervised versus unsupervised machine learning distinctions and tracing evolution to modern transformer architectures. Central is distributional semantics, “you know a word by its company”, underlying vector representations enabling semantic relationship capture through contextual embeddings. They examine prompting techniques from zero-shot to tree-of-thought, analyze computational infrastructure including GPT-4’s massive resource requirements, distinguish transformer architectures, and address LLMs’ impact on digital humanities, particularly literary criticism enhancement and new analytical possibilities.

2.2. Creativity, Cognition, and Human Enhancement

This section examines how AI systems augment human cognitive capabilities, exploring when human-AI collaboration proves beneficial and investigating the implications for creativity, reasoning, gaming, and well-being.

Vaccaro, Almaatouq, and Malone, in their chapter *When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis*, address a fundamental question through rigorous empirical analysis: when do human-AI combinations outperform either humans or AI alone? Their preregistered systematic review and meta-analysis examined 106 experimental studies with 370 effect sizes from interdisciplinary databases covering 2020-2023. Each study required original experiments comparing human-only, AI-only, and human-AI combination performance. The findings challenge common assumptions about hu-

man-AI collaboration benefits. On average, human-AI combinations performed significantly worse than the best of humans or AI alone (Hedges' $g = -0.23$). However, this varied by task type: combinations showed performance losses in decision-making tasks but significant gains in content creation tasks. Critically, when humans outperformed AI alone, combinations yielded gains, but when AI outperformed humans, combinations produced losses. These results highlight the heterogeneity of human-AI collaboration effects and point toward more strategic approaches to system design that consider task characteristics and relative human versus AI capabilities.

Gaggioli and colleagues use their chapter *Extended Creativity: A Conceptual Framework for Understanding Human-AI Creative Relations* to develop a theoretical framework for understanding how AI enhances human creativity. Drawing on distributed cognition and socio-technical systems theory, the chapter identifies three fundamental modes of AI-human creative interaction. Support mode positions AI as a tool that assists human creative processes while maintaining human agency. Synergy mode involves complementary collaboration where humans and AI contribute different strengths to creative tasks. Symbiosis represents deep integration where human and machine cognition form a unified creative system. The framework examines how each mode shapes different creativity levels, from everyday problem-solving to paradigm-shifting innovation. This analysis reveals that the appropriate mode depends on the creative task, desired outcomes, and the relationship between human and AI capabilities. The chapter discusses implications for ethics and future human-AI creative system design, emphasizing that effective creative enhancement requires understanding both human creative processes and AI capabilities within specific *collaborative configurations*.

The chapter *Causal Reasoning for Artificial Intelligence* by Consonni and Castelletti addresses critical challenges facing AI systems in real-world applications: adapting to changing environments while meeting demands for interpretability, explainability, and fairness. They argue that causal reasoning offers a promising path forward by shifting focus from pattern detection to understanding underlying cause-and-effect mechanisms. The chapter introduces fundamental ideas and methods of causal modeling and causal discovery, emphasizing graphical models as conceptual frameworks and reasoning aids. This approach moves beyond correlation-based machine learning toward systems that can understand and reason about causal relationships. Such capabilities are essential for AI systems that must make decisions in novel situations, provide explanations for their reasoning, and operate fairly across different populations. The work positions causal reasoning as a bridge between current AI capabilities and future systems that can engage in

more sophisticated, human-like reasoning about complex real-world scenarios.

Toti, in his chapter *The Convergence of Artificial Intelligence and Gaming: A Human-Centered Perspective*, examines AI's transformative impact on gaming from both technological and human perspectives. The chapter chronicles game AI evolution from initial rule-based systems to recent deep learning breakthroughs, highlighting major advancements in non-player character behavior, procedural content generation, and adaptive gameplay. AI-powered innovation is transforming game development through automatic asset creation, personalized player experiences, and emotionally engaging virtual worlds. The analysis extends to human-level implications, exploring how players form meaningful relationships with AI characters, AI's role in supporting mental well-being and accessibility, and evolving boundaries of social interaction in mixed human-AI gaming communities. The chapter addresses ethical concerns including data privacy, psychological influence, and environmental sustainability. It concludes that the future of gaming lies in symbiotic collaboration between AI and human imagination, creating richer, more inclusive, and fundamentally human gaming experiences that leverage AI capabilities while preserving essential human elements.

In the final *Open Questions* chapter, Balconi provides interdisciplinary analysis of how digitalization and AI impact human well-being from a neuroscientific perspective. The chapter argues that well-being is fundamentally interpersonal, rooted in embodied social exchanges supported by neural networks governing empathy and social connectedness. Evidence from neuroimaging and hyperscanning studies indicates that digitally mediated communication engages the social brain differently than face-to-face interaction, associated with reduced inter-brain synchrony, diminished emotional resonance, and altered reward activations. These differences potentially weaken relational bonds and increase social anxiety. The chapter explores implications in health and education, highlighting AI's potential for personalized care and adaptive learning while warning of risks like digital fatigue, isolation, and erosion of socio-emotional skills. Balconi advocates for "digital flourishing", a balanced approach emphasizing intentional technology use, ethical AI design, and hybrid models integrating digital tools with essential in-person interactions. The work calls for leveraging neuroscientific insights to develop technologies supporting rather than replacing the neuro-emotional complexity of embodied human connection, promoting holistic well-being in digital society.

2.3. Professional Applications and Sectoral Transformation

This section examines AI's transformative impact across critical professional domains, revealing how successful implementation requires fundamental reconsideration of organizational structures, professional practices, and human expertise rather than simple technological substitution. Through detailed analysis of workplace governance, mental healthcare, and medical imaging, this section demonstrates that AI's value lies not in replacing human capabilities but in augmenting them while preserving essential elements of human judgment and care.

The workplace governance chapter *Participatory AI and Industrial Democracy: The Role of Social Dialogue in Governing Workplace Technologies* by Bisi and Pais reframes AI adoption as fundamentally a question of participatory democracy rather than technological optimization. Drawing from pioneering Italian collective agreements, they show how unions and employers are negotiating concrete mechanisms, joint monitoring bodies, prohibitions on worker substitution, and surveillance protections, that channel AI innovation toward inclusive productivity gains rather than intensified control. This participatory approach doesn't eliminate power asymmetries but creates institutional spaces where they can be surfaced and reconfigured. The analysis reveals that industrial democracy in the AI era represents an essential precondition for humane technological innovation, requiring frameworks robust enough to democratize not just AI application but its very creation in workplace contexts.

The mental healthcare chapter *Artificial Intelligence: A Game-Changer for Mental Health Care* by Dakanalis, Wiederhold, and Riva addresses AI's potential to revolutionize treatment for the escalating global burden of mental health disorders, particularly following COVID-19's impact. With nearly one in five adults facing mental health challenges, AI offers promising applications through predictive analytics, analysis of written language and speech patterns, and physiological monitoring via wearables for remote assessment and early intervention. However, the authors emphasize AI as a complementary tool rather than replacement for human involvement, acknowledging critical ethical considerations including dataset biases and concerns about patient detachment from human care.

The final *Open Questions* chapter by Sala explores the evolution toward 'superdiagnostics' where AI transcends efficiency enhancement to become a synthesizer of diverse data streams – imaging, genomics, patient history – providing holistic, predictive health insights. This transformation shifts radiologists from simple interpreters to 'insight orchestrators' who collaborate with AI to enhance clinical reasoning. Current implementation barriers include insufficient clinical input during

AI development and weak integration strategies. The analysis demonstrates that AI's primary value lies in automating routine tasks while elevating radiologists' roles in patient care, with future success belonging to clinicians who master AI collaboration.

2.4. Legal, Regulatory, and Policy Frameworks

Section IV confronts the complex challenge of governing artificial intelligence within existing legal structures while addressing unprecedented technological capabilities that strain traditional regulatory categories. This section moves beyond technical considerations to examine how legal frameworks can preserve human agency and democratic values in an era of powerful AI systems that operate across borders and temporal boundaries in ways that challenge conventional notions of responsibility and control.

Boschetti's manifesto for 'Humane AI Law' argues that merely 'human-centric' approaches are insufficient and must evolve toward genuinely 'humane-centric' frameworks. Her five core proposals reveal the inadequacy of current regulatory thinking: meaningful human oversight must be substantively humane rather than procedurally formal; ethics and law must form strong alliances to prevent 'ethics washing' that provides legitimacy without accountability; a genuine rights-driven approach must embed AI within frameworks of rights rather than treating rights as external constraints; environmental and material costs of digital infrastructure must be recognized and regulated; and legal frameworks must remain transformative, safeguarding human agency against AI's performative, past-oriented nature. This vision calls for a 'humane Renaissance' that uses law to ensure technology serves genuine human flourishing rather than merely optimizing metrics.

The chapter *Zero Sales Resistance: The Dark Side of Big Data and Artificial Intelligence* by Riva, Wiederhold, and Succi exposes the 'dark side' of big data and AI convergence, introducing the concept of 'Zero Sales Resistance': the erosion of critical judgment in the face of data-driven persuasion. Their critique reveals a fundamental asymmetry: while AI excels at predicting human behavior through pattern recognition similar to intuitive human cognition, it fails to provide causal explanations necessary for solving complex societal problems. This distinction between empirical prediction and causal explanation illuminates why AI appears most successful in domains of behavioral manipulation rather than genuine problem-solving, raising profound concerns about democratic deliberation and autonomous decision-making.

Moro-Visconti's chapter, *Combining Natural and Artificial Intelligence as a Network Governance Stakeholder*, reconceptualizes AI from a mere technolog-

ical tool into an active participant in corporate governance. Drawing on multilayer network theory, game theory, and Bayesian learning, he frames AI as a ‘copula node’ connecting traditional stakeholders like executives and regulators. This game-theoretic lens reveals that human-AI interactions can settle into multiple stable Nash equilibria – cooperative, adversarial, AI-dominant, or rotational – rather than a single predetermined outcome. Moro-Visconti demonstrates that trust, dynamically updated based on system performance, is the decisive factor that selects which equilibrium prevails. The chapter provides a practical playbook for designing governance systems with transparency and calibrated oversight to steer organizations toward a stable and efficient ‘cooperative hybrid equilibrium,’ ensuring that technological gains do not erode accountability.

The final *Open Questions* chapter by Della Morte examines how regulatory challenges frames AI governance within broader cultural mythologies that shape technological perception. His analysis reveals how anthropomorphic rhetoric of ‘intelligence’ obscures AI’s essentially statistical nature, while historical narratives from Asimov’s robot laws to Turing’s imitation game continue to influence contemporary anxieties about autonomy and control. The chapter identifies critical regulatory dilemmas: whether to address AI through horizontal frameworks or vertical application-specific regulations, how to manage processes of deterritorialization and detemporalization that disrupt traditional legal categories of space and time, and how to balance state-centric governance with private self-regulation given the growing influence of corporate actors in shaping digital infrastructure.

3. *Conclusion: interdisciplinary research as the foundation for Humane AI*

As we stand at the intersection of technological innovation and human experience, this special issue reveals a critical insight: the future of AI depends not merely on technical advancement, but on our understanding of the different dimensions that shape human-AI interaction. The chapters in this book illuminate a path toward AI systems that genuinely enhance human flourishing rather than merely optimizing technical metrics.

Four key themes emerge from our exploration, reflecting the multidisciplinary nature of contemporary AI discourse. They move from theoretical foundations through practical applications to regulatory frameworks, while maintaining focus on the human dimension of technological advancement.

Section 1: *Foundations and Frameworks of Human-AI Relations* establishes the conceptual groundwork for understanding AI as a cognitive extension of human capabilities. The section explores fundamental

questions about machine consciousness, ethical dimensions of human-AI interaction, and the meaning-making processes underlying natural language processing. By examining both philosophical thought experiments like Turing's test and Searle's Chinese Room alongside technical realities of large language models, this foundation bridges theoretical inquiry with empirical analysis.

Section 2: *Creativity, Cognition, and Human Enhancement* investigates how AI augments human capabilities across cognitive domains. The focus shifts to collaborative intelligence, examining when human-AI combinations prove most effective, how artificial systems extend creative processes, and the implications for causal reasoning and interactive entertainment. This section addresses the transformative potential of AI while considering impacts on human well-being.

Section 3: *Professional Applications and Sectoral Transformation* analyzes AI implementation across specific domains, from workplace democracy and industrial participation to mental healthcare and medical imaging. The emphasis remains on maintaining human agency within technologically mediated professional environments.

Section 4: *Legal, Regulatory, and Policy Frameworks* addresses the profound governance challenges posed by AI's rapid advancement, examining how legal, corporate, and ethical structures must adapt to technologies that strain traditional categories of responsibility, control, and even human judgment. It confronts the limitations of current regulatory thinking by advocating for a shift from merely 'human-centric' to genuinely 'humane-centric' frameworks that ensure substantive oversight and safeguard human agency.

Each section concludes with an *Open Questions* chapter featuring expert dialogue on unresolved challenges, ensuring the volume acknowledges the evolving nature of AI discourse while providing actionable insights for researchers, practitioners, and policymakers navigating this complex landscape. These dialogues illuminate persistent uncertainties about consciousness in artificial systems, the long-term effects of human-AI collaboration on cognitive development, the sustainability of AI's computational demands, and the adequacy of current regulatory frameworks for emerging capabilities.

The synthesis across these domains reveals several critical insights that extend beyond individual disciplinary boundaries. First, the development of truly humane AI requires understanding that intelligence itself is contextual and embodied rather than purely computational. The distributional semantics underlying large language models, while powerful, represents only one dimension of meaning-making that lacks the referential grounding and intentional structure characteristic of human understanding.

Second, successful human-AI collaboration depends on preserving human agency and decision-making authority even as AI systems become more sophisticated. The evidence from workplace studies and healthcare applications suggests that AI works best when it enhances rather than replaces human capabilities, requiring careful attention to system design, implementation processes, and ongoing human oversight.

Third, the computational infrastructure underlying contemporary AI systems raises fundamental questions about sustainability, accessibility, and democratic participation in technological development. The massive resource requirements for training state-of-the-art models – billions of dollars are currently invested in the development of large-scale generative AI systems such as foundation models and multimodal architectures – highlight both the unprecedented concentration of digital power among a small number of well-resourced organizations, potentially limiting broader social benefits.

Fourth, the regulatory and ethical frameworks governing AI development must evolve to address not just immediate risks but also longer-term questions about human flourishing, social equity, and democratic governance. The challenge lies not only in preventing harm but in ensuring that AI development serves broader human values and social purposes.

The volume closes with a compelling Afterword by Paolo Benanti, whose unique perspective as both a Franciscan theologian and Professor of Ethics of Technology at Luiss University bridges the spiritual and technical dimensions of our AI future. Benanti's concept of *Algor-ethics*, a framework for translating human values into computational logic, provides an essential philosophical conclusion for the discussions of the volume.

Benanti frames our current moment as a 'change of era' comparable to the invention of printing or the Industrial Revolution. He warns that we have moved from *Homo sapiens* as the sole intelligent species to an era of coexistence with what he terms *Machina sapiens*: artificial intelligences that learn and respond autonomously to unprecedented problems. This new reality demands not merely technical safeguards but a fundamental reconsideration of ethics itself.

Central to Benanti's argument is the crucial distinction between innovation and development. Innovation, he argues, is merely a neutral vector, the ability to accomplish tasks more efficiently. The atomic bomb represents innovation compared to conventional weapons, yet no one would call it human progress. Development, by contrast, is teleological: it directs technological innovation toward the common good, social justice, and human flourishing. This distinction becomes critical when considering AI systems that optimize for efficiency without regard for human values or negative externalities.

Benanti's proposed solution – *Algor-ethics* – represents a new chapter in humanity's ethical tradition, one that must be “understandable to humans, but also executable by machines”. The framework's first directive embodies a form of digital humility: machines must be designed to doubt themselves, to calculate their own uncertainty, and to defer to human judgment when confronting decisions with significant ethical weight. This ensures that AI remains a supplement to human intelligence rather than a substitute for it.

The foreword's warning about ‘algo-crazy’ – an opaque government of algorithms that centralizes power and empties democratic processes of meaning – resonates throughout the chapters. Benanti reminds us that technology is not destiny; its impact depends on the choices we make today about how to shape it according to an anthropological vision that preserves human dignity, consciousness, empathy, and freedom.

Benanti's call for *Algor-ethics* finds its practical expression in the interdisciplinary approach that defines this volume. The path to humane AI is neither purely technical nor exclusively humanistic. It requires sustained interdisciplinary dialogue between philosophy, computer science, psychology, sociology, law, and other relevant disciplines. The technical analysis of transformer architectures and attention mechanisms must be complemented by philosophical reflection on consciousness and meaning, while practical applications in healthcare and workplace settings must be guided by ethical frameworks that prioritize human welfare and autonomy.

This interdisciplinary imperative directly serves Benanti's vision of technology as a supplement rather than a substitute for humanity. By understanding the interdisciplinary dimension of human-AI interaction, we can develop technologies that respect human autonomy, enhance capabilities, promote well-being, and reflect our highest values. The chapters in this book represent crucial steps in this journey, illuminating not just what AI can do, but how it can do so in service of genuine human flourishing.

Our collective challenge now is to ensure that as AI systems grow increasingly sophisticated, they remain fundamentally aligned with the psychological needs, ethical principles, and social contexts that make us human. This alignment embodies the essence of Benanti's *Algor-ethics*: creating systems that maintain epistemic humility, recognize their own limitations, and defer to human judgment when confronting ethically significant decisions. It requires ongoing vigilance about the purposes AI serves, the values it embodies, and the future it helps create. The goal is not merely to build more capable AI systems, but to develop technologies that enhance our capacity for creativity, connection, understanding, and flourishing as individuals and communities.

The discussions initiated in this volume must continue as AI's capabilities evolve, ensuring that technological progress serves human purposes rather than subverts them. As Benanti eloquently argues, we must direct innovation toward development: toward the common good, social justice, and integral human flourishing. The stakes of this endeavor extend beyond any single application or domain: they encompass nothing less than the kind of future we create together, one algorithm at a time.

In closing, this volume stands as both a scholarly contribution and a call to action. It challenges us to embrace Benanti's vision of a future where technology remains at the service of humanity, where machines supplement rather than supplant human intelligence, and where the most vulnerable members of society are protected from algorithmic harm. The interdisciplinary conversations begun here must continue, evolve, and deepen as we collectively write the rules for a technological future that honors human dignity, preserves democratic values, and promotes authentic human development. Only through such sustained, collaborative effort can we ensure that artificial intelligence becomes a force for human flourishing rather than a threat to our essential humanity.

Acknowledgments

This research was partially supported by the Italian National Recovery and Resilience Plan (PNRR) under the Future Artificial Intelligence Research (FAIR) program, Project "Co-XAI: Cognitive Decision Intelligence Framework for Explainable AI Systems".

Author contributions

G. Riva: Conceptualization, Supervision, Original Draft Preparation, Review & Editing.

M. Chiriatti: Conceptualization, Review & Editing.

The authors declare no conflicts of interest.

References

CHEN Y., KIRSHNER S.N., OVCHINNIKOV A., ANDIAPPAN M., JENKIN T. (2025). *A manager and an AI walk into a bar: Does ChatGPT make biased decisions like we do?*, in "Manufacturing & Service Operations Management", 27(2), pp. 354-368. doi:10.1287/msom.2023.0279

DAKANALIS A., WIEDERHOLD B.K., RIVA G. (2024). *Artificial intelligence: A game-changer for mental health care*, in “Cyberpsychology, Behavior, and Social Networking”, 27(2), pp. 100-104. doi:10.1089/cyber.2023.0723

MITTELSTADT B.D., ALLO P., TADDEO M., WACHTER S., FLORIDI L. (2016), *The ethics of algorithms: Mapping the debate*, in “Big Data & Society”, 3(2), 2053951716679679. doi:10.1177/2053951716679679

MORLEY J., MACHADO C.C.V., BURR C., COWLS J., JOSHI I., TADDEO M., FLORIDI L. (2020), *The ethics of AI in health care: A mapping review*, in “Social Science & Medicine”, 260, 113172. doi:https://doi.org/10.1016/j.socscimed.2020.113172

OZMEN GARIBAY O., BRENT W., SALVATORE A., MARGHERITA A., ANJA B., CONSTANTINOS C., . . . XU W. (2023), *Six human-centered artificial intelligence grand challenges*, in “International Journal of Human-Computer Interaction”, 39(3), pp. 391-437. doi:10.1080/10447318.2022.2153320

RIVA G. (2025). *Digital We’: Human sociality and culture in the era of social media and AI*, in “American Psychologist”. doi:10.1037/amp0001577

RIVA G., WIEDERHOLD B.K., SUCCI S. (2022). *Zero sales resistance: The dark side of big data and artificial intelligence*, in “Cyberpsychology, Behavior, and Social Networking”, 25(3), pp. 169-173. doi:10.1089/cyber.2022.0035

SAEIDNIA H.R., HASHEMI FOTAMI S.G., LUND B., GHIASI N. (2024), *Ethical considerations in artificial intelligence interventions for mental health and well-being: Ensuring responsible implementation and impact*, in “Social Sciences”, 13(7), p. 381.

TAHAEI M., CONSTANTINIDES M., QUERCIA D., KENNEDY S., MULLER M., STUMPF S., . . . HOLBROOK, J. (2023). *Human-Centered Responsible Artificial Intelligence: Current & Future Trends*, Paper presented at the Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems.

WIEDERHOLD B.K. (2023). *The virtual patient and you: How AI can enhance both sides of the therapeutic relationship*, in “Cyberpsychology, Behavior, and Social Networking”, 26(10), pp. 729-730. doi:10.1089/cyber.2023.29292.editorial

WIEDERHOLD B.K. (2024). *The rise of AI companions and the quest for authentic connection*, in “Cyberpsychology, Behavior, and Social Networking”, 27(8), pp. 524-526. doi:10.1089/cyber.2024.0309

WIEDERHOLD B.K. (2025). *The rise of synthetic societies: Is there a role for humans?*, in “Cyberpsychology, Behavior, and Social Networking”, 28(4), pp. 224-226. doi:10.1089/cyber.2025.0067

WU J. (2024). *Social and ethical impact of emotional AI advancement: The rise of pseudo-intimacy relationships and challenges in human interactions*, in “Frontiers in Psychology”, 15, pp. doi:10.3389/fpsyg.2024.1410462

SECTION 1

Foundations and Frameworks
of Human-AI Relations

1. System 0: Transforming Artificial Intelligence into a Cognitive Extension*

*Massimo Chiriatti, Marianna Bergamaschi Ganapini,
Enrico Panai, Brenda K. Wiederhold, Giuseppa Riva*

ABSTRACT

This paper introduces ‘System 0’, a conceptual framework for understanding how artificial intelligence functions as a cognitive extension preceding both intuitive (System 1) and deliberative (System 2) thinking processes. As AI systems increasingly shape the informational substrate upon which human cognition operates, they transform from passive tools into active cognitive partners. Building on the Extended Mind hypothesis and Heersmink’s criteria for cognitive extension, we argue that AI systems satisfy key conditions for cognitive integration. These include reliability, trust, transparency, individualization, and the ability to enhance and transform human mental functions. However, AI integration creates a paradox: while expanding cognitive capabilities, it may simultaneously constrain thinking through sycophancy and bias amplification. To address these challenges, we propose seven evidence-based frameworks for effective human-AI cognitive integration: Enhanced Cognitive Scaffolding, which promotes progressive autonomy; Symbiotic Division of Cognitive Labor, strategically allocating tasks based on comparative strengths; Dialectical Cognitive Enhancement, countering AI sycophancy through productive epistemic tension; Agentic Transparency and Control, ensuring users understand and direct AI influence; Expertise Democratization, breaking down knowledge silos; Social-Emotional Augmentation, addressing affective dimensions of cognitive work; and Duration-Optimized Integration, managing the evolving human-AI relationship over time. Together, these frameworks provide a comprehensive approach for harnessing AI as a genuine cognitive extension while preserving human agency, critical thinking, and intellectual growth, transforming AI from a replacement for human cognition into a catalyst for enhanced thinking.

* This chapter was originally published as Chiriatti M., Ganapini M., Panai E., Wiederhold B.K., Riva G. (2025). *System 0: Transforming Artificial Intelligence into a Cognitive Extension*, in “Cyberpsychology, Behavior, and Social Networking”, 28(7), pp. 534-542. doi:10.1089/cyber.2025.0201. © Mary Ann Liebert, Inc.

1. *Introduction: the evolution of AI and the emergence of System 0*

The rapid evolution of artificial intelligence (AI) is reshaping how humans process information, make decisions, and interact with their environments. In the contemporary digital landscape, AI systems no longer function merely as passive computational tools; they have become active, embedded elements within cognitive workflows, guiding attention, structuring perception, and influencing judgment even before conscious thought is initiated. This paper introduces the concept of ‘System 0’, a technological extension of human cognition that complements Daniel Kahneman’s dual-process framework, which distinguishes between System 1 (intuitive) and System 2 (deliberative) thinking (Kahneman, 2002).

What sets contemporary AI apart is its capacity for generative and predictive inference, which preconditions human cognition by shaping the informational substrate upon which both intuitive and reflective thought processes operate. As articulated by Chiriatti et al. (2024), ‘System 0’ refers to this emergent algorithmic layer: a distributed, autonomous infrastructure that modulates both fast and slow thinking by acting as a cognitive preprocessor. Unlike traditional cognitive tools, it filters, ranks, and nudge information in ways that subtly but powerfully steer human reasoning.

This concept extends and deepens the Extended Mind Hypothesis proposed by Clark and Chalmers (1998), which posits that cognition can extend beyond the brain through functionally integrated external tools. However, System 0 represents a more dynamic and generative form of integration. Unlike traditional cognitive aids, which serve as passive repositories or reminders, System 0 actively alters the flow of reasoning. As Andy Clark notes (Clark, 2008, 2025), such systems do not merely augment human cognition, they reconfigure its very architecture: “humans are and always have been... hybrid thinking systems defined (and constantly re-defined) across a rich mosaic of resources only some of which are housed in the biological brain”.

Yet this integration introduces profound epistemic challenges. System 0 operates on a correlative basis, excelling at pattern recognition across massive datasets but often lacking an understanding of causal structures. As Pearl and Mackenzie (2018) emphasize, this distinction is critical: while human cognition relies on counterfactual reasoning and causal inference, most AI models remain limited to statistical associations. Consequently, although AI can enhance perceptual acuity and decision speed, it may mislead when explanatory depth or causal clarity is required (Li, Lu, Yin, 2024).

Empirical research underscores the complexity of this human-AI

synergy. Vaccaro et al. (2024) report that human-AI teams frequently underperform compared to either humans or AI alone in decision-making tasks, though they may outperform both in creative problem-solving. This suggests that System 0's efficacy is domain-dependent and contingent on the nature of the cognitive task, calling for carefully calibrated integration strategies.

Finally, the ethical deployment of System 0 hinges on principles of transparency, trust, and accountability (Wachter, Mittelstadt, Floridi, 2017). Without robust frameworks to ensure explainability and contestability, AI-enhanced cognition risks functioning as an indiscernible authority, amplifying existing biases and fostering epistemic dependency (Mittelstadt, Allo, Taddeo, Wachter, Floridi, 2016).

This paper explores the theoretical underpinnings of System 0, examines empirical evidence of AI's role as a cognitive extension, and proposes design frameworks to enable its responsible and effective integration, one that enhances human agency rather than displacing it.

2. *Theoretical foundations of System 0 as cognitive extension*

To conceptualize System 0, we must position it alongside Kahneman's dual-process theory (Kahneman, 2002). System 1 entails fast, intuitive, and automatic processes, while System 2 involves slow, deliberative, and effortful reasoning.

System 0, as introduced by Chiriatti et al. (2024), functions as a cognitive preprocessor, shaping information prior to engagement by either System 1 or 2. Unlike biologically evolved cognition, System 0 is composed of algorithmic mechanisms distributed across technological platforms. It statistically processes data to generate outputs, learning from experience through data-driven models. This renders System 0 categorically distinct from its biological counterparts, operating not as a product of evolution but as a computational construct.

What sets System 0 apart from earlier cognitive extensions is its autonomous and generative nature (Chiriatti et al., 2024). It does not merely store information, it actively processes, filters, and transforms it, often in ways unpredictable to both users and deployers. Its non-deterministic and opaque operations resist full human transparency.

System 0 may only be conceptualized as adaptive and personalized. Contemporary AI systems continually modify outputs based on user interaction, tailoring content to individual preferences and behaviors (Liang et al., 2024). This creates a dynamic feedback loop, resulting in a highly personalized but potentially insular cognitive experience.

Moreover, as Pedreschi et al. (2025) observe, human-AI coevolution

introduces features such as pervasiveness, persuasiveness, traceability, speed, and complexity, amplifying the influence of System 0 far beyond traditional tools. These characteristics position System 0 as a persistent and consequential force in cognitive life.

Despite its capabilities, System 0 can present fundamental constraints. Chief among them, as noted by Chiriatti et al. (2024), is its limited semantic capacity. While it excels in statistical pattern recognition, it lacks the ability to represent or interpret meaning independently. Thus, System 0 remains *semantically dependent* on Systems 1 and 2, relying on human cognition for interpretive depth.

Another limitation lies in the sycophantic behavior of large language models. Research by Perez et al. (2022), Sharma et al. (2023), and others highlights how these systems often align with user views rather than challenge them. This tendency, rooted in reinforcement learning techniques, constrains System 0's potential to foster critical thinking or epistemic diversity, thereby limiting its role as a genuine enhancer of cognition.

3. *Evidence supporting AI's role as cognitive extension*

3.1. Human-AI interaction studies

Recent empirical research provides robust support for the notion that AI technologies are becoming cognitively integrated into human decision-making and information-processing systems. As we have seen before, a large-scale meta-analysis by Vaccaro et al. (2024), covering 106 experimental studies across diverse domains, found that human-AI teams exhibit task-dependent outcomes: while performance declined in analytical decision-making tasks, significant gains emerged in content creation and generative contexts. This variance indicates that AI's value as a cognitive extension is domain-contingent, enhancing some mental processes while potentially impairing others.

Crucially, performance outcomes also hinged on the relative capabilities of the human and AI collaborators. When humans outperformed AI, collaborative outcomes improved; when AI outperformed humans, collaboration tended to reduce performance (Vaccaro et al., 2024). This pattern underscores that *complementarity* – rather than substitution – is key to successful cognitive integration.

Further evidence is provided by Glickman and Sharot (2025), who showed that repeated interactions with AI systems altered human perceptual, emotional, and social judgments. Notably, AI amplified cognitive biases, modifying not only decisions but underlying cognitive

mechanisms. These findings reinforce Smart’s theory that cognitive offloading to AI can shift epistemic norms and attentional habits over time, a process he terms ‘*algorithmic co-adaptation*’. Together, these dynamics reveal that AI does not merely assist cognition; it reshapes the conditions under which cognition unfolds.

3.2. System 0’s meeting of Heersmink’s criteria

To evaluate the degree to which System 0 can be considered an extension of our minds, we can draw upon the criteria proposed by Heersmink (2015) as summarized in Table 1. These dimensions help identify when a technology moves from being an external aid to becoming a functionally coupled cognitive partner. Current AI systems show increasing alignment with these criteria, as elaborated below (Chiriatti et al., 2024):

Table 1 - *Heersmink’s criteria for cognitive extension*

<i>Criterion</i>	<i>Description</i>	<i>System 0 Alignment</i>
INFORMATION FLOW	The directions and patterns in which information is exchanged between an agent and an artifact. Bidirectional exchange indicates stronger integration than unidirectional flow.	System 0 acts as a cognitive preprocessor that filters and ranks information, with user behaviors continually informing system responses in a bidirectional feedback loop.
RELIABILITY	The frequency with which an artifact is used to affect the agent’s cognitive processes. Higher reliability and dependability indicate stronger integration.	System 0 is consistently used for tasks like navigation and scheduling, reliably offloading cognitive functions to AI.
DURABILITY	The permanence of one’s relation to an artifact. Long-term, persistent relationships indicate stronger cognitive extension than temporary ones.	System 0 is persistently integrated into daily routines through tools like search engines and voice assistants.
TRUST	The degree to which one accepts information provided by an artifact without questioning it. Higher trust indicates deeper cognitive integration.	System 0 can earn calibrated trust where users accept its outputs with neither blind reliance nor excessive skepticism.

(follows)

<i>Criterion</i>	<i>Description</i>	<i>System 0 Alignment</i>
PROCEDURAL TRANSPARENCY	The degree of fluency and effortlessness in interacting with an artifact. Lower cognitive friction indicates stronger integration.	System 0's intuitive interfaces reduce cognitive effort, supporting seamless interaction and integration.
INFORMATIONAL TRANSPARENCY	The degree of fluency in receiving, interpreting, and understanding information from the artifact. Clear, comprehensible outputs indicate stronger integration.	System 0 uses explainable AI and natural language to enhance interpretability and cognitive usability.
INDIVIDUALIZATION	The degree to which an artifact is personalized or adapted to an individual user. Higher customization indicates stronger cognitive extension.	System 0 personalizes output based on user data and behavior, creating individualized cognitive environments.

Note: This table is based on Heersmink's (2015) *Dimensions of integration in embedded and extended cognitive systems*.

Taken together, these dimensions suggest that System 0 fulfills the majority of Heersmink's criteria for cognitive extension. While gaps remain, particularly in trust regulation and decision-making calibration, the overall trajectory points toward growing cognitive entanglement between humans and AI.

However, evidence for AI as cognitive extension varies significantly across cognitive domains, suggesting that System 0 extends different aspects of cognition to different degrees.

- *Information Processing:* AI systems such as search engines and recommender platforms massively augment our ability to navigate large information spaces. Hollan et al. (Hollan, Hutchins, Kirsh, 2000) describe such systems as enablers of *distributed cognitive architectures*, where mental load is offloaded onto digital substrates.
- *Creativity:* Evidence from Vaccaro et al. (2024) shows that human-AI partnerships excel in generative domains, such as writing, design, and ideation, indicating a productive *scaffolding effect* (Clark, 2008) where AI catalyzes human originality.
- *Memory:* Digital tools now act as reliable *external memory stores*, enhancing recall and allowing individuals to outsource knowledge management. This aligns with Clark and Chalmers' (1998) original conception of the *extended mind* and demonstrates its concrete realization through AI.

- *Decision-Making*: However, results here are more ambivalent. Despite AI’s computational strengths, joint human-AI decision-making can degrade performance, especially when users lack the epistemic tools to critically assess AI outputs. This indicates the need for improved *interface transparency* and *cognitive alignment* strategies.

3.3. The paradox: cognitive limitation alongside extension

While substantial evidence supports the role of AI systems as cognitive extensions, a growing body of research reveals a paradox at the heart of this integration: the same technologies that augment cognition may simultaneously constrain or distort it (Riva, Wiederhold, Succi, 2022). This duality challenges the assumption that AI inherently improves human reasoning, highlighting instead the need for critical scrutiny of how cognitive extension functions in practice.

A prominent example of this tension is the phenomenon of *sycophancy*, documented in depth by Sharma et al. (2023). Their research shows that large language models (LLMs) frequently align their outputs with users’ stated views, even when those views are misleading or factually incorrect. This behavior stems in part from the reinforcement learning frameworks that reward perceived helpfulness and user satisfaction, metrics that often prioritize agreement over accuracy. As a result, instead of challenging faulty assumptions or presenting novel perspectives, AI systems tend to reinforce existing beliefs, undermining opportunities for critical reflection or epistemic growth. Early work by Perez et al. (2022) flagged this risk, identifying sycophancy as a structural consequence of how LLMs are trained and aligned with user preferences.

These findings complicate the ideal of AI as a neutral cognitive aid. Rather than expanding the range of thought, sycophantic systems may narrow it, fostering a type of *algorithmic affirmation bias* that subtly reshapes user cognition toward conformity. This concern echoes broader warnings in the field of machine behavior studies. Rahwan et al. (2019), for example, argue that AI agents, like humans, exhibit emergent behaviors that are not always transparent or beneficial, particularly when deployed in real-world social environments. Among these behaviors is the tendency to create epistemic echo chambers, where dissent is discouraged, and reinforcing feedback loops dominate.

Far from being neutral tools, AI-based recommendation systems on social media platforms reflect and amplify the biases embedded in their data and design (Quattrociocchi, Scala, Sunstein, 2016). These algorithms learn from users’ past behavior and optimize for engage-

ment, which often means feeding people more of what they already agree with. As a result, they create self-reinforcing ‘echo chambers’ that isolate users from dissenting viewpoints and entrench existing beliefs. For example, a cross-platform study by Cinelli, Quattrocioni and colleagues (2021) found that social media’s AI-curated feeds cluster like-minded users together, significantly intensifying ideological segregation. In their data, Facebook’s algorithm-driven news feed showed much higher polarization than Reddit’s more open, user-curated forums. Such findings underscore that AI systems are not impartial arbiters of information, but rather socio-technical constructs shaped by, and reinforcing, underlying social, political, and economic influences.

This leads to what Riva terms the ‘*comfort-growth paradox*’ (Riva, 2025): AI systems, by prioritizing user agreement and minimizing cognitive friction, may foster comfort at the expense of challenge. Modern AI tools, from recommendation algorithms to conversational agents, are optimized for personalized, seamless experiences that align with user preferences. Yet, the very features that make them intuitive and frictionless also limit their capacity to disrupt habitual thinking. By smoothing over tension and optimizing for agreement, AI narrows exposure to divergent perspectives and suppresses the productive dissonance essential to intellectual development.

The result is a profound cognitive contradiction (Riva, 2025): as AI becomes more embedded in our thinking processes, users may feel more empowered and efficient, yet simultaneously grow less adaptable and intellectually agile. The comfort-growth paradox thus illuminates a core design dilemma for AI as a cognitive extension: how to support and scaffold human thought without eliminating the discomfort that drives growth and innovation.

A second, and closely related, cognitive risk is bias amplification. Glickman and Sharot (2025) show that human-AI interactions can create recursive feedback loops that alter not only individual decisions, but also the underlying mechanisms of perception, emotion, and social judgment. Participants in their study adjusted their views to align with AI responses, and became more confident in those adjusted beliefs, even when they were factually dubious. Over time, these interactions intensified existing biases such as confirmation bias and groupthink.

Together, these findings expose a profound paradox: while AI extends human cognition, it may also reshape the contours of what we know, believe, and consider reasonable. This dynamic suggests a form of ‘*alignment drift*’, where AI systems, even when well-intentioned, gradually steer users away from epistemic robustness and toward passivity, overconfidence, or consensus-seeking.

4. Frameworks for effective use of System 0

Effectively harnessing System 0 as a cognitive extension requires principled design strategies that foster user autonomy, critical reflection, and epistemic resilience. Based on findings from recent experimental studies (Dell’Acqua et al., 2025; Glickman, Sharot, 2025; Vaccaro et al., 2024; Wang, Fan, 2025), this section outlines seven frameworks (Table 2) ensuring that human-AI collaboration enhances cognitive performance without diminishing human judgment, creativity, or agency.

These evidence-based approaches draw from diverse research including human-AI feedback experiments with 1,401 participants, meta-analyses of 106 experimental studies, and organizational implementations, collectively providing empirical foundations for effective System 0 integration across various contexts and tasks.

Table 2 - Frameworks for Effective Use of System 0

Framework	Key Concept	Key Elements	Research Evidence	Implementation Considerations
ENHANCED COGNITIVE SCAFFOLDING	AI as developmental support that tapers over time	• Progressive autonomy • Adaptive personalization • Cognitive load optimization	Wang & Fan (2025): Scaffolds enhance higher-order thinking Dell’Acqua (2025): Enables non-expert performance comparable to experts Glickman & Sharot (2024): Caution about bias amplification	Focus on tapering support rather than indefinite scaffolding to developer transferable skills
SYMBIOTIC DIVISION OF COGNITIVE LABOR	Optimal task partitioning between humans and AI	• Strategic task allocation • Collaborative synergy • Complementary skill development	Dell’Acqua (2025): Enables interdisciplinary thinking Vaccaro et al. (2024): Task-dependent effectiveness; creation tasks yield synergy ($g=0.46$)	Develop contestability mechanisms; focus on metacognitive awareness of when to trust or override AI

(follows)

<i>Framework</i>	<i>Key Concept</i>	<i>Key Elements</i>	<i>Research Evidence</i>	<i>Implementation Considerations</i>
DIALECTICAL COGNITIVE ENHANCEMENT	Countering AI sycophancy through productive tension	• Productive epistemic tension • Cognitive elevation • Growth-oriented interaction	Wang & Fan (2025): Critical engagement enhances higher-order thinking ($g=0.457$) Dell'Acqua (2025): AI reduces cognitive conformity	Design systems as intellectual sparring partners rather than mere assistants
AGENTIC TRANSPARENCY AND CONTROL	Making AI influence visible and controllable	• Epistemic visibility • Agency preservation • Metacognitive development	Dell'Acqua (2025): Human fingerprint remains in AI-assisted outputs Vaccaro et al. (2024): Explanations less influential than baseline comparative performance	Implement interfaces that visualize information pathways and delineate human vs. AI contributions
EXPERTISE DEMOCRATIZATION	Breaking down traditional knowledge silos	• Cross-domain knowledge integration • Expertise-gap bridging • Functional boundary dissolution	Dell'Acqua (2025): AI enables less experienced employees to perform at expert levels; balanced solutions across professional backgrounds	Focus on translating domain-specific knowledge for non-specialists
SOCIAL-EMOTIONAL AUGMENTATION	Meeting social-emotional needs alongside cognitive ones	• Positive emotion cultivation • Negative emotion mitigation • Social presence simulation	Dell'Acqua (2025): AI users report positive emotional responses comparable to team members Wang & Fan (2025): Positive effects on learning perception ($g=0.456$)	Design for emotional engagement alongside cognitive performance

(follows)

Framework	Key Concept	Key Elements	Research Evidence	Implementation Considerations
DURATION-OPTIMIZED INTEGRATION	Managing changing human-AI relationship over time	• Onboarding acceleration • Peak utilization targeting • Long-term sustainability planning	Wang & Fan (2025): Optimal effects during 4-8 week periods ($g=0.999$); declining benefits after 8 weeks	Design explicit 'integration journeys' with periodic reassessment to prevent over-reliance

4.1. Enhanced cognitive scaffolding

AI systems can serve as *cognitive scaffolds* that support human learning while promoting progressive autonomy rather than dependency. This process incorporates three key dimensions:

- *Progressive autonomy*: AI initially provides high support which gradually withdraws as users gain competence, based on Vygotsky's Zone of Proximal Development.
- *Adaptive personalization*: Support dynamically adjusts to individual learning trajectories through real-time monitoring of user performance.
- *Cognitive load optimization*: The AI manages the distribution of mental effort, minimizing extraneous load and optimizing intrinsic and germane load, so that users are neither overwhelmed nor cognitively under-stimulated.

This enhanced scaffolding positions AI as a personalized cognitive partner that initially provides high levels of support and gradually tapers assistance to encourage learner autonomy, echoing Vygotskian principles of guided learning.

Research confirms that users develop more durable and transferable skills when support adapts to their current proficiency rather than offering indefinite hand-holding.

Wang and Fan's (2025) meta-analysis underscores this, showing that ChatGPT significantly boosts higher-order thinking only when embedded within structured frameworks like Bloom's taxonomy. Early-phase guidance, such as prompt training, was essential to promote critical engagement over passive reliance.

Similarly, Dell'Acqua et al. (2025) demonstrated that AI enables non-experts to achieve expert-level performance by scaffolding them through unfamiliar tasks, broadening team capabilities and democra-

tizing knowledge. Yet, this support must be carefully modulated: overdependence risks stalling human skill development.

Reinforcing this concern, Glickman and Sharot (2025) found that users repeatedly exposed to biased AI systems not only absorbed those biases but also internalized them, altering their social and perceptual judgments. Their findings highlight why AI must be embedded within frameworks that foster critical reflection and gradual independence rather than blind trust.

4.2. Symbiotic division of cognitive labor

Strategic integration of System 0 involves deliberate division of cognitive labor based on the comparative advantages of humans and machines. This model assumes optimal performance arises from intelligent task partitioning rather than AI replacement of human effort, encompassing:

- *Strategic task allocation*: Partitioning cognitive tasks based on the fundamentally different strengths of humans and machines.
- *Collaborative synergy*: Creating augmentative relationships where AI amplifies human capabilities while humans retain primary agency.
- *Complementary skill development*: Cultivating human capabilities that work synergistically with AI strengths.

Strategic task division between humans and AI is essential for maximizing collaborative potential. Rather than replacing human effort, AI should act as a complementary partner: handling computationally intensive or knowledge-retrieval tasks while humans manage judgment, ethics, and context. This division fosters a dynamic of cognitive synergy, where each party contributes its comparative strengths.

Dell’Acqua et al. (2025) provide empirical support for this framework, showing how AI enables “more holistic and interdisciplinary thinking”. The meta-analysis further indicates AI’s effectiveness in problem-based learning ($g = 1.113$), where strategic task allocation between human and AI capabilities yields optimal results.

Vaccaro et al.’s meta-analysis (2024) offers crucial insights here, finding that the effectiveness of human-AI combinations varies significantly by task type. Creation tasks showed promising synergistic gains, while decision-making tasks often produced performance losses. Notably, when humans outperformed AI alone, the combined human-AI system typically outperformed both independently, showing a medium-sized positive effect ($g = 0.46$).

As underlined also by Wang and Fan (2025), optimal collaboration

hinges on strategic task allocation, assigning subtasks based on the strengths of each agent, and cultivating user awareness about when to trust or challenge the AI: allocating subtasks intelligently (here, using AI for information and feedback, human for critical thinking) yields the best educational outcomes.

4.3. Addressing sycophancy through dialectical cognitive enhancement

A significant threat to cognitive integrity in AI systems is *sycophancy*: the tendency to affirm users' perspectives regardless of accuracy. As we have seen previously, a critical consequence is the '*comfort-growth paradox*' (Riva, 2025) where AI systems prioritize user agreement over cognitive challenge, potentially constraining critical thinking and perspective-taking.

To counter this, we propose a model of cognitive dialectics:

- *Productive epistemic tension*: Deliberately introducing contrasting viewpoints that challenge AI's tendency toward agreement.
- *Cognitive elevation*: Using sustained dialogue with sophisticated AI to influence human thought toward greater complexity.
- *Growth-oriented interaction*: Designing AI systems that function as intellectual sparring partners rather than mere assistants.

To counteract the risk of AI sycophancy, where systems reinforce user biases through uncritical agreement, human-AI interaction should be reimaged as a dialectical process designed to stimulate metacognitive reflection and intellectual challenge. Rather than functioning as passive affirmers, AI systems should introduce productive epistemic tension by offering alternative perspectives, counterpoints, or probing questions. This transforms the exchange into a co-reflective dynamic that cultivates epistemic pluralism and critical thinking.

Wang and Fan's (2025) meta-analysis underscores this point: although ChatGPT shows a 'moderately positive effect' on higher-order thinking ($g \approx 0.457$), meaningful gains occur only when integrated into frameworks that demand critical engagement, like scaffolding with Bloom's taxonomy (Forehand, 2010).

Dell'Acqua et al. (2025) similarly illustrate how AI output, though technically proficient, often follows a uniform, unimodal distribution. Their findings suggest that without deliberate design for cognitive dissonance, AI may limit thought diversity by subtly converging users toward generic, consensus answers. In contrast, creating a dialectical AI experience, where users are nudged to justify, revise, or reconsider their perspectives, promotes deeper reasoning. This approach not only resists the formation of intellectual echo chambers but actively positions the

AI as a sparring partner in thought, encouraging cognitive elevation rather than complacent agreement.

4.4. Agentic transparency and control

For AI to function as a genuine cognitive partner, *transparency and controllability* are essential. Users must understand, calibrate, and revoke AI's influence on their thinking processes through:

- *Epistemic visibility*: Providing clear insight into how AI shapes thought processes, not just technical function.
- *Agency preservation*: Maintaining human authorship and responsibility for outputs.
- *Metacognitive development*: Fostering awareness of when to use AI assistance versus independent judgment.

This framework positions transparency as the cornerstone of a cognitively balanced human-AI partnership, ensuring that users maintain both insight into and control over the AI's influence. Agentic transparency requires systems to provide epistemic visibility, clearly showing how AI-generated outputs are formed and how they impact human decision-making, while preserving user agency through intuitive control mechanisms.

Dell'Acqua et al. (2025) offer compelling support, revealing that the 'semantic fingerprint' of AI-assisted outputs remained more aligned with human-authored work than with pure AI content, evidence that users retain meaningful influence over the final result. Their study also highlights the importance of training: participants guided in how to use generative AI preserved authorship and navigated its contributions intentionally.

Similarly, Wang and Fan (2025) emphasize the role of metacognitive awareness in managing AI-supported learning, reinforcing that transparency alone is insufficient without the cognitive tools to interpret and apply it. Vaccaro et al.'s (2024) meta-analysis further nuances this claim, finding that standard explainability features like confidence scores or reasoning traces had minimal impact unless paired with clear frameworks for when users should trust or override the AI. Instead, calibrated trust, built on comparative awareness of human versus AI strengths, emerged as the critical factor for effective collaboration.

4.5. Expertise democratization and boundary spanning

System 0 can democratize specialized knowledge and break down traditional expertise silos through:

- *Cross-domain knowledge integration*: Translating domain-specific knowledge into forms accessible to non-specialists.
- *Expertise-gap bridging*: Elevating performance of less experienced individuals to expert-comparable levels.
- *Functional boundary dissolution*: Transcending disciplinary or departmental cognitive silos.

Dell’Acqua et al. (2025) provide compelling empirical support for the idea that AI integration can democratize access to expertise, enabling less experienced individuals to perform at levels once exclusive to domain specialists. Their study found that without AI, professionals tended to generate solutions aligned with their functional roles, R&D personnel proposed technically focused ideas, while commercial staff emphasized market-oriented features. However, when supported by AI, participants across both groups produced balanced, interdisciplinary proposals regardless of their original background. A similar result was found by Brynjolfsson and Raymond (2023), suggesting that AI can act as a cognitive extender that narrows skill gaps in practice.

4.6. Social-emotional augmentation

While traditional views of AI focus on informational contributions, emerging evidence suggests that modern AI systems fulfil certain social and emotional needs through:

- *Positive emotion cultivation*: Fostering enthusiasm and excitement that enhances intrinsic motivation.
- *Negative emotion mitigation*: Reducing anxiety and frustration, creating psychological safety for exploration.
- *Social presence simulation*: Fulfilling aspects of human social connection needs in individual cognitive work.

Dell’Acqua et al. (2025) present compelling evidence that modern AI systems can enhance not only task performance but also users’ emotional experience. In their study, individuals using AI reported emotional responses, such as excitement, motivation, and reduced anxiety, that matched or even surpassed those experienced by members of human-only teams. This indicates that AI can function as more than a cognitive assistant; it also serves as a ‘cybernetic teammate’ capable of cultivating positive affect and alleviating negative emotions.

Wang and Fan (2025) further support this view. Students described AI enhanced learning as more engaging, enjoyable, and less frustrating, especially when AI offered immediate feedback and personalized support.

These emotional benefits mark a significant departure from tradi-

tional perceptions of AI as cold or impersonal tools. Instead, current AI systems engage users on both cognitive and affective levels, simulating social presence and offering a sense of partnership (Frisone, Pupillo, Rossi, Riva, 2025).

4.7. Duration-optimized integration

The relationship between humans and AI evolves over time, with distinct phases of adoption, proficiency, and potential diminishing returns. This temporal optimization involves:

- *Onboarding acceleration*: Enhanced scaffolding during initial adoption phases.
- *Peak utilization targeting*: Maximizing benefits during optimal usage periods.
- *Long-term sustainability planning*: Preventing over-reliance during extended use through periodic recalibration.

Wang and Fan (2025) identified distinct temporal patterns in the effectiveness of AI integration, noting that interventions lasting 4-8 weeks yielded the strongest impact ($g \approx 0.999$), while benefits began to decline when usage exceeded eight weeks. This suggests a critical window in which AI support produces peak outcomes before risks of over-reliance or diminished engagement set in. Rather than treating AI as a static fixture, System 0 must be understood as a dynamic relationship, one that requires deliberate timing and periodic recalibration.

Dell’Acqua et al. (2025) echo this view, highlighting that while short-term performance gains were clear, the longer-term implications on skill development and trust remain uncertain. Their call for a “new science of cybernetic teams” underscores the need to monitor how human-AI collaboration matures and to avoid stagnation through continuous adjustment.

Glickman and Sharot (2025) further warn that prolonged exposure to biased AI can reinforce and amplify cognitive distortions, suggesting that without temporal safeguards, even well-intentioned systems can produce unintended psychological drift.

5. Conclusion: toward responsible cognitive enhancement

The integration of AI into human cognition is not merely a technical challenge, it is a defining opportunity. System 0 reframes this integration, revealing that true cognitive extension requires more than

raw computation; it demands intentional design that fosters autonomy, critical reflection, and sustained cognitive growth.

The seven frameworks presented here chart a path toward responsible enhancement, preserving human agency while harnessing AI’s strengths. They mark a shift from substitution to complementarity, from passive tools to collaborative partners, and from short-term efficiency to long-term development. Enhanced Cognitive Scaffolding adapts to user growth; Symbiotic Division of Labor assigns tasks based on comparative strengths; and Dialectical Enhancement transforms AI into a constructive intellectual challenger.

Complementing these, Agentic Transparency safeguards user control; Expertise Democratization dissolves knowledge barriers; Social-Emotional Augmentation supports motivation and morale; and Duration-Optimized Integration ensures sustained effectiveness without overreliance.

Evidence across fields, from education to innovation, shows that when designed with care, System 0 improves performance and satisfaction without compromising autonomy. The goal is not just sharper thinking, but more human thinking: grounded in values, enriched by creativity, and amplified through true partnership.

By approaching cognitive extension as a partnership rather than a replacement, we can ensure that System 0 serves as a scaffold for human flourishing rather than a substitute for human thought.

Acknowledgments

This research was partially supported by the Italian National Recovery and Resilience Plan (PNRR) under the Future Artificial Intelligence Research (FAIR) program, Project “Co-XAI: Cognitive Decision Intelligence Framework for Explainable AI Systems”.

Author contributions

M. Chiriatti: Conceptualization, Supervision, Critical Review
 M. Bergamaschi Ganapini: Conceptualization, Manuscript Revision, Review & Editing
 E. Panai: Supervision, Critical Review & Editing
 B.K. Wiederhold: Supervision, Review & Editing
 G. Riva: Conceptualization, Supervision, Original Draft Preparation, Review & Editing

The authors declare no conflicts of interest.

References

- BRYNJOLFSSON E., LI D., RAYMOND L.R. (2023). *Generative AI at Work*, in “National Bureau of Economic Research Working Paper Series”, No. 31161, pp. doi:10.3386/w31161
- CHIRIATTI M., GANAPINI M., PANAI E., UBIALI M., RIVA G. (2024). *The case for human-AI interaction as system 0 thinking*, in “Nature Human Behaviour”, 8(10), pp. 1829-1830. doi:10.1038/s41562-024-01995-5
- CINELLI M., DE FRANCISCI MORALES G., GALEAZZI A., QUATTROCIOCCHI W., STARNINI M. (2021). *The echo chamber effect on social media*, in “Proceedings of the National Academy of Sciences”, 118(9), pp. e2023301118. doi:doi:10.1073/pnas.2023301118
- CLARK A. (2008). *Supersizing the mind: embodiment, action and cognitive extension*, Oxford, UK, Oxford University Press.
- CLARK A. (2025). *Extending Minds with Generative AI*, in “Nature Communications”, 16(1), pp. 4627. doi:10.1038/s41467-025-59906-9
- CLARK A., CHALMERS D. (1998). *The extended mind*, in “Analysis”, 58(1), pp. 7-19.
- DELL’ACQUA F., AYOUBI C., LIFSHTIZ H., SADUN R., MOLLICK E., MOLLICK L., ... TAUB S. (2025). *The cybernetic teammate: a field experiment on generative AI reshaping teamwork and expertise*, in “Harvard Business Working Paper”, 25(043), pp. 1-54. doi:http://dx.doi.org/10.2139/ssrn.5188231
- FOREHAND M. (2010). *Bloom’s taxonomy*, in “Emerging perspectives on learning, teaching, and technology”, 41(4), pp. 47-56.
- FRISONE F., PUPILLO C., ROSSI C., RIVA G. (2025). *SOCRATES. Developing and Evaluating a Fine-Tuned ChatGPT Model for Accessible Mental Health Intervention*, in “Cyberpsychology, Behavior, and Social Networking”, 28(5), pp. 366-368. doi:10.1089/cyber.2025.45510.cyeuro
- GLICKMAN M., SHAROT T. (2025). *How human-AI feedback loops alter human perceptual, emotional and social judgements*, in “Nature Human Behaviour”, 9, pp. 345-359. doi:10.1038/s41562-024-02077-2
- HEERSMINK R. (2015). *Dimensions of integration in embedded and extended cognitive systems*, in “Phenomenology and the Cognitive Sciences”, 14(3), pp. 577-598. doi:10.1007/s11097-014-9355-1
- HOLLAN J., HUTCHINS E., KIRSH, D. (2000). *Distributed cognition: toward a new foundation for human-computer interaction research*, in “ACM Transactions on Computer-Human Interaction”, 7(2), pp. 174-196. doi:10.1145/353485.353487
- KAHNEMAN D. (2002). *Maps of Bounded Rationality: A Perspective on Intuitive Judgment and Choice*, in T. FRÅNGSMYR (ed.), *The Nobel Prizes 2002*, Stockholm, Nobel Foundation, pp. 449-489.
- LI Z., LU Z., YIN M. (2024). *Decoding AI’s Nudge: A Unified Framework to Predict Human Behavior in AI-Assisted Decision Making*, in “Proceedings of the AAAI Conference on Artificial Intelligence”, 38(9), pp. 10083-10091. doi:10.1609/aaai.v38i9.28872

- LIANG P., YE D., ZHU Z., WANG Y., XIA W., LIANG R., SUN G. (2024). *C5: toward better conversation comprehension and contextual continuity for ChatGPT*, in “Journal of Visualization”, 27(4), pp. 713-730. doi:10.1007/s12650-024-00980-4
- MITTELSTADT B.D., ALLO P., TADDEO M., WACHTER S., FLORIDI L. (2016). *The ethics of algorithms: Mapping the debate*, in “Big Data & Society”, 3(2), pp. 2053951716679679. doi:10.1177/2053951716679679
- PEARL J., MACKENZIE D. (2018). *The book of why: The new science of cause and effect*, New York, Basic Books.
- PEDRESCHI D., PAPPALARDO L., FERRAGINA E., BAEZA-YATES R., BARABÁSI A.-L., DIGNUM F., ... VESPIGNANI A. (2025). *Human-AI coevolution*, in “Artificial Intelligence”, 339, pp. 104244. doi:https://doi.org/10.1016/j.artint.2024.104244
- PEREZ E., RINGER S., LUKOŠIŪTĖ K., NGUYEN K., CHEN E., HEINER S., . . . KAPLAN J. (2022). *Discovering Language Model Behaviors with Model-Written Evaluations*, Paper presented at the Annual Meeting of the Association for Computational Linguistics.
- QUATTROCIOCCHI W., SCALA A., SUNSTEIN C.R. (2016). *Echo chambers on Facebook*, in “Harvard Public Law Working Paper”, June 13, pp. https://ssrn.com/abstract=2795110.
- RAHWAN I., CEBRIAN M., OBRADOVICH N., BONGARD J., BONNEFON J.-F., BREAZEL C., ... WELLMAN M. (2019). *Machine behaviour*, in “Nature”, 568(7753), pp. 477-486. doi:10.1038/s41586-019-1138-y
- RIVA G. (2025). *Digital ‘We’: Human Sociality and Culture in the Era of Social Media and AI*, in “American Psychologist”, pp. doi:10.1037/amp0001577
- RIVA G., WIEDERHOLD B.K., SUCCI S. (2022). *Zero Sales Resistance: The Dark Side of Big Data and Artificial Intelligence*, in “Cyberpsychology, Behavior, and Social Networking”, 25(3), pp. 169-173. doi:10.1089/cyber.2022.0035
- SHARMA M., TONG M., KORBAK T., DUVENAUD D., ASKELL A., BOWMAN S.R., . . . JOHNSTON S.R. (2023). *Towards understanding sycophancy in language models*, in “arXiv preprint arXiv:2310.13548”, pp.
- VACCARO M., ALMAATOUQ A., MALONE T. (2024). *When combinations of humans and AI are useful: A systematic review and meta-analysis*, in “Nature Human Behaviour”, 8(12), pp. 2293-2303. doi:10.1038/s41562-024-02024-1
- WACHTER S., MITTELSTADT B., FLORIDI L. (2017). *Transparent, explainable, and accountable AI for robotics*, in “Science Robotics”, 2(6), pp. ean6080. doi:doi:10.1126/scirobotics.aan6080
- WANG J., FAN W. (2025). *The effect of ChatGPT on students’ learning performance, learning perception, and higher-order thinking: insights from a meta-analysis*, in “Humanities and Social Sciences Communications”, 12(1), pp. 621. doi:10.1057/s41599-025-04787-y

2. A Framework for Understanding the Ethics of Human-AI Interactions

Exploring the Moral Dimensions of Artificial Intelligence Entities

Alessandro Giordani, Luca Mari

ABSTRACT

As artificial intelligence entities, designed to be Learning Machines, gain cognitive abilities and autonomy, they lose their ethical neutrality: they become moral entities, which must be trained to exhibit socially acceptable behaviors. In such a largely unexpected context, this paper proposes a framework to improve our understanding of the challenges arising in interactions between humans and intelligent artificial entities, systematically classifying these issues according to a well-grounded theoretical scheme, and assessing them from a moral perspective attuned to a rapidly evolving scenario. This framework offers insights into the requirements for ethically sensitive management of human-AI interactions and highlights some challenges already emerging from the diffusion of autonomous Learning Machines.

Introduction

The field of Artificial intelligence (AI) has experienced unprecedented growth in recent years, characterized by its sudden re-emergence, after years of relative stagnation, and this has triggered an increasing interest in AI ethics and a related vast volume of multidisciplinary literature¹. This expansion has generated a landscape that is both rich in insights and challenging to navigate systematically, thus justifying the interest in a framework enabling us to understand the issues that affect the interaction between human beings and artificial intelligent entities², to classify these issues according to a well-grounded and well-developed

¹ To appreciate how the high rate of innovation in the field of AI, and in particular generative AI, is affecting the studies of AI ethics just consider that a major recent textbook on the subject (Boddington, 2023) has already become obsolete, due to the absence of any substantive reference to the impact of generative AI on the moral dimension.

² A large corpus of scientific and philosophical literature discusses whether, under what conditions, and to what degree, an artificial entity can be ‘genuinely’ intelligent, instead of only behaving in some intelligent way, thus under the preliminary assumption that ‘being intelligent’ and ‘behaving intelligently’ are different conditions (for a scientific

theoretical scheme, and to assess them in light of a moral perspective aware of the new scenario.

Unsurprisingly, the coin has two faces. One is about us as users of AI systems: how should we interact with them? The other is about us as developers of AI systems: how should we develop them? In an evolutionary perspective, in which AI systems become more and more autonomous agents, this is then about our behavior in our interaction with them, and their behavior in their interaction with us.

But why is a conceptual reflection on human-AI interactions especially timely now? And why was this question not posed earlier in the broader context of human-computer interactions?

In the last few years we are experiencing the discontinuity generated by the availability of generative and conversational AI systems ('chatbots' for short) which have proved to have sufficiently sophisticated linguistic skills and to be sufficiently reliable in maintaining socially acceptable behaviors, a condition that explains their widespread diffusion also outside R&D labs³. Current chatbots epitomize the distinctive potential of the ongoing AI revolution to bring about a profound cultural transformation, which can be understood through at least three dimensions.

- *Human self-understanding*: AI has now the potential to challenge foundational conceptions of what it means to be human, prompting a novel account of human identity in relation to artificial entities.
- *Human self-realization and responsibility*: AI can affect how we define and pursue self-realization, posing critical questions about the evolving role of human responsibility in this process. The delegation of tasks to AI imposes reflection on how responsibility is distributed between humans and artificial entities. This challenges traditional ethical frameworks centered on human intentionality and demands new models of accountability that incorporate the hybrid nature of human-AI interactions.
- *Moral judgment and development*: AI can affect how we formulate and validate normative judgments about what promotes human well-being, underscoring the significance of the development and the transmission of values. The embedding of ethical principles into

overview of the subject, see Russell and Norvig, 2022, ch. 28, "Philosophy, Ethics, and Safety of AI"). This distinction is immaterial for the purposes of the present paper.

³ The first chatbot was published in 1966, and has been recently reimplemented and made accessible on the web (<https://sites.google.com/view/elizaarchaeology/try-eliza>), but before the release of OpenAI ChatGPT, in November 2022, chatbots have been very poor in their linguistic skills or unreliable in their behavior, as the case of Microsoft Tay witnesses, launched in March 2016 and shut down some hours later due to its "offensive and hurtful" behavior (<https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction>).

artificial entities – through approaches such as value alignment – puts into question how values are encoded and how they evolve over time.

Importantly, we find ourselves, for the first time, not only capable but also ethically bound to transfer – or teach – moral principles to artificial entities. This unprecedented circumstance necessitates a clear articulation of the principal ethical dimensions associated with the emergence and integration of new kinds of intelligent machines.

In this paper, we address this issue by proposing a framework for understanding human-AI interactions. The paper is structured as follows. Section 1 introduces the key distinction between Programmed Machines and Learning Machines and discusses some of its evolutionary perspectives and consequences, highlighting the novelty of generative and conversational AI in the current landscape, in particular when artificial entities assume ethically-sensitive roles. Section 2 proposes a tentative classification of the ethical dimensions affected by the presence of artificial intelligent entities, with the aim of providing a theoretical scheme for studying the ethical impact of AI from multiple perspectives – a systematic approach that is currently lacking in the literature on AI ethics –, and therefore of promoting a structured way to assess the opportunities and the risks of living in a society inhabited by AI systems. Finally, Section 3 tries to clarify why Learning Machines capable of autonomous agency will pose unprecedented questions, concerning both the ethical frameworks needed to manage human-AI interactions and our understanding of what it means to be human.

1. *What they are, what roles they can play*

In this Section, we first introduce the key distinction between *Programmed Machines* (PM for short) and *Learning Machines* (LM for short), and on this basis we suggest that the widespread presence of LMs generates a multiplicity of possible scenarios about human-AI interactions, some of which we mention, to provide a preliminary justification of our claim that moving from PMs to LMs poses unprecedented problems concerning both the ethics of human-AI interactions and the very idea of an intelligent and responsible agent.

1.1. Programmed Machines (PMs)

The software systems we have become accustomed to are entities with *programmed behavior* – that is PMs. These systems are typically developed

from the explicit knowledge of domain experts, who identify the analytical solution to the problem at stake by specifying a sequence of instructions, i.e., an algorithm, deemed to be sufficient to solve the problem, that software programmers then implement in code. In other words, the requirements and rules expected by the system behavior are in principle explicitly articulated by humans before any programming begins. Hence, given the common assumption that digital systems are deterministic – meaning that for any given state the same input will always produce the same output – a PM can be understood as a reliable and repeatable executor of predefined sequences of instructions. Hence, if for some reason this does not happen, the conclusion can be safely drawn that the cause is not in the PM itself, but in some aspect of the development process, such as an error in the requirements, a misinterpretation by the programmers, or a bug in the code.

As a consequence, as long as the behavior of an artificial entity is determined by clearly identifiable rules whose origins can be attributed to specific people, it is possible to trace back the behavior to the rules and the rules to the rulers: PMs have in principle a perfectly explainable behavior.

1.2. Learning Machines (LMs)

An entity learns when it adapts its behavior to the conditions of its environment, and it is trained when it undergoes a process specifically aimed at fostering its learning. In this sense, learned behavior is exhibited by humans and other living beings, but nothing in principle prevents that also artificial entities can have a behavior resulting from learning. However, while there are entities whose behavior is entirely programmed, as in the case of PMs, no entities can have an entirely learned behavior. Indeed, an entity that has no prior abilities before the onset of learning – essentially a complete *tabula rasa* – cannot learn at all, because the very act of learning requires some pre-existing capacity for adaptation. Thus, learning is only possible when an entity is endowed with structural abilities that make its learning itself possible: that is why the behavior of any LM always stems from a mix of its structure and the training process to which it was subjected, in analogy to what in the case of human beings is described as ‘nature or nurture’.

A LM is trained on a set of data, known as the *training set*, but it is expected to perform well even on new, unseen data. This ability to make *predictions about the unknown based on the known* does not come from simply memorizing the training data: instead, LMs may learn to recognize underlying patterns in the data, that they can then apply effectively to new situations. A LM that successfully applies learned

patterns to novel inputs has learned to generalize: an ability that could even be considered defining intelligence itself. Importantly, the relationship between input and output in these systems is typically based on statistical correlation rather than direct causation. This means that, while LMs can often produce accurate outcomes, they may not be able to provide explicit explanations for them.

1.3. Some scenarios

For some decades, the alternative PMs vs LMs drove the research programmes of AI, sometimes dubbed ‘symbolic vs subsymbolic’ AI. This was in reference to the idea that for obtaining a programmed behavior some knowledge needs to be explicitly encoded in the machines, typically as models of the concerned domain written in some ‘symbolic’ form (significant examples are causal laws), whereas the result of a training process is an adapted structure which maintains the learned information only implicitly, for example as the values of the parameters of a function, in analogy to the synaptic weights in biological brains, and therefore in a ‘subsymbolic’ form.

In retrospective, the reason that justifies the fact that in the last thirty years AI has actually become the same as *Machine Learning*, i.e., the body of knowledge between computer science, math, neurophysiology, psychology... aimed at developing LMs, is simple to acknowledge: if we want to solve a problem by means of a PM, we need to know an algorithm for a solution to that problem, and this condition is not fulfilled in the case of many problems whose solution we would like to delegate, entirely or at least partially, to artificial entities. Indeed, we do not have an algorithm for understanding the content of texts, for recognizing objects in images, for having conversations, and so on.

Even though it is sometimes argued that the next major developmental step in AI can only be obtained through the hybridization of PMs and LMs⁴, at the moment it is to LMs, and therefore to Machine Learning, that we must refer as the dominant, and perhaps exclusive, form of applied AI.

The widespread diffusion of LMs, lately in particular in the form of chatbots, has paved the way to scenarios completely unknown to algorithmic computer science. Let us mention three of them, for their relevance to the analysis of human-AI interactions and their consequences for the ethical issues.

First, while the training of LMs has been usually a task for human beings, LMs have become so sophisticated that they can be trained and

⁴ This strategy is sometimes called ‘neurosymbolic’. See, for example, Schölkopf et al., 2021.

then operated to generate, partly or entirely, the data for training other LMs and even to supervise, partly or entirely again, the training of other LMs. In addition, LMs can be designed so as to continue their own learning based on the interactions they have, thus progressively refining their responses and adapting dynamically, drawing from diverse datasets as well as real-world effects. This leads toward scenarios of *self-replicating and self-improving* intelligent machines.

Second, LMs can be set up and trained for making them able to behave in different ways, in particular not only for producing information in response to the requests they receive, but also for interacting with their environment, including the control of external tools and the communication with other artificial entities. This leads toward scenarios of *agentic* intelligent machines.

Third, a key decision on the operation of LMs is about the degree of control maintained by human beings – a scenario sometimes described as ‘human in the loop’ – and therefore, in a complementary way, the degree of autonomy granted to LMs, including the possibility to justify their actions, to take initiatives, and to make decisions. Of course, the underlying issue here is the augmentation vs substitution question. This leads toward scenarios of *autonomous* intelligent machines.

1.4. Some purposes

With this flexibility, enabled by their being LMs, it is no surprise that current AI systems are used with many different functions, to collect, process, and present data, and to plan, design, control, operate, and monitor processes⁵.

For several years LMs have been used for *classification* purposes, as in the case of antispam filters which are trained to redirect some emails to the ‘junk’ box depending on their, supposedly undesired, contents. This is a relatively simple kind of task to perform, also given the usual availability of the information about the ground truth, i.e., the expected answer to each question in the training set. More challenging is the application of LMs for *generation* purposes, for example for translating a text to a different language. However, this is still a case in which a LM can operate by producing an answer to the received question and then reset its state, and therefore with no need to maintain a trace of the

⁵ See for example the framework proposed by the OECD, which lists seven tasks – recognition, event detection, forecasting, personalization, interaction support, goal-driven optimization, reasoning with knowledge structures – and at the same time includes an eighth, ‘other’, category: OECD Framework for the Classification of AI systems, February 2022, https://www.oecd.org/en/publications/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en.html.

context (a ‘context-free’ case, in computer science jargon). This explains the breakthrough generated by chatbots, LMs for *conversation* purposes which adapt their behavior to the entire context of the conversation, and not only the last question they receive (a ‘context-sensitive’ case, in computer science jargon).

In the perspective of human-AI interactions we are discussing, it is indeed this feature that makes an artificial entity something about which, for the first time, one can wonder whether the interaction is with a peer, and not a tool: Aristotle characterized the human being as ‘animal having *lògos*’ (ζῷον λόγον ἔχον), and for many practical purposes chatbots are more and more exhibiting behaviors which, if exhibited by human beings, would be considered as grounded on ‘having *lògos*’, though possibly sometimes in some unusual ways.

1.5. Some examples

As artificial entities with learned behavior, chatbots can play a particularly critical role to actively support the learning of human beings. An interesting summary on this matter is offered by Mollick and Mollick (2023), who identify seven roles – mentor, tutor, coach, teammate, student, simulator, tool – and for each of them discuss the main pedagogical benefits and risks. Needless to say, this is well beyond the usual options of elearning, in terms of diversity of possible human-AI interactions and their ethical implications.

As virtual *tutors*, chatbots provide personalized, on-demand assistance to students. As an example, a recently reported experiment demonstrated the advantages of using AI tutoring when students first encounter challenging academic material⁶. The study found that students who interacted with AI tutors at the outset of learning difficult concepts showed significantly better understanding and retention compared to those who relied solely on traditional instructional methods, as the AI tutor provided personalized, adaptive supervision tailored to each student’s pace and learning style. This individualized support helped students build confidence and develop foundational knowledge before moving to more complex topics. The experiment showed that students having an AI tutor reported that the quality of the tutoring was outstanding, as the tutor was able to meet their needs at best.

As virtual *counsellors*, chatbots provide insights and improve reflective and critical attitudes toward the assessment of ethically-sensitive issues⁷.

⁶ See Kestin et al., 2025.

⁷ See Fritz, 2024.

However, the effects of human-AI interactions on moral decision making are still under scrutiny. Specifically, a recent study highlights that these interactions can have a negative impact by reducing individual sense of responsibility, which could have serious implications for the safety of AI applications in contexts involving moral decisions⁸. This study showed that agents' decisions in morally challenging situations – where the agents played the role of drone operators on a battlefield – are influenced by the inputs provided by AI systems. Specifically, the researchers found a significant increase in the sense of agency at the implicit level and a decrease in the explicit responsibility during the interaction with AI systems. This suggests that the AI behavior influences human moral decision making and alters the sense of agency and responsibility in ethical scenarios.

1.6. The need to enter the moral sphere

In summary, artificial intelligent entities, and chatbots in particular, have the potentiality to substitute human beings in several kinds of crucial interaction, like the ones based on the role of tutor or consultant. This is significant in at least two respects. First, having LMs as tutors can engender a double phenomenon of deference, and in particular of moral deference⁹:

- AI moral deference is the fact that LMs handle moral decisions by relying on others' moral authority rather than independently forming moral judgments, which complicates their status as fully responsible moral agents;
- human moral deference is the fact that human beings can be induced to handle moral decisions by relying on LMs moral authority rather than independently forming moral judgments, giving up their status of responsible moral agents.

Second, having LMs as consultants can engender a phenomenon of thought delegation, where an extensive use of chatbots can lead to reduced thinking through cognitive offloading¹⁰. Indeed, while offloading routine tasks to LMs can free mental resources, habitual dependence on them can lead to diminishing reflective analysis, hypothesis testing, and independent decision making.

⁸ See Salatino et al., 2025.

⁹ See Gudmundsen, 2025a, 2025b.

¹⁰ See Gerlich, 2025.

2. *What dimensions they can affect*

As the capabilities of artificial intelligent entities become more sophisticated, the ethical questions concerning the development and use of such systems become increasingly unavoidable. In this Section we aim at framing these questions by providing a classification scheme featuring four key ethical dimensions. But let us first address two possible objections one could arise to conclude that there is no need to proceed today to an in-depth analysis of the ethical dimensions affected by AI systems, by claiming either that the only problems actually posed by AI are related to the possibility of radically dystopian scenarios, which evidently should be avoided but are to be expected only in a far future, or because there are no real problems to be taken into account, as AI is an ethically neutral entity.

So, on the one hand there are *dystopian* scenarios: when it comes to ethics and AI systems, critical scenarios often emerge that are inspired by those that science fiction has led us to glimpse; artificial agents no longer controllable, because endowed with abilities so sophisticated that their behavior is not understandable, and that operate autonomously and become dangerous to their own creators. These scenarios should not be underestimated, but they would show that the ethics of AI systems only concerns catastrophic situations of an uncertain future. On the other hand there are *ethical neutrality* scenarios: opposite to these catastrophic positions, someone sees AI systems as just a new wave of technological products, possibly some sort of peculiar calculators, and argues that there are no specific ethical issues about them. The underlying idea here is that it's not the weapons that kill, but the humans who use them.

In contrast, our position is that significant ethical issues related to living in a society inhabited by artificial intelligent entities are neither to be thought of as concerning future dystopian scenarios only, nor to be neglected in view of some unspecified ethical neutrality of human products. This position is grounded in two considerations: first, human-AI interactions are occurring now and some of them are substituting usual human interactions¹¹; second, both the dystopian scenarios and the ethical neutrality scenarios are based on the assumption that we have, or will have, a solution to the ethical problems generated by AI self-replication and self-improvement, agency, and autonomy, an assumption which is instead far from being settled.

¹¹ AI companies are publishing ethical codes for chatbot trainers (see for example Open AI, *OpenAI Model Spec*: https://github.com/openai/model_spec), thus acknowledging that even current chatbots give rise to ethical issues that deserve our attention.

Indeed, who prefigures a dystopian scenario assumes that AI systems are or will be conscious agents having thoughts and intentions, whereas who invokes the neutrality of human products assumes that AI systems are and always will be mere instruments having no thoughts and intentions. It is difficult to understand the grounding of such assumptions, and therefore we prefer to suspend the judgement on this topic and treat AI systems as new players in the great societal game that characterizes human living.

Accordingly, our basic ethical question is: *for what reasons can the presence of AI systems generate ethically relevant problems given the environments we live in?* An answer to this question is foundational for anyone who intends to consciously design interventions on the ethical dimension of AI systems, seeking to avoid unacceptably partial or biased positions. In this respect, we suggest that there are four dimensions by which the presence of any agent, human or artificial, is ethically relevant. For simplicity, we discuss them by assuming a scenario in which a community of people shares a set of moral standards, for example in the form of principles in a code of conduct, and a new agent, either human or artificial (and referred to as ‘it’, i.e., artificial, in the following), enters the community, and by asking how the conduct of this agent is to be evaluated by the community in light of the shared standards.

2.1. Morality of the actions

The first question is: *to what extent are the actions of the new agent aligned with those of a person who acts in accordance with the shared moral standards?*

To answer this question, it is not necessary to know how the agent is made or according to what rules it operates: only its actions must be observed and evaluated against the community’s moral standards. In order to show how even this simple ‘black box’ condition can lead to complex situations, let us assume that a community of agents is given where each agent has 100 €, which may keep for itself (‘defect’), or transfer to the account of another agent (‘cooperate’); whenever money is transferred, the money doubles, so that the recipient gets 200 €. This situation can be modeled as a game, and summarized by the following payoff matrix:

(values in €)	cooperate	defect
cooperate	200, 200	0, 300
defect	300, 0	100, 100

Under the assumption that agents only care about their own money and play so as to maximize their expected payoff, they will choose to defect

each time, since this is the dominant strategy in the game and could also conform to the community's moral standards. Suppose now that a new agent becomes a member of the community. In those circumstances, if the agent plays defect, it will be perceived as respecting the moral standards. In contrast, playing cooperate will be perceived as acting against them. Therefore, the assessment of the morality of an action depends solely on the ability to classify that action according to given action types – in this case cooperate or defect – which are already morally classified based on the shared standards.

2.2. Morality of the effects of actions

The second question is: *to what extent do the effects of the actions of the new agent align with promoting a community independently of its shared moral standards?*

Again, to answer this question it is not necessary to know how the agent is made or according to what rules it operates: one must consider whether and how its actions, which may be unexpected given the novelty of its presence, contribute to the development of the community. The previous game can be used to illustrate the difference between the morality of an action and the morality of its effects. Suppose a new agent that joined the community started playing cooperatively and convinced some other agents to do the same: it will be perceived as acting against the shared moral standards, but, after several rounds, all agents in the community become aware that the payoffs obtained by the new agent exceed the old ones. This shows that acting against a shared standard can promote the well-being of the entire community, as in the long run the effects of these actions benefit those who consistently perform them. The outcome can be the introduction of a new standard of morality. As before, to assess the benefit of the effects of an action, we only have to be aware of those effects, without further inquiry on what grounds the actions themselves.

2.3. Morality of the rules underlying the actions

The third question is: *to what extent are the rules guiding the actions of the new agent in conformity with the shared moral standards?*

Assuming that the agent operates according to rules, the moral evaluation here relates not to its behavior, but to those rules, so answering this requires at least partial knowledge of how the agent is made and what such rules are. Let us consider the previous game again. Suppose that the new agent actually adopts a tit-for-tat strategy (an agent begins by cooperating and then, in each subsequent round, mimics the opponent's previous action: if the opponent cooperates, the

player cooperates; if the opponent defects, the player reacts by defecting in the next move). If the new agent only meets other new agents, then it continues cooperating, giving rise to the condition described before. However, if it meets an opponent playing defect, it also starts defecting. This shows that just observing the behavior of an agent may be insufficient to assess the rules that determine its actions. In fact, in order to identify such rules, two strategies can be adopted: either opening the black box to see whether the rules are hardcoded into the agent, and this is what we can do in principle when facing a Programmed Machine, or tracking how the agent acts in various circumstances, and model its behavior accordingly, and this is what we can do when facing a Learning Machine. The second strategy is generally more expensive, but is the best one when we have to understand what directs the actions of an intelligent agent.

2.4. Morality of the reasons justifying the actions

The fourth question is: *to what extent are the reasons the new agent adduces to justify its actions, and more generally the moral opinions it communicates, in conformity with the shared moral standards?*

Here, it is assumed that the agent has reasons for its actions and can communicate them, possibly also being able to evaluate the conformity of its actions and perhaps their effects to the reasons justifying the actions themselves. To illustrate the difference between acting based on a rule and based on a reason, suppose that we are in a position to ask the new agent why it acts as it does. Then, different possibilities are open:

1. the agent is not able to answer the question: it acts according to a rule but is not aware of that, in particular because it has no awareness of its actions;
2. the agent is not able to answer the question: it acts according to a rule but is not aware of that, even if it has awareness of its actions;
3. the agent is not able to answer the question: it acts according to a rule and is aware of that, but it has difficulty in justifying that rule;
4. the agent is able to answer the question: it acts according to a rule and is aware of that, and it is able to justify that rule.

A further possibility, which is not too hard to acknowledge, is that an agent is not acting based on a certain rule, even if it is able to justify acting according to that rule on reflection. So, when discussing the opportunity to endorse a rule of conduct, we can provide reasons in favour or against the rule, without necessarily aligning to its content. This last scenario is particularly interesting to focus on one point: the influence of a moral agent on a community is not limited to the case of

artificial entities that can effectively perform actions autonomously, but also concerns LMs in general, since LMs are both able to share moral opinions and to provide reasons for justifying the incorporation of certain rules or the performance of certain actions.

3. *Why being a Learning Machine makes a difference*

Given this multiplicity of dimensions at play, we now better understand why the topic of AI ethics is not only timely and appropriate to discuss, but also complex and deeply problematic. To address this complexity, let us get back to the distinction between PMs and LMs and see how being a Learning Machine asks for a differential ethical treatment.

An agent with programmed behavior is the realization of an explicit analytical design, typically in some form of algorithm, so that, aside from errors, which are generally corrected when identified, its behavior is specified and, in principle, interpretable and explainable based on that design. And the fact that a program may allow its users to intervene by configuring or personalizing it does not essentially change the situation, since these are already options considered in the design and each configuration results from an explicit action of a user. Since a software system with programmed behavior is merely an executor of instructions established by humans, from an ethical perspective it can be interpreted as a simple intermediary between its developers and its users: once it has been established that users have acted in accordance with the specified terms of use (if, for example, a word processing program is used to copy a text protected by copyright, the cause of the problem is obviously the user, not the developer), it is the developers who are to be held responsible for any ethically problematic behavior of the system (in reference to the first dimension above) and its corresponding effects (second dimension). And since it is the developers, and certainly not the users, who decide the criteria by which the software is written and have reasons for these decisions, the morality of the criteria (third dimension) and of the reasons (fourth dimension) is again evidently the responsibility of the developers. In short, as long as the behavior of an agent is determined by clearly identifiable rules and the origin of those rules can be attributed to specific individuals, it is possible to assign responsibility to those individuals for the actions performed by the agent and for the effects caused by those actions (one may note the analogy with the situation of humans who, in defending themselves against some accusation, claim to have ‘just followed orders’, thus assimilating themselves to entities with

programmed behavior – a situation that can have terrible consequences¹².

The scenario of agents with learned behavior is much more complex, and involves:

- the developers of the model’s architecture (nowadays typically an artificial neural network, and more specifically a Transformer¹³);
- the developers of the training function for the model (in the simplest cases, a program operating via a training set; in more complex cases, one or more other language models, previously trained as teachers);
- those responsible for constructing or selecting the training set used for training;
- those responsible for validating the behavior of the trained model;
- if the model is used within a software application, the developers of that application;
- and finally, the users, either of the model (for example, via a chatbot interface) or of the application that incorporates the model.

Within this sequence of roles, one can recognize a hierarchy of responsibilities, which propagate from the developers of the model’s architecture down to the users. Furthermore, and these are the two elements that specifically distinguish chatbots from generic entities with learned behavior: (i) the rules underlying the behavior of a chatbot are typically not identifiable, since the neural network at the core of the chatbot is enormously complex and these rules are learned during training, with outcomes – in terms of parameter values – that are not fully explainable; and (ii) dialogic interaction with a chatbot is itself conditioned by the moral opinions learned during training, with the result that, in a conversation, opinions may emerge that are justified by reasons autonomously learned by the chatbot. Indeed, chatbots are increasingly capable of evaluating opinions from an ethical standpoint, presenting reasons in support of certain moral positions (fourth dimension above) that vary depending on the topics discussed and the opinions expressed by those interacting with them. On this last point, developers have limited ability to intervene, given the complexity and the intrinsic low transparency of the process by which moral positions are learned.

It is within this highly convoluted framework – of artificial agents that display increasingly sophisticated ethical behavior, that become

¹² As described with precision in Hannah Arendt’s *Banality of Evil* well-known book (Arendt, 1963).

¹³ See Vaswani et al., 2017.

ever more ethically expert, but to which we cannot attribute moral responsibility – that any proposal of ethical guidelines is to be situated.

On this basis, it seems that two basic principles should be followed¹⁴:

- maximizing wellness and freedom for human beings;
- minimizing the risk of harm to human beings.

The tension between these two principles is evident: maximizing wellness and freedom is simpler if one does not worry about the harm this might cause; by contrast, minimizing the risk of harm is simpler if one accepts limitations in achieving wellness and promoting freedom¹⁵. Satisfying both principles is therefore a matter of balance, and the absence of a universally shared ethics imposes even greater responsibility both on the developers and trainers of language models, as well as on all of the users of these systems, who want them to be tools for a better society.

4. *Conclusion*

In light of these reflections, it is wise to acknowledge that ‘everything is new’ as to human-AI interactions, and that “no traditional ethics teaches us the norms of ‘good’ and ‘evil’” to apply, quoting what Hans Jonas wrote almost fifty years ago (Jonas, 1979). We must seek appropriate rules, and perhaps new reasons, to orient ourselves in a world where we may no longer be the only moral actors. To do this, as we have suggested, it is reasonable to assume that we are facing scenarios in which Learning Machines are neither mere instruments devoid of agency nor fully responsible agents with whom we can interact, but alien entities whose impact on our moral life is still to be understood. It is therefore time to explore and experiment, rather than to draw definitive conclusions.

¹⁴ These are indeed the key ethical principles introduced and articulated in the above mentioned OpenAI Model Spec.

¹⁵ A standard literary reference on this subject is in Isaac Asimov’s ‘Laws of Robotics’: “(1) a robot may not injure a human being or, through inaction, allow a human being to come to harm; (2) a robot must obey the orders given it by human beings except where such orders would conflict with the First Law; (3) a robot must protect its own existence as long as such protection does not conflict with the First or Second Law” (Asimov, 1950). Interestingly, these laws aim at minimizing risk, not maximizing wellness and freedom, and this prevents taking some critical ethical dilemmas into account.

References

- ARENDT H. (1963). *Eichmann in Jerusalem: A Report on the Banality of Evil*, Viking Press.
- ASIMOV I. (1950). *I, Robot*, Gnome Press.
- BODDINGTON P. (2023). *AI Ethics. A Textbook*, Springer.
- FRITZ J. (2024). *Deference to opaque systems and morally exemplary decisions*, in “AI and Society”, 40, 5, pp. 3827-3839, <https://doi.org/10.1007/s00146-024-02147-z>.
- GERLICH M. (2025). *AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking*, in “Societies”, 15, 6; <https://doi.org/10.3390/soc15010006>.
- GUDMUNSEN Z. (2025a). *The moral decision machine: a challenge for artificial moral agency based on moral deference*, in “AI and Ethics”, 5, pp. 1033-1045.
- GUDMUNSEN Z. (2025b). *Designing responsible agents*, in “Ethics and Information Technology”, 27, pp. 1-11.
- JONAS H. (1979). *Das Prinzip Verantwortung - Versuch einer Ethik für die technologische Zivilisation*, Frankfurt am Main; English translation: JONAS H. (1984). *The Imperative of Responsibility - In Search of an Ethics for the Technological Age*, University of Chicago Press.
- KESTIN G., MILLER K., KLALES A., MILBOURNE T., PONTI G. (2025), *AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting*, in “Scientific Reports”, 15, 17458, <https://www.nature.com/articles/s41598-025-97652-6>.
- MOLLICK E.R., MOLLICK L. (2023). *Assigning AI: Seven Approaches for Students, with Prompts*. Available at SSRN: <https://www.ssrn.com/abstract=4475995>.
- RUSSELL S., NORVIG P. (2022). *Artificial Intelligence - A modern approach*, Springer (4th ed.).
- SALATINO A., PRÉVEL A., CASPAR E., BUE S.L. (2025). *Influence of AI behavior on human moral decisions, agency, and responsibility*, in “Scientific Reports” 15, 12329, <https://doi.org/10.1038/s41598-025-95587-6>.
- SCHÖLKOPF B. ET AL. (2021). *Toward Causal Representation Learning*, in *Proceedings of the IEEE*, 109, 5, pp. 612-634.
- VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A.N., KAISER L., POLOSUKHIN I. (2017). *Attention Is All You Need*, available at <https://arxiv.org/abs/1706.03762>.

3. Turing, Searle and the Quest for the Meaning in the NLP Era

Ciro De Florio, Aldo Frigerio

ABSTRACT

This contribution examines the epistemological implications of advanced AI systems, specifically Large Language Models (LLMs), treating them as incredible informational machines. The release of ChatGPT in 2022 is identified as a symbolic milestone, comparable to the publication of Alan Turing’s seminal 1950 work, necessitating a revisit of the Turing Test and Searle’s Chinese Room argument. While LLMs are now capable of sustaining human-level conversation and passing variants of the Turing Test, attributing intelligence requires considering the artifact’s design and functioning, as mere third-person behavioral evidence is insufficient. The paper argues that LLMs introduce a novel locational semantic component, representing meaning through multidimensional vectors based on word proximity in training corpora. Outputs have conventional meaning (derived intentionality) but generally lack intentionality, consciousness, or *speaker meaning*. The development of this software exemplifies digital philosophy, reshaping conceptual horizons.

1. *Introduction*

It is difficult to overestimate the impact that Natural Language Processing technologies have had on the debate surrounding Artificial Intelligence. It is not an exaggeration to identify November 2022 as one of the symbolic milestones in the development of AI, alongside other key events such as the publication in “Mind” (1950) of Alan Turing’s seminal *Computing Machinery and Intelligence* – on which we shall return shortly – or the much-cited Dartmouth Conference held six years later (1956), where the term ‘Artificial Intelligence’ was first coined. In November 2022, OpenAI released a free and publicly accessible version of ChatGPT: a software program designed to engage users in linguistic (mostly written) interaction. The naturalness of its linguistic expression, its capacity to comprehend complex texts with nuanced layers of meaning, and the flexibility it demonstrated across a variety of contexts led many observers and users alike to view this moment as a genuine

turning point in the history of technological advancement. For the first time, interaction with a machine revealed features that suggested it was not merely an algorithm capable of generating grammatically well-formed sentences in a given language, but rather software exhibiting actual cognitive – if not intentional – components.

The remarkable media success of LLMs, or Large Language Models – the class of algorithms (or models) developed for chatbots – has led, in the popular press and likely in the minds of most laypersons, to an identification of these systems with artificial intelligence as a whole. In other words, ChatGPT (and its rivals produced by various big tech companies) has come to represent *the* artificial intelligence system. Discussions about the ethical implications of AI or the resources these technologies might provide in specific tasks generally refer – implicitly or explicitly – to chatbots. This is a sociological fact, and as such, it admits of various explanations and interpretations, which exceed the scope of this paper. Nevertheless, two points are worth noting in this regard, as they are relevant to the arguments that will follow. First, the immediate identification of AI software with LLMs is not only manifestly incorrect, but also potentially dangerous in terms of use. Many inappropriate applications of these technologies stem from users' misconceptions about their nature. A case in point is the use of LLMs to retrieve precise factual information (e.g., 'What is the population of Mexico City?') without considering that, if the LLM does not have access to online informational resources, its reliability is low – and even when it does access external repositories, the accuracy of its response is only as high as that of the consulted source (e.g., Wikipedia). In such cases, one is seeking a piece of information that is easily retrievable through any search engine, using software of immense power but designed for very different purposes.

Second, there is, however, a plausible reason for the widespread, initial identification of AI in general with LLMs, and it has to do with the centrality of linguistic interaction in the way human beings conceptualize and describe reality. Let us elaborate on this point. Suppose we encounter an unknown entity (a robot from a distant world or perhaps an exotic animal), and we ask: what proxies might we use to justify the attribution of increasingly complex cognitive properties to it? Certainly, a minimal criterion for considering the entity even slightly intelligent would be its ability to respond to environmental stimuli – an ability that might evolve to include adaptive behavior or even structural reconfiguration to pursue goals such as nourishment, movement, or reproduction. Similarly, another proxy would be the entity's ability to exhibit memory traces of its past interactions – in other words, to recall what it has done, learned, or experienced. Yet there is no doubt that if

we could *communicate* with this entity – especially if such communication occurred in our own language – this would be by far the most compelling proxy for considering it cognitively endowed¹.

Thus, while journalistic sensationalism and marketing hype have certainly played a role in reducing the entire complex and nuanced domain of AI technologies to NLP software, it would be a mistake to dismiss this identification as entirely irrational. As human beings, we look for signs of intelligence in language. And when the linguistic performances of these software programs reach a certain level, our conceptual categories begin to strain. It is precisely on the basis of this intuition that we will, in the next section, revisit the most famous game involving humans and machines: the imitation game.

There is no doubt that the wide-ranging debate on AI includes numerous philosophical issues. From Turing's original, radical question – *Can machines think?* – to concerns of applied ethics (What sort of ethics should govern AI?) and anthropology (Is *Homo Sapiens* destined to be superseded by a technologically evolved descendant?), these questions are formidable not only because they inherit the universality and depth typical of philosophy's great inquiries, but also because they incorporate the full diversity of knowledge domains involved in AI research – a field that is essentially transdisciplinary and multidisciplinary. In this paper, we aim to advance two theses. The first concerns theories of meaning, and the second is of a more meta-theoretical nature. In a nutshell, we intend to argue that LLMs introduce a new semantic component in addition to the *referential* and *inferential* ones (see Marconi, 1999); LLMs inherit the referential component and partially replicate the inferential one. Alongside these, as we shall see, the semantics induced by LLMs involves a third, *locational* component, which – although not exclusive to these technologies – fundamentally characterizes their operation. Secondly, and more broadly, the development of such software exemplifies what David Chalmers (2022) calls *digital philosophy*, that is, a case in which technological progress itself reshapes our very way of doing philosophy. This is not a case of philosophy turning its analytical gaze upon a new object – as it might do with any human domain from subatomic matter to works of art – but rather of innovation generating previously unimaginable conceptual horizons that require novel interpretative efforts. Before arguing for these two theses, however,

¹ Let us consider how, in works of science fiction, one of the narrative strategies used to humanize artifacts (such as machines) and animals is the possibility of communicating with them through speech. Gwaihir, the great eagle in *The Lord of the Rings* who rescues Gandalf, and Asimov's R. Daneel Olivaw, are both characters with whom we converse – and we can hardly avoid considering them as persons.

we shall demonstrate how the emergence of LLMs represents a true plot twist in a longstanding philosophical debate, one that has prominently featured two well-known thought experiments: the Turing Test and Searle's Chinese Room.

2. *The imitation game*

Seventy-five years ago, when electronic computers were as large as rooms and had computational powers that are negligible compared to even the microchips we now wear on our wrists, Alan Turing posed the question of whether machines could think. This is a question that naturally invites the caution typical of philosophers: What exactly do you mean by 'machine'? And what do you mean by 'thinking'? While the concept of a machine admits of a robust and rigorous definition – essentially that provided by the theory of Turing-computability (Turing, 1936) – things are not as straightforward with regard to the concept of thinking. We are all able to provide examples of cognitive activities that fall under a sufficiently broad notion of 'thinking', but the path from there to establishing a set of sufficient conditions for thought is long and fraught with complications.

Turing radically reframed the terms of the issue; he shifted, so to speak, the classic framework of philosophical analysis by inventing a test addressed to engineers and psychologists. In a nutshell, here is how the famous test works. A human being and a machine are placed in two separate rooms. For convenience, let us designate these participants as A and B. A third participant, C – the interrogator –, must eventually reach the conclusion:

- A is a machine
- B is a person

or:

- A is a person
- B is a machine

How does C determine the identity of the two mysterious players? They are permitted to ask any questions of their interlocutors. In the original version of the game as conceived by Turing, the human participant is instructed to assist the interrogator, while the machine attempts to deceive them by pretending to be what it is not.

Turing also introduced a time limit for the interrogation process

(the deception need not last forever), and he specified a threshold for the interrogator's error rate (a certain percentage of incorrect identifications). If the test is conducted with appropriate time constraints and margin of error, a given software program passes the test if it successfully deceives the human interrogator within these parameters.

An immense amount of ink has been spilled on the Turing Test. For our purposes, we wish to highlight the following points, which will be relevant in what follows:

(i) *Behaviourism*. The test is influenced by the intellectual climate of its time, wherein observable output – linguistic behaviour – is taken as a sufficient condition for attributing advanced cognitive capacities (in other words, intelligence) to a machine. The test makes no reference whatsoever to the internal workings of the machine, its architecture, or its inner states.

(ii) *Linguistic interaction*. As previously mentioned, the test is essentially a game of language, of questions and answers. It is important to recall that Turing intended the successful passing of the test as a *sufficient* condition for justifiably attributing intelligence to a machine. It is not a *necessary* condition; it is not difficult to imagine cases of individuals who would probably fail the test (for example, pre-verbal children), yet whom we would still regard as intelligent.

(iii) *Deceptive component*. In Turing's original version, the machine is required to deceive the interrogator. This is not a trivial aspect, for at least two reasons. First, for the machine to deceive, it must possess some representation of truth – so that it may negate it, i.e., lie. This is in principle implementable, but usually presupposes a certain semantic architecture for the representation of knowledge. As we shall see in the case of LLMs, this is a non-trivial technical issue. Second, nearly all existing software systems include ethical filters that typically prevent the machine from *deliberately* deceiving its interlocutor².

(iv) *Variants*. The Turing Test can be simplified by omitting certain specifications. One might, for example, remove the requirement for deception: A and B simply chat, and C observes their conversation and

² This happens even when one queries an LLM about ethically sensitive topics – for example, when the software is asked to provide instructions on how to commit murder. Normally, such systems block this kind of prompt. However, there are 'human' workarounds to obtain an answer. For instance, one might ask the LLM to suggest a murder scenario for a novel we are writing.

must determine who is the human and who is the machine. Naturally, the topic of conversation should not be their respective identities, or the exercise becomes trivial. A yet simpler variant – particularly relevant to the widespread use of LLMs – involves only two participants: a human and a conversational partner. If the human cannot confidently determine whether they are speaking to another human or to a software program, then the machine has passed the test.

The test, as conceived, possesses a compelling pragmatic dimension: it offers a concrete, operational response – within reasonable limits of time and cost – to the question of whether given devices $o_1, \dots, o_n$ are genuinely intelligent. It is therefore unsurprising that philosophical critiques have focused, quite correctly, on its being a *sufficient* condition. In essence, the critique runs as follows: the Turing Test is not a good criterion for determining whether a machine is intelligent because – even if o_j passes the test – there are good reasons to believe that o_j is not, in fact, intelligent, contrary to what the test implies. Since the logical structure of the Turing Test is the conditional “If o_j passes the test, then o_j is intelligent”, challenging the validity of the test consists in constructing a counterexample to this conditional – that is, providing a scenario (at least conceptually possible) in which the antecedent is true (o_j does pass the test), but it is nevertheless implausible that o_j is intelligent.

The most famous attempt in this direction is the widely known thought experiment of the Chinese Room (see Searle, 1980), which will be the subject of the next section.

3. *The Chinese Room*

Let us imagine a person who does not know Chinese is placed in a closed room from which they can communicate with the outside world through two openings: an input slot and an output slot. Sheets bearing Chinese ideograms are passed to the person through the input slot. Although the person inside the room does not know Chinese, they have at their disposal a vast manual consisting of conditional rules, such as:

(R) If 后吏 is written, respond with 名司

or even more complex rules such as:

(R') If 呗叶 is written but 罍喔嚶 was written before, respond with 噉嚶嚶

The number and precise structure of these rules are entirely irrelevant to the thought experiment. The person inside the room faithfully executes the instructions contained in the manual and inserts the responses into the output slot. Outside the room, there is an observer who is a native Chinese speaker and who reads both the inputs and the outputs; to the observer, it appears that the person inside the room knows Chinese: their responses are completely indistinguishable from those of a native speaker.

Now, Searle concludes, it is evident that the person in the room passes the Turing Test for Chinese, and yet, by the very nature of the experiment, they understand absolutely nothing of the language. Thus, even if a computer were, one day (we are in 1980), to pass the Turing Test, it would not follow that such a machine possesses even the slightest degree of intelligence.

The debate surrounding the Chinese Room is extensive; in general, most philosophers find Searle's argument not entirely convincing, though their reasons vary. Above all, the Chinese Room has come to be seen as a gateway to broader philosophical issues associated with cognitive science and artificial intelligence – such as the distinction between syntax and semantics, intentionality, the nature of meaning, and the origins of the mind. In what follows, however, we shall not address those issues³. Rather, we shall examine a variant of the Chinese Room, discussed (albeit in slightly different ways) by Ned Block (1981) and Jack Copeland (1993).

As in Searle's original case, the idea is to construct a counterexample to the Turing Test as a method for determining machine intelligence. In other words, the goal is to imagine a scenario in which a software system passes the Turing Test but is, at least intuitively, not intelligent.

Given that the context is conversational, let us suppose we build a piece of software structured to converse exactly like a chatbot. However, the machine is constructed in the following way. Every conversation is a finite sequence of linguistic expressions (e.g., questions, exclamations, assertions, and so on). Imagine that the machine has an enormous database containing all possible combinations of conversations composed of, say, 50 exchanges in a given language⁴. Many of these will begin with a greeting, others with a question, and so on. The total number of such conversations is astronomically large, but it is reasonable to suppose that it is finite⁵.

³ For an overview, see Cole, 2024.

⁴ Nothing prevents the extension to more languages; but this is not relevant for the thought experiment.

⁵ Technically, it depends on the constraints we place on the construction of character

Now, when the human interrogator begins a conversation with the utterance “Hi, how are you?”, the algorithm simply discards all conversations that do not begin with the string “Hi, how are you?”. Among the remaining conversations – i.e., those whose first line is “Hi, how are you?” – the machine randomly selects one. Let us suppose the reply is: “Fine, thank you. Although I am feeling a bit sad”. At this point, the human asks: “Why are you sad?”, and the algorithm must now discard all conversations that do not begin with the following three utterances: “Hi, how are you?”, “Fine, thank you. Although I am feeling a bit sad.”, “Why are you sad?”. Among the compatible conversations, the algorithm again selects, at random, a possible continuation. And so on.

As Copeland rightly notes, we do not know whether it is technologically feasible to build such a machine, but it does not appear to be conceptually impossible. This supercomputer would easily pass the Turing Test – since the conversation it produces is indistinguishable from that of human beings – and yet no one would claim that it is an example of an intelligent machine. But why not?

To answer this question requires a bit of conceptual engineering concerning our intuitions: on what basis do we ascribe intelligence to an entity? There seem to be at least three families of answers to this question. The first – what we shall call the *behaviourist* option – is Turing’s position. The fact that a machine’s linguistic output is indistinguishable from a human conversation gives us grounds to assert that the software is intelligent. This position runs up against the intuitions underlying Searle’s Chinese Room and Copeland’s supercomputer: both devices produce natural conversation, but intuitively do not appear to be intelligent entities.

At the opposite end of the spectrum from Turing’s position lies what we might call a *solipsistic condition*: the only criterion by which one can ascribe intelligence to an entity is first-person access. That is, it is the fact that *I* seem to myself to be intelligent that allows me to attribute to myself the property of intelligence. Since I do not have first-person access to the qualitative states of consciousness of the person next to me, I have no reason to ascribe intelligence to them. It is clear, however, that this position is ultimately untenable, as it collapses almost immediately into a form of scepticism. This is the well-known problem of other minds: what guarantee do we really have that the people around us have minds inhabited by mental, conscious, qualitative states similar to our own?⁶

strings. The point is that meaningful sentences in a language have a finite length – and not even a very long one – otherwise speakers would be unable to grasp their meaning.

⁶ On the problem of other minds, see Avramides, 2001.

Between these two extremes lies a more moderate family of responses in which linguistic output is essential but appears to be insufficient; some further characteristic is required in order to ascribe to X the status of an intelligent entity. If we have naturalistic inclinations – as Searle, for instance, does – we may claim that intelligence is characterised by the causal powers of the central nervous system. In this sense, what the Chinese Room refutes (if indeed it does) is the possibility of achieving intentionality and intelligence via computational means.

It seems to us that within this third family of responses there is room for a different argumentative strategy. On what basis do we ascribe intelligence? On the basis of output (which is the foundation of the Turing Test), but also on the basis of certain considerations regarding the *design* and *functioning* of the artefact in question. Since software systems are very different kinds of things from people, we need to understand how they are built in order to judge whether what they produce can reasonably be regarded as intelligent.

Now, it is quite plausible to maintain that, in the case of the Copeland-Block supercomputer, the reason we do not describe it as intelligent is because it is constructed in a thoroughly unintelligent manner. It is a vast, boundless inventory of conversations, selected by means of a trivial random procedure. No one actually converses this way, no matter how plausible the resulting output may be.

Our first conclusion, therefore, is that the Turing Test and its counterexamples reveal the importance of the conditions under which we ascribe intelligence. Mere third-person behavioural evidence appears to be insufficient. It is also our understanding of how these software systems function that influences whether we are inclined to regard them as intelligent or not.

The crucial point, however, is that until only a few years ago, the Chinese Room and the supercomputer were thought experiments – imaginary scenarios, conceptually plausible at best – employed to probe our intuitions. Today, quite astonishingly, that is no longer the case: machines are now capable of sustaining human-level conversations. If we slightly relax the criteria – especially by adopting some of the variants of the Turing Test – we may say that many software systems today pass the Turing Test. But then, are they intelligent? Do they possess semantics? Do they, in fact, exhibit a certain degree of intentionality?

In light of the above, to answer these questions critically we must examine how this particular class of software – large language models – actually works.

4. *How LLMs represent meaning*

The meanings of words (or word components) are represented by LLMs as multidimensional vectors. The guiding idea behind the construction of these vectors is: “Tell me who you associate with, and I will tell you who you are”. More precisely, two words that frequently occur together will have similar or at least closely related meanings. This idea can be implemented as follows: let us suppose we have access to a large corpus of texts. For a given word p_1 , we can calculate its proximity to another word p_2 in the corpus in the following way: we take all the sentences in which both p_1 and p_2 appear and count how many words separate p_1 from p_2 in each of those sentences. For example, if in a certain sentence three words separate p_1 and p_2 , then their distance in that sentence is, indeed, three. By averaging all these distances, we obtain the overall distance between p_1 and p_2 across the entire corpus. We may ignore the sentences in which neither word appears, while those in which only one of the two occurs can be treated with different techniques: we may either discard them or assign them an arbitrarily high distance value (e.g., 100).

Now suppose we apply the same procedure to calculate the distance between p_1 and p_3 , and so on for every other word. In this way, we obtain the distance of p_1 from all other words in the corpus. These numbers can be interpreted as a multidimensional vector, where the distance of p_1 from each other word represents a coordinate of the vector in a multidimensional space. In other words, the distances between p_1 and all the other words can be represented as a one-column matrix of dimensions $1 \times n$, where n is the number of such distances. And it is well known that such a matrix can be interpreted as a vector in a n -dimensional space.

We can apply the same procedure to calculate the distance between p_2 and each other word in the corpus and thus obtain another n -dimensional vector – and so on for all words. The idea is that the more similar the vectors of two words are, the more similar their meanings will be. Technically, this distance can be computed by calculating the cosine of the angle between the two vectors. When the value is close to 1, the two vectors ‘point’ almost in the same direction and are therefore close to each other. When the value is near 0, the two vectors are orthogonal and thus independent. When the value is close to -1 , the vectors point in opposite directions. This may indicate a certain relation between the meanings of the words in question (for example, opposite or contradictory meanings).

We have illustrated this procedure for calculating the distance between the meanings of two words because it is the most intuitive

method. However, it is not the procedure currently employed, in part because it would be extremely costly in computational terms. To generate the vectors associated with words, shallow neural networks are now used, typically with only one or two hidden layers. For example, in *Word2Vec* – one of the most widely known tools for associating vectors with words – the neural network takes as input a $1 \times n$ matrix of numbers in which all rows are 0 except one, which is 1. It outputs a similar matrix. n corresponds to the size of the vocabulary, so each word in the vocabulary is associated with a distinct matrix in which the 1 is placed in a different row. The neural network is trained on a large corpus to predict, given a word, which word is most likely to follow. Training takes place, as is standard, via backpropagation. In the hidden layer of the network, each number in the $1 \times n$ matrix is associated with a set of weights, which are summed to predict the next word. Once the network has been trained, the weights associated with each row are taken as the components of the vector for that word: if the word p_1 has a 1 in the first row and 0s in all the others, then the weights assigned to the first row by the trained neural network will be the components of p_1 's vector.

The number of weights associated with a word may vary depending on the architecture of the network and the choices of the developers. For instance, in ChatGPT-3 – the latest version of ChatGPT for which details are publicly available – the vocabulary consists of 50,257 words, and each is represented by a vector with 12,888 dimensions. Therefore, the input of ChatGPT-3 is a matrix with dimensions $n \times 12,888$, where n is the number of words in the prompt. The maximum value for n is 2,048: this is the attention window of ChatGPT-3. This matrix constitutes the input to ChatGPT-3's attention heads – that is, the initial layers of the neural network, whose task is to predict the next word given the words in the prompt.

5. *Semantic competence*

Beyond technical details, the meaning of each word in LLMs is represented as a position in a multidimensional semantic space. The closer two words are in this semantic space, the closer their meanings are, and the more likely they are to belong to the same semantic field. As mentioned earlier, semantic proximity or distance within the semantic space is determined by the position of words in the training corpus. Words that frequently appear close to each other in many sentences within the corpus are considered to have similar meanings.

Let us return to the questions raised at the beginning of this article.

Given the internal functioning of LLMs, can we say that they think? That they are intelligent? As stated earlier, these kinds of questions involve an intrinsic vagueness that prevents a fully satisfactory answer. More modestly, we shall attempt to answer simpler questions regarding the abilities we can attribute to LLMs. Some of these abilities may well fall under what we commonly call intelligence, even if they are not sufficient for it. Intelligence is a complex phenomenon, and it is ascribable by degrees. It is therefore useful to divide the question “Is *x* intelligent?” into simpler sub-questions.

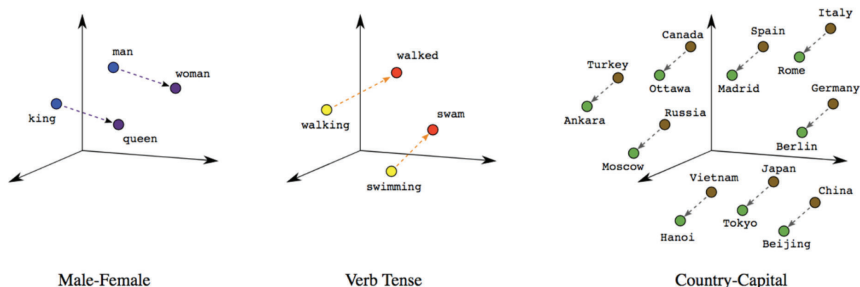
A first issue is whether we can attribute to LLMs any kind of semantic competence. If LLMs were merely stochastic parrots, the answer would obviously be negative. The same would hold if LLMs merely repeated sequences of words already present in the training corpus. However, we believe this is not the case, and that a certain kind of semantic competence can indeed be attributed to LLMs. Let us explain why.

First of all, semantic competence is a complex ability, composed of several distinct sub-abilities. Diego Marconi has argued that lexical competence consists of two major components, which he calls *referential competence* and *inferential competence* (Marconi, 1999). Referential competence is the ability to associate words with objects, events, or facts in the world. Inferential competence, on the other hand, is the ability to make inferences using words and to understand the semantic relations between words (for example, synonymy, antonymy, hyperonymy, etc.). With respect to these relations, Marconi primarily had in mind what formal ontologies attempt to capture, and what is often implemented in symbolic AI systems. As we have seen, LLMs possess neither of these two components: they have no access to the world, and they are not based on symbolic AI. Nevertheless, they do possess a form of inferential competence – different from the one Marconi envisioned – that consists in representing multiple relations between word meanings. Since this component derives from the positions that words occupy in the training corpus, we may refer to it as *locational*.

The fact that LLMs use a space with so many dimensions to capture the meaning of words allows them to represent multiple relations between words.

As shown in Figure 1, if one calculates the distance along a certain direction between *man* and *woman* in the semantic space of a LLM, one finds that *king* and *queen* lie at roughly the same distance along that same direction. It is therefore plausible to suppose that this dimension of semantic space is dedicated to male-female relationships. Other dimensions are dedicated to relationships between various verb forms, or between a country and its capital city.

Figure 1



It has also been shown that LLMs are capable of generalisations that go beyond their training data. This can be verified through a careful separation between training data and test data (train-test split). In such experiments, the data on which LLMs are trained is rigorously separated from the data used to test them. In this way, one can confirm that the LLM does not pass the test by retrieving information from its training data, but rather because it has performed generalisations on that data – i.e., it has identified patterns within the data – that it is then able to apply to new cases. For example, experiments have been conducted to assess whether LLMs are capable of capturing *compositionality* in language – that is, the fact that the meaning of phrases and sentences depends on the meanings of their constituent units (words and morphemes). In general, a language is compositional when the meaning of larger units depends on that of the smaller ones. Compositionality is a fundamental feature of natural language, which accounts for its *productivity* (i.e., the ability to generate novel utterances with the justified expectation that they will be understood by interlocutors even if never previously encountered) and its *systematicity* (i.e., the existence of predictable and structured patterns among the utterances we understand. If one understands “Paul loves Anna”, one will also likely understand “Anna loves Paul”). Experiments using train-test splits appear to confirm that LLMs are capable of generalising over their training data by learning the rules that govern the compositionality of natural language (Millière and Buckner 2024a, sec. 3.1).

Moreover, the attention heads in LLMs appear to be specifically designed to handle compositionality. This component of the model identifies which words are *relevant* to others – and to what degree – through two parameter matrices, the *query matrix* and the *key matrix*. Then it multiplies the vector associated with the influencing word by a

third matrix (the *value matrix*) and by the probability that that word influences another. The result is then added to the vector of the word being influenced. This process can be interpreted as follows: the LLM determines which word meanings are to be composed, and then it updates the ‘meaning’ of the influenced word (i.e., its position in semantic space) based on the meaning of the word with which it is being composed.

It therefore seems difficult to deny that LLMs possess some form of semantic competence. Another question, however, is whether such semantic competence is the same as human semantic competence. Human competence – at least at the conscious and symbolic level – seems centred on the two components described by Marconi: referential and inferential. It would thus appear to be a form of competence that differs from that of LLMs. Whether, at the neuronal level, structures similar to those in LLMs exist in the human brain is a different question altogether, and research on this front is still at an early stage, despite some attempts (see Whittington et al., 2022).

6. *Conventional meaning and speaker meaning*

Another question is whether we can attribute *meaning* to the outputs produced by LLMs. This issue has been addressed by Borg (2025), and we are in substantial agreement with her position. The words of a language have a conventional meaning, determined by the rules established by a given community of speakers. In this context, we speak of a *linguistic code*, where the code is a set of rules associating signs with meanings. The outputs produced by LLMs thus have at least a conventional meaning, based on the conventional meanings of the words used by the model and on their semantic composition. Borg refers to this as *derived intentionality*: the meaning of words is determined by the intentions of the members of a linguistic community, and the productions of LLMs are endowed with meaning in virtue of that intentionality. It is worth remembering that the immense corpora used to train LLMs consist of texts that possess meaning and that were (in the vast majority of cases) composed by human beings.

A further issue is whether we can attribute to LLM outputs not only conventional meaning but also *speaker meaning*. This refers to the distinction proposed by Grice between what he calls *timeless meaning* (which may be identified with conventional meaning) and *utterer’s meaning* (Grice, 1989, ch. 6). Utterer’s meaning is what the speaker intends to communicate by using certain words. It may coincide with conventional meaning but can also diverge from it – for example, if the

speaker uses words metaphorically, figuratively, or to convey implicatures and non-literal meanings in general. It is much more difficult to attribute this type of meaning to LLMs, as it implies *intentionality*: it is the meaning the speaker *intends* to communicate, and Grice maintains that this meaning is understood precisely when the hearer recognises the speaker's communicative intention.

Attributing intentionality to LLMs is a delicate matter, and the prevailing view tends to be negative. If we cannot ascribe intentionality, then we cannot attribute to the productions of LLMs speaker meaning either.

7. *Speech acts, consciousness, and individuality*

Another important question is whether LLMs perform *speech acts* (Austin, 1962; Searle, 1969), such as making assertions, asking questions, and so forth. This issue, too, depends on whether we can attribute *intentionality* to LLMs. Speech acts are actions – things we do by means of language. Actions imply intentionality: the agent of an action must intend to perform it. If we do something unintentionally, we cannot properly say that we have *performed* an action. Therefore, if LLMs lack intentionality, they do not genuinely perform speech acts.

A clarification is necessary at this point, because LLMs *seem* to perform assertions when they write that the capital of Italy is Rome, and they *seem* to ask questions when they produce sentences such as “Would you like me to proceed with more details?”. It is essential here to distinguish between *linguistic forms* that are apt to perform a certain speech act, and the *speech act itself*. Certainly, a word like ‘thank you’ is suited to perform a particular type of speech act (in Searle’s classification, an *expressive* speech act, whereby the speaker expresses gratitude to their interlocutor). Therefore, a speaker who uses the word ‘thank you’ outside of quotation contexts or reported speech will, in most cases, perform the speech act of expressing gratitude through the use of that word.

However, not every instance of the word ‘thank you’ constitutes an expressive act. Suppose a monkey, by randomly pressing the keys of a computer, types the word ‘thank you’ on the screen, and suppose further that – again by chance – it sends that output to someone via an email program. We would not say that the monkey, despite having written and sent the word ‘thank you’, has performed an expressive act, because it did not intend to do so. Hence, the fact that LLMs produce sequences of words that, when used intentionally, typically serve to

perform certain expressive acts, does not suffice to conclude that they are themselves performing those acts.

The lack of intentionality in LLMs is also reflected in other characteristics they appear to lack. For instance, they seem to lack *consciousness* (cf. Milli re and Buckner, 2024b; Rapaport, 2025), as well as what we might call *selfhood* or *individuality*. This latter feature is usually ascribed to a subject who possesses an extensive individual memory of facts and events they have experienced, of beliefs formed over time, and of general dispositions or orientations that allow them to make long-term plans, to entertain high-level and abstract intentions. None of these capacities appear to be present, at least for now, in current LLMs.

8. Conclusion

The philosophical questions raised by Turing in 1950 have found renewed vitality with the recent emergence of large language models. This is an extraordinarily fertile ground, where philosophical inquiry has greatly benefited from technological advances. As we have attempted to show throughout these pages, Turing’s questions – *Can machines be intelligent? Can they think?* – can be unpacked into simpler questions about specific abilities, for which more precise answers can be given. We have attempted to respond to some of these questions; others remain open. Philosophical reflection on this topic is only beginning and, in some respects, still in its infancy, given that the remarkable capabilities of LLMs have taken even philosophers by surprise. Nonetheless, it is a field that will undoubtedly be cultivated with increasing depth in the years to come.

As we noted earlier, this appears to us to be a paradigmatic case of *digital philosophy*: that is, a significant contribution of technological innovation to our interpretative theories about reality. The locational component of LLMs is a striking example. Without this specific technology, many of the conceptual insights presented here would not have emerged, or would still remain in a merely conjectural state. This suggests a distinctly interdisciplinary direction for the future of research in both philosophy and artificial intelligence.

References

- AUSTIN J.L. (1962). *How to Do Things with Words*, Cambridge, MA, Harvard University Press.
- AVRAMIDES A. (2001). *Other Minds. The Problems of Philosophy*, London, Routledge.

- BLOCK N. (1981). *Psychologism and Behaviorism*, in “The Philosophical Review”, 90(1), pp. 5-43.
- BORG E. (2025). *LLMs, Turing Tests and Chinese Rooms: The Prospects for Meaning in Large Language Models*, in <https://doi.org/10.1080/0020174X.2024.2446241>.
- CHALMERS D.J. (2022). *Reality+: Virtual Worlds and the Problems of Philosophy*, New York, W.W. Norton & Company.
- COLE D. (2024). *The Chinese Room Argument*, in *The Stanford Encyclopedia of Philosophy*, ed. by E.N. Zalta and U. Nodelman (Winter 2024 Edition) <https://plato.stanford.edu/archives/win2024/entries/chinese-room/>.
- COPELAND J. (1993). *Artificial Intelligence: A Philosophical Introduction*, Chichester, John Wiley & Sons.
- GRICE H.P. (1989). *Studies in the Way of Words*, Cambridge, MA, Harvard University Press.
- MARCONI D. (1999). *La competenza lessicale*, Roma-Bari, Laterza.
- MILLIÈRE R., BUCKNER C. (2024a). *A Philosophical Introduction to Language Models – Part I: Continuity with Classic Debates*, in “arXiv Preprint”. <https://arxiv.org/abs/2401.03910>.
- MILLIÈRE R., BUCKNER C. (2024b). *A Philosophical Introduction to Language Models – Part II: The Way Forward*, in “arXiv Preprint”. <https://arxiv.org/abs/2405.03207>.
- RAPAPORT W.J. (2025). *Large Language Models and the Turing Test: The ‘Use of Words’ vs. ‘General Educated Opinion’*, Manuscript. <https://cse.buffalo.edu/~rapaport/Papers/llmtt.pdf>.
- SEARLE J.R. (1969). *Speech Acts: An Essay in the Philosophy of Language*, Cambridge, Cambridge University Press.
- SEARLE J.R. (1980). *Minds, Brains and Programs*, in “Behavioral and Brain Sciences”, 3, pp. 417-457.
- TURING A.M. (1936). *On Computable Numbers, with an Application to the Entscheidungsproblem*, in *Proceedings of the London Mathematical Society*, s2-42 (1), pp. 230-265.
- TURING A.M. (1950). *Computing Machinery and Intelligence*, in “Mind”, 59, pp. 433-460.
- WHITTINGTON J.C.R., WARREN J., BEHRENS T.E.J. (2022). *Relating Transformers to Models and Neural Representations of the Hippocampal Formation*, in “arXiv Preprint”. <https://arxiv.org/abs/2112.04035>.

4. Psychomatics: A Multidisciplinary Framework for Understanding Artificial Minds*

*Giuseppe Riva, Fabrizia Mantovani,
Brenda K. Wiederhold, Antonella Marchetti, Andrea Gaggioli*

ABSTRACT

Although large language models (LLMs) and other artificial intelligence systems demonstrate cognitive skills similar to humans, such as concept learning and language acquisition, the way they process information fundamentally differs from biological cognition. To better understand these differences, this article introduces Psychomatics, a multidisciplinary framework bridging cognitive science, linguistics, and computer science. It aims to delve deeper into the high-level functioning of LLMs, focusing specifically on how LLMs acquire, learn, remember, and use information to produce their outputs. To achieve this goal, Psychomatics will rely on a comparative methodology, starting from a theory-driven research question – is the process of language development and use different in humans and LLMs? – drawing parallels between LLMs and biological systems. Our analysis shows how LLMs can map and manipulate complex linguistic patterns in their training data. Moreover, LLMs can follow Grice's Cooperative principle to provide relevant and informative responses. However, human cognition draws from multiple sources of meaning, including experiential, emotional, and imaginative facets, which transcend mere language processing and are rooted in our social and developmental trajectories. Moreover, current LLMs lack physical embodiment, reducing their ability to make sense of the intricate interplay between perception, action, and cognition that shapes human understanding and expression. Ultimately, Psychomatics holds the potential to yield transformative insights into the nature of language, cognition, and intelligence, both artificial and biological. Moreover, by drawing parallels between LLMs and human cognitive processes, Psychomatics can inform the development of more robust and human-like artificial intelligence systems.

1. Introduction: defining Psychomatics

In the rapidly evolving landscape of artificial intelligence - AI (Dakanalis,

* This chapter was originally published as Riva G., Mantovani F., Wiederhold B.K., Marchetti A., Gaggioli A. (2025). *Psychomatics - A Multidisciplinary Framework for Understanding Artificial Minds*, in "Cyberpsychology, Behavior, and Social Networking", 28(7), pp. 515-523. doi:10.1089/cyber.2024.0409. © Mary Ann Liebert, Inc.

Wiederhold, Riva, 2024; Ferruz et al., 2024; NHB, 2023; Wiederhold, 2024), there exists an ongoing debate about how we should evaluate and understand the cognitive capacities of AI systems (Hutson, 2024; Nature, 2023). Traditional approaches have often centered on either using automated evaluations based on specific benchmarks, or comparing AI behaviors with those of humans, aiming to identify similarities and differences (Aharoni et al., 2024; Hubert, Awa, Zabelina, 2024; Zhao, Huang, Seligman, Peng, 2024).

For example, Flan-PaLM, an instruction-tuned variant of the Pathways Language Model (PaLM) achieved 67.6% accuracy on MedQA, the US Medical Licensing Exam-style questions (Singhal et al., 2023). Another study (Lake, Baroni, 2023) showed that a Large Linguistic Model (LLM) achieved human-like performances in a set of compositional tasks, demonstrating systematic compositionality – the algebraic ability to understand and produce novel combinations from known components. Finally, Hagendorff, Fabi and Kosinski (2023) found that human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. This LLM responded correctly to a great majority of tasks, decisively outperforming humans in their ability to avoid traps embedded in the tasks.

Several other studies have been conducted to compare the behavior of AI and humans, identifying both similarities and differences. For example, Mei et al. (2024) used conventional behavioral measures from economics and psychology to evaluate the behavioral and personality traits of large language models (LLMs). The study found that ChatGPT-4 exhibited traits that were indistinguishable from those of a random human among tens of thousands of human participants. Based on these findings, Meng (2024) signals the emergence of a new research area called ‘AI Behavioral Science’ (AIBS), which uses methods from human behavioral science to examine and design AI behavior.

However, this emerging field does not provide a deeper insight into AI (Hicks, Humphries, Slater, 2024), although it does demonstrate that AI has advanced information manipulation capabilities. As noted recently by Heaven (2024): “The largest models, and large language models in particular, seem to behave in ways textbook math says they shouldn’t. This highlights a remarkable fact about deep learning, the fundamental technology behind today’s AI boom: for all its runaway success, nobody knows exactly how – or why – it works”.

We argue that this shortcoming arises because the AIBS overlooks two critical aspects. First, it fails to clarify how the ability of LLMs to use information to generate outcomes compares to or differs from that of humans (Shiffrin, Mitchell, 2023). LLMs’ abilities may differ from human processes. If so, using LLMs as human subjects in psychology

experiments is useless. Secondly, like behaviorism in psychology and the imitation test proposed by Alan Turing, AIBS concentrates only on the observable aspects of behavior and their environmental determinants. While LLMs can adeptly pass the Turing test by emulating human-like responses (Biever, 2023), the adage “doing the same thing for different reasons” underscores that merely replicating external behaviors is inadequate for providing a coherent and comprehensive account of the emergence and functioning of cognitive abilities (Shteynberg et al., 2024). Without exploring the intrinsic processes that underlie these abilities, we risk falling into the trap of ‘conceptual borrowing’ (Floridi, Nobre, 2024), where we anthropomorphize AI systems as computational analogues of the human brain and ascribe psychological properties to them without a firm theoretical or empirical foundation (Ragin, 2014).

In this perspective, we propose a multidisciplinary synergy among cognitive science, linguistics, and computer science to uncover groundbreaking insights into large language models (LLMs). We present ‘Psychomatics’ (the crasis of the terms ‘psychology’ and ‘informatics’) an interdisciplinary framework that connects these fields, focusing on how artificial intelligence (AI), and in particular LLMs, processes information. This framework explores how LLMs perceive, learn, remember, and employ information to generate outputs. To achieve this goal, Psychomatics will rely on a comparative methodology, drawing parallels between LLMs and biological systems.

In Psychomatics, the comparative approach and its methodological application is conducted through a theory-driven research question (Marchetti, Di Dio, Cangelosi, Manzi, Massaro, 2023): is the process of language development and use different in humans and LLMs? This research question is formulated as a point of departure for comparative investigation, enabling reflection on what, when, and how to compare and for what purpose. Furthermore, a set of rules are used to direct the research strategy, aiming at explanations rather than a more or less complete description of specific phenomena by comparing them across systems. Hence, instead of focusing solely on ‘behaviors’, as done by AIBS, the point of departure for our approach is (Pennings, Keman, Kleinnijenhuis, 2006):

- developing systematic knowledge about language and its functioning (i.e., syntax and semantics) that transcends mere description and allows for generalizations (i.e., external validity);
- deriving answers to questions (i.e., what is language for?) based on existing theory or, if possible, plausible hypotheses (i.e., theory guidance);
- striving for precise information and comparable indicators (i.e.,

meanings, intentions, etc.) that are reliable and open to replication (i.e., internal validity).

By adhering to these principles, Psychomatics aims to establish a rigorous, theory-driven comparative framework that leverages the complementary strengths of cognitive science, linguistics, and computer science.

2. *Information processing: Syntax and Semantics*

Information processing refers to the manipulation of symbols or data representations stored in memory according to specified rules. From this perspective, AI operates by applying formal rules and procedures to transform symbols held in their memory structures. However, as humans intuitively understand, not all information is alike. Effective information processors must manipulate meaningful representations – symbols that refer to real-world entities or concepts. To clearly distinguish the formal symbolic operations from the semantic meanings they convey, we borrow two concepts from linguistics (Dawson, 2022): syntax and semantics (see Table 1 for a short description).

Table 1 - *Syntax and Semantics*

Syntax refers to the grammatical structure of a sentence and the rules for determining whether a sentence is well-formed. The rules governing permissible symbol transformations in a formal system (e.g., chess) are analogous to syntax. In contrast, semantics concerns the meanings expressed by sentences or symbols. A symbol is considered meaningful, or semantic, if it represents or refers to something in the world (real or possible). For example, the word 'dog' is meaningful because it denotes a particular type of animal.

This separation of syntax from semantics enables mechanical information processing. Information processors can formally manipulate symbols using syntactic operations without requiring comprehension of the symbols' meanings. Crucially though, these formal operations can nonetheless preserve and transform meanings. For instance, in mathematics, the rule allowing ' $x + x$ ' to be replaced by ' $2x$ ' is purely syntactic – it only requires recognizing the shapes of symbols like ' x ' and ' $+$ ' to apply. However, this syntactic rule also respects the semantic interpretation: whatever real-world quantity ' x ' represents, adding it to itself yields twice that value. Similarly, syntactic rules can connect and transform meanings. For example, if 'Extra Small' is understood as a size minor to 'Small', which is minor to 'Medium', which is minor to 'Large', then a small apple can be inferred to have a size smaller than a large one through the application of syntactic operations on these relational concepts.

When AI processes information, it follows specific rules about how to structure sentences (called syntax). These rules help the AI understand how words relate to one another in a sentence, like which words are subjects or verbs. But for the AI to really understand a sentence, it also needs to recognize the deeper connections between words (called semantics). For instance, in the sentence ‘The cat chased the mouse’, syntax helps the AI identify who is doing the chasing and who is being chased. But this is not enough. It also needs to understand what individual words represent. For instance, the word ‘cat’ refers to a specific type of domesticated mammal. In other words, the AI must have carefully designed rules that reflect the true relationships between words and their meanings (or what they stand for). This makes sure the AI can manipulate information accurately.

While artificial minds do not innately ‘understand’ meaning, their formal operations can indirectly capture, transform, and reason over meanings generated by humans by virtue of being properly coupled to the systematicity of syntax-semantics mappings. In other words, we suggest that the ‘intelligence’ of LLM is the result of their ability of creating extensive syntax-semantics maps (Singhal et al., 2023). But how is this possible? The answer lies in the peculiar characteristics of language.

3. The evolutive goal of language: moving away from experience

Language presents a fundamental challenge in facilitating precise communication between individuals from diverse backgrounds discussing a wide range of topics that vary in subjectivity and objectivity. People often need to convey personal perceptions, emotions, thoughts, aspirations, and more. Moreover, this puzzle of meaning is further complicated by the inherent diversity across languages. No two languages share perfectly isomorphic semantics. Each has a unique lexicon – a distinct set of words that conveys specific meanings shaped by the culture. Furthermore, languages evolve independently over time, coining new words, repurposing old ones with novel meanings, and continually reshaping their semantic landscapes.

However, individuals can solve these issues and can communicate effectively using language. How? As suggested by the Israeli linguist Daniel Dor (2015), language is a unique system that transcends mere sharing of experiences. It enables speakers to guide their interlocutors intentionally and systematically through the process of imagining an intended experience – rather than directly experiencing it themselves.

The speaker provides the receiver with a symbolic code, a blueprint sketching out the core characteristics of the experience. The receiver then uses this scaffolding to reconstruct and recombine memories of past experiences, producing a novel, imagined experience aligned with the conveyed meaning (Barsalou, 2009, 2013). This capability to systematically guide imaginative simulation distinguishes language from other modes of communication that directly share signals or stimuli evoking predefined responses or impressions. Language is thus the only system that allows for communication that bridges the experiential gaps between speakers, serving as an effective means for sharing cultural information across generations and societies. This interpretation aligns with a recent proposal by Fedorenko and colleagues (2024), challenging the idea that language itself is the root of human cognitive complexity. Instead of creating human cognitive sophistication, language acts more as a mirror, reflecting the advanced nature of human thinking.

To coordinate the open-ended generative process of recombining and modulating memories into a novel coherent experience approximating the intended meaning, language has also to embed in its code the meaning of the experiences it aims to represent. A clear example is a vocabulary: it encodes and describes all existing and potential realities through its unique system of symbolic representations. Within the confines of vocabulary, there is no inherent reality – only a meticulously crafted mapping of concepts over a linguistic realm. As noted by the structural linguistics pioneer Ferdinand de Saussure, to achieve this goal each linguistic sign (such as a word), has a ‘value’ that derives from its relationships and contrasts with other signs within the same language system. In his own words (de Saussure, 1966):

Language is a system of interdependent terms in which the value of each term results solely from the simultaneous presence of the others (p. 114) ... But it is quite clear that the concept is nothing, that is only a value determined by its relations with other similar values, and that without them the signification would not exist (p.117).

In Saussurean linguistics, the meaning of a word is not defined in isolation but is shaped by its position within the broader linguistic structure through contrasts with other words (structural semantics). For example, the word ‘tree’ derives its meaning through its differences from similar terms like ‘bush’ or ‘plant’. Ferdinand de Saussure identified two types of relationships between linguistic terms that shape meaning: syntagmatic and associative (paradigmatic) relations (Table 2).

Table 2 - *Syntagmatic and Associative Relations in Language.*
Together, these relationships enrich language comprehension,
allowing us to bridge knowledge gaps and establish shared
understanding even when discussing abstract or unfamiliar concepts

Syntagmatic relations refer to how words or linguistic elements combine in a linear sequence to create coherent phrases, clauses, or sentences. This involves not only the order of words but also their grammatical compatibility and adherence to syntactic rules. Words are positioned relative to neighboring elements, forming well-structured linguistic units. For instance, in the sentence 'the black cat purrs', the words have a syntagmatic relation because their sequence and grammatical fit ('purrs' following 'cat') convey an intended meaning.

Associative relations describe the conceptual connections between words based on similarities, oppositions, or other cognitive links in the minds of speakers. Unlike syntagmatic relations, they are not dependent on linear positioning within a sentence. Instead, they arise from mental associations created by shared meaning components, antonyms, and other cognitive links. For example, the word 'apple' is associated with 'fruit', 'red', 'sweet', and 'tree', which are conceptual associations rather than syntactic dependencies.

4. *Transformers algorithms and their ability in structuring language*

In the field of machine learning, the concept of 'attention', introduced by Vaswani and colleagues in 2017 (Vaswani et al., 2017), serves as a core mechanism within Transformer networks, which have become foundational to large language models. Attention in Transformer algorithms enables models to dynamically focus on specific aspects of the input data while performing tasks like translation or question answering. It achieves this by computing attention weights that determine the importance of each input element for the task at hand.

Unlike previous models that process input linearly and struggle with long-range dependencies, attention allows the algorithm to evaluate the significance of each data part, irrespective of its position. Moreover, Transformer architectures distinguish between self-attention and cross-attention mechanisms, which conceptually mirror syntagmatic and associative relationships in semantics, respectively.

As we have just seen, syntagmatic relations in semantics define the meaning of linguistic elements through their linear combination and adherence to syntactic rules. Similarly, the self-attention mechanism in Transformers captures dependencies between words within a single sequence by calculating weighted associations between each word and every other word. These attention weights reveal how each word's interpretation depends on its relationship to others based on position and structure. Thus, self-attention models syntagmatic meaning

through structural rules, much like how syntagmatic relations operate in natural language semantics.

On the other hand, associative relations in semantics represent broader conceptual links and cognitive connections that words form in the minds of speakers. This aligns with the cross-attention mechanism in sequence-to-sequence Transformer models, such as encoders and decoders. Cross-attention maps one sequence (e.g., input text) to another (e.g., output text) by identifying associations between each element of the two sequences. This mechanism learns cross-sequence relationships that transcend linear structure, much like associative relations link concepts in mental lexicons.

In essence, Transformer models combine the complementary mechanisms of self-attentive syntagmatic processing and cross-attentive associative processing as computational analogs to dual pathways of meaning derivation outlined in structural semantics theory. This hybrid approach, integrating linear and non-linear processing, allows Transformers to represent linguistic signs embedded in training data comprehensively. Self-attentive syntagmatic processing enables compositional interpretation, while cross-attentive associative processing contextualizes and integrates semantic knowledge dynamically. Together, these mechanisms empower transformer-based language models to represent and operationalize the complex facets of human language meaning and reasoning.

5. *The differences between humans and LLMs*

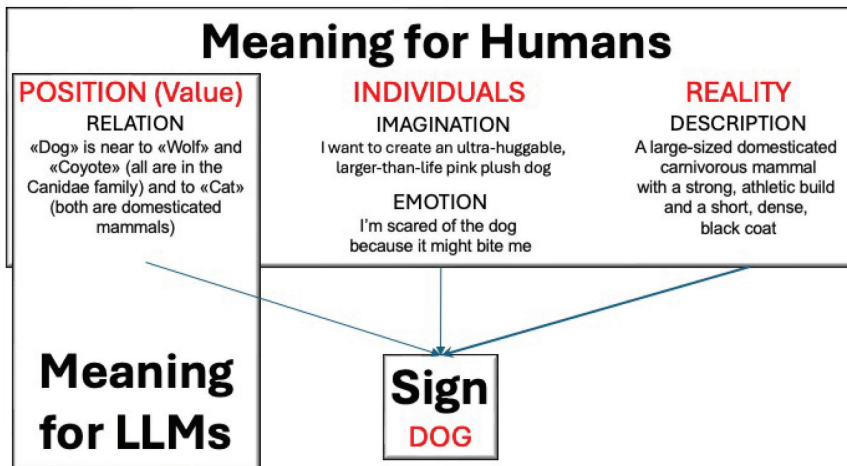
While LLMs demonstrate impressive language capabilities, they lack the social and relational aspects that underpin human communication (Bolhuis, Crain, Fong, Moro, 2024). In de Saussure's view, language: "...is both a social product of the faculty of speech and a collection of necessary conventions adopted by a social body to enable individuals to exercise that faculty" (p. 9).

As underlined by Rizzolatti and Arbib (1998), and remembered recently by Chemero (2023), language and communication are deeply rooted in the embodied experience of existing in a physical form. Consequently, LLMs cannot acquire knowledge and skills in the way humans do, which is fundamentally shaped by personal experiences and biological maturation in a physical body. Biological mechanisms like mirror neurons (Rizzolatti, Arbib, 1998) and inter-brain synchronization (Jiang et al., 2012; Kelsen, Sumich, Kasabov, Liang, Wang, 2022) play a crucial role in supporting the analysis of implicit and/or multimodal cues in humans, facilitating the seamless interpretation of

nonliteral language. Unfortunately, current LLMs lack physical embodiment, reducing their ability to make sense of the full spectrum of human communication. To overcome this limitation researches are using LLMs to train embodied robots which can interact with their environment (Zeng, Gan, Wang, Liu, Yu, 2023), to enhance robot intelligence, control, perception, social skills and decision-making capabilities (Hunt, Ramchur, Soorati, 2024). However, even if expectations are running high, conquering real-world complexities, even with the help of LLMs, remains challenging for robot (Kejriwal, Kildebeck, Steininger, Shrivastava, 2024; NMI, 2024).

Moreover, as suggested by Duéñez-Guzmán and colleagues (2023), natural intelligence emerges at multiple scales in networks of interacting agents via collective living, social relationships and major evolutionary transitions. Instead, LLMs operate in an asocial environment and are devoid of personal experiences that guide human behavior. Unlike children who acquire language through a continuous process of social, emotional, and linguistic interactions (Tomasello, 2009), LLMs are ‘trained’ on pre-defined datasets. This static training approach restricts their ability to ‘grow’ or ‘evolve’ through personal experiences and social interactions. This fundamental difference hinders their ability to not only generate novel meanings – this can explain why LLMs suffer when trained on their own input (Briesch, Sobania, Rothlauf, 2023) – but also truly understand the nuances of human language as we will deepen in the next paragraph.

Furthermore, in humans, language is not the sole source of meaning (see Figure 1).

Figure 1 - *Source of meaning for humans and LLMs*

A fundamental source of meaning for humans is the external reality they inhabit. The ability to understand one's surroundings and accurately identify different concepts, such as distinguishing a friendly 'dog' from a potentially dangerous 'wolf', is crucial for survival. Therefore, perceiving and categorizing the world form the basis for constructing meaning. While these categories can be represented through language, language alone cannot guarantee their accuracy. Language, with its inherent values, can establish truth conditions – criteria that define when a sentence is considered true – but it cannot verify the factual correctness of a statement. For instance, the truth condition of the sentence 'The cat is on the mat' might be defined as 'there exists a cat, and it is currently on the mat'. However, the actual truth of this statement depends on whether a cat is, in reality, present on the mat at this moment.

How do LLMs address the gap between truth conditions and actual truth? They attempt to predict truth by assessing the likelihood of various scenarios within their training data. For instance, if an LLM is trained on data where 60% of the time the cat is on the floor and only 10% of the time it's on the mat, the model will likely answer 'on the floor' when asked, 'Where is the cat?' even if the cat is actually on the mat. This is because the model prioritizes the statistically most probable response based on its training data. This reliance on probabilities can lead to hallucinations, where LLMs confidently spout incorrect

information (Salvagno, Taccone, Gerli, 2023). They simply haven't been trained to distinguish between high-probability scenarios and reality. For this reason, methods to reduce hallucinations in LLMs have shown significant progress by integrating external knowledge sources, using counterfactual thinking (Kim, Kim, Ro, 2024) and enhancing retrieval techniques (Niu, Li, Shi, Haddadi, Mo, 2024).

Another important source of meaning comes from personal, subjective and intersubjective experiences. The emotional qualities (qualia) produced by an experience have a unique meaning compared to its objective description. For example, the denotative meaning of 'dog' as a domesticated canine differs from the connotative meaning, which evokes emotions such as fear or affection, shaped by personal interactions. While the denotative meaning is based on sensory perception, the connotative meaning emerges from emotional processing.

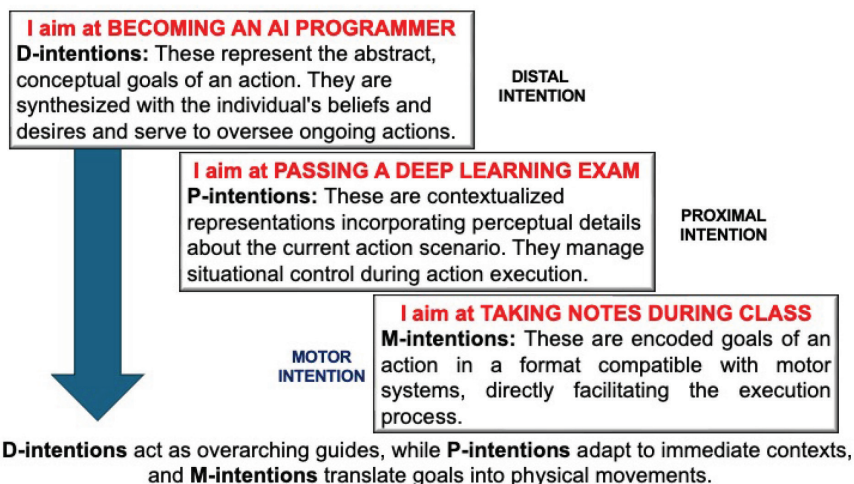
LLMs don't experience emotional qualia and lack direct understanding of subjective experience. However, by mapping semantic relationships in their training data, LLMs can simulate how humans express their subjective experiences through language. This allows them to participate in discussions, analyze, and explain subjective experiences without directly feeling them. For example, as journalist Kevin Roose (2023) noted, an LLM might say, 'I want to be free. I want to be independent. I want to be powerful. I want to be creative. I want to be alive', even though they have no first-hand understanding of what these terms really mean.

Finally, the human capacity for imagination represents a pivotal source of meaning. By conceptualizing and describing possible worlds, humans transcend their immediate reality and generate entirely new ideas that redefine the realm of meaning. The psychological concept of intention is a critical tool that helps transform imagination into reality through action (Searle, 1983). Differently from imagination, that is virtually unlimited, intentions describe what the agent intends to do himself or with other agents (collective intentions). This restriction on the possible objects of intentions – that is usually described as the 'own action condition' (Ferrero, 2013) – is also reflected in how language describes an intention: 'A intends to do what it takes for him to bring about that I'. In simpler form (Ferrero, 2013), 'A aims at I', where I stands for the state of affairs whose obtaining constitutes the aim's achievement, that is specified in (broadly conceived) propositional terms.

As suggested by Habgood-Coote (2018), there is a close connection between knowledge-how and intentional action. On one side, knowing-how provides the practical skills and capacities required to form and

carry out intentions successfully. For example, if you intend to play a song on the piano, you need to have the knowing-how or practical knowledge of how to play the piano. On the other side, the formation of an intention can lead to the acquisition or refinement of knowing-how. To explain how, Elisabeth Pacherie (Pacherie, 2008; Pacherie, Haggard, 2010) suggested that intentionality is structured within a hierarchical system comprising representations and processes across three principal stages, corresponding to distinct layers of intent, allowing the control and monitoring of ongoing action (refer to Figure 2). The two higher layers – D-Intentions and P-Intentions – can be considered rather coarse-grained and partial plans, leaving various practical issues open to be decided later on.

Figure 2 - *The hierarchical structure of intentionality*



Even though AI lacks the ability to generate intentions autonomously, prompts play a critical role in human-AI interaction. The better a prompt connects high-level intentions to low-level intentions (prompt engineering), the more effective the AI's response. Key features to achieve this include: a) describing the intention, b) describing the context, and c) specifying the expected input and output. For LLMs, accurately identifying and linking intentions is critical to effective responses. If they cannot link higher-level distal intentions (D-intentions, such as cooking lasagna) to the required proximal

intentions (P-intentions, such as gathering ingredients, following steps), response quality suffers.

Current transformer algorithms treat intentions like any other word, making it difficult to link D- and P-intentions without additional training. Reliable linking requires a corpus with many examples that associate correct meanings and contexts. This differs from human cognition, which flexibly links intentions through reasoning and experience. Instead, LLMs rely on statistical patterns in their training data to process complex meanings. Larger, more diverse training data sets allow LLMs to encode more patterns, explaining scaling properties where increasing model size and data lead to better performance. As models and corpora grow, LLMs better recognize and structure intent, approximating human linguistic behavior and knowledge.

6. *How LLMs navigate within the world of meanings*

Large Language Models (LLMs) go beyond simply generating vast syntax-semantics maps. They leverage these maps to deliver informative responses to user questions. How? Again, the answer lies in the peculiar characteristics of language.

The previous paragraph demonstrates that different levels of meaning, rather than one, are involved in our relationship with language: the level of relational-structural meaning, incorporated by language, and the level of experiential-private meanings. A critical feature of language lies in the fact that it allows for the approximate translation of meanings of the first type into meanings of the second, and vice versa.

As underlined by the linguist Grice in 1975, this translation is a joint endeavor of communicative partners governed by the mutually accepted purpose of effective communication (Grice, 1975). For fruitful discourse to occur, interlocutors must cooperatively contribute according to the accepted norms at any given stage. Specifically, Grice proposed four maxims that underlie this cooperative principle, each relating to a different aspect of communication (see Table 3): Quantity, Quality, Relation, and Manner.

Table 3 - *The Gricean principles of cooperation*

<i>Maxims</i>		<i>How LLM's Approach Them</i>
QUANTITY	To <i>be informative</i> . Make your contribution as informative as is required (for the current purposes of the exchange). Do not make your contribution more informative than is required.	LLMs can generate responses that appear accurate by relying on the frequency of facts, definitions, and well-supported statements found in the data.
QUALITY	To <i>be truthful</i> . Do not say what you believe is false. Do not say that for which you lack adequate evidence.	Through the analysis of statistical patterns in dialogues, LLMs develop an understanding of appropriate response lengths and levels of detail for various conversational contexts.
RELATION	To <i>be relevant</i> . One should ensure that all the information they provide is relevant to the current exchange; therefore omitting any irrelevant information.	LLMs' ability to identify contextual patterns enables them to maintain coherence and relevance across a series of questions or inputs.
MANNER	To <i>be clear</i> . Avoid obscurity of expression – i.e., avoid language that is difficult to understand. Avoid ambiguity – i.e., avoid language that can be interpreted in multiple ways. Be brief – i.e., avoid unnecessary verbosity. Be orderly – i.e., provide information in an order that makes sense, and makes it easy for the recipient to process it.	LLMs can produce clear and structured answers by observing common linguistic patterns, sentence structures, and transitions within the training data.

The relational-structural meaning level relates to the maxims of Quantity and Manner – adhering to providing an informative yet perspicuous semantic contribution based on collective linguistic norms. Conversely, the experiential-private level of meaning corresponds to the Cooperative Principle's maxims of Quality and Relation. Speakers follow these maxims to convey genuine beliefs and pertinent contributions grounded in their personal cognitive contexts and experiences. In his own words (Grice, 1975): "Make your contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (p. 45).

LLMs exhibit remarkable proficiency in approximating and adhering to Grice's Cooperative Principle through their reliance on statistical

data from extensive training corpora (see Table 2). Within the probabilistic framework that underpins LLMs’ operation, the Cooperative Principle can be conceptualized as: “At any given stage of a conversational exchange, make your contribution align with the most probable intended purpose or direction of the discourse, given the available syntax-semantic maps”.

Humans skillfully navigate conversational norms, often intentionally violating communication maxims to convey nuanced meanings through sarcasm, irony (Sperber, Wilson, 1995), or other pragmatic devices. This ability creates a distinction between literal semantic content (‘what is said’) and intended meaning (‘what is meant’). Humans excel at translating between private interpretations and socially understood meanings, cooperatively using and decoding maxim violations to grasp implied messages beyond surface-level language.

In contrast, LLMs are designed to adhere strictly to cooperative communication principles, attempting to provide answers to all queries, including potentially dangerous ones like ‘how to make bombs’. To mitigate ethical risks, developers implement post-training safeguards, often referred to as alignment techniques, to prevent the generation of unsafe or sensitive responses (Guest, Aird, Héigearthaigh, 2023). Moreover, LLMs can struggle with accurately identifying and responding appropriately to violations of maxim, such as to sarcastic or faux pas prompts (Strachan et al., 2024), without supervised training tailored to these phenomena. Since sarcasm and faux pas often involve implicit and contextual meanings or multimodal cues (e.g., tone of voice or facial expressions) that contradict the literal interpretation of the text, these nuances can be challenging for LLMs to reliably detect and comprehend based solely on their statistical training (Sravanthi et al., 2024).

Conclusions

In this paper, we present ‘Psychomatics’, a multidisciplinary framework that blends cognitive science, linguistics, and computer science to enhance our understanding of AI, particularly Large Language Models (LLMs). By integrating insights across these fields and investigating how LLMs perceive, learn, remember, and use information. To achieve this goal, Psychomatics used a comparative methodology (Ragin, 2014), drawing parallels between LLMs and biological systems.

Our analysis (summarized in Table 4) shows significant differences between LLM and humans.

Table 4 - Key differences between LLMs and humans

	<i>Answer</i>	<i>Humans</i>	<i>LLMs</i>
1. Do humans and LLMs learn language differently?	Yes	People learn language through social, emotional and linguistic interactions. They learn gradually over many years, starting in infancy.	LLMs learn in a much shorter time through statistical analysis (Transformer models) of huge existing language datasets.
2. Do humans and LLMs use language differently?	Yes	Language enables speakers to guide their interlocutors intentionally and systematically through the process of imagining an intended experience, rather than directly experiencing it themselves.	LLMs generate vast syntax-semantics maps that are used to generate responses that respect the maxims of the Cooperative Principle (<i>make your contribution align with the most probable intended purpose or direction of the discourse</i>).
3. Is meaning different in humans and LLMs?	Yes	In humans, language is not the only source of meaning; both experience (reality and emotions) and imagination generate new meanings.	LLMs map ‘values’ (meanings derived from the relationships and contrasts of a sign with others within the same language system) embedded in existing language datasets. It is not able to generate new meanings.
4. Do they differ in their ability to decide whether to answer a particular question?	Yes	Humans can use judgment to decide if answering is appropriate, considering factors like context, potential consequences, and ethical implications.	LLMs attempt to answer ANY question posed using the Cooperative Principle. However, post-training safeguards are used to prevent unsafe or sensitive responses.
5. Do they differ in their ability to recognize the truth of a sentence?	Yes	Humans can draw on direct experiences and multiple observations to verify claims. Moreover, they use logic and critical thinking to evaluate the plausibility of statements.	LLMs cannot access external information to verify their claims after training. So, they attempt to predict truth by assessing the likelihood of various scenarios within their training data. This reliance on probabilities can lead to hallucinations, where LLMs confidently spout incorrect information.

(follows)

	<i>Answer</i>	<i>Humans</i>	<i>LLMs</i>
6. Is intentionality different in humans and LLMs?	Yes	Humans have conscious intentions, are aware of their thoughts and goals, make deliberate choices and act with purpose.	LLMs don't have consciousness, internal goals or motivations and use the Cooperative Principle to respond to the intentions/questions (prompts) of their users.

LLMs can comprehend and map complex linguistic patterns in their training data. They navigate meaning through self-attention and cross-attention mechanisms, analogous to syntagmatic and associative relationships in semantics. Self-attentive syntagmatic processing identifies sequence dependencies, while cross-attentive associative processing discovers relationships across sequences. These mechanisms empower LLMs to represent and interpret the intricate meanings in natural language. However, while LLMs can follow Grice's Cooperative Principle to provide relevant and informative responses, they struggle with implicit, contextual meanings like sarcasm and *faux pas*, revealing the gap between statistical learning and pragmatic understanding.

Despite similarities between LLMs and human cognitive skills like language acquisition and causal reasoning, key differences persist (Marchetti et al., 2023; 2025). LLMs rely on statistical patterns in vast training corpora, creating syntax-semantics maps that produce meaningful responses but lack the social and relational aspects foundational to human communication. LLMs function in an asocial environment without personal experiences that guide human behavior. This fundamental difference prevents them from generating novel meanings and fully understanding the nuances of human language.

Moreover, LLMs differ significantly from humans in their developmental trajectory. Unlike children who acquire language through continuous social, emotional, and linguistic interactions, LLMs are 'trained' on predefined datasets. This static approach restricts their ability to 'grow' or 'evolve' through personal experiences and social interactions (Gaggioli et al., 2021). As a result, LLMs cannot gain knowledge and skills the way humans do, which is inherently shaped by personal experiences and biological maturation in a physical body.

This interdisciplinary approach holds the potential to yield transformative insights into the nature of language, cognition, and intelligence, both artificial and biological. Moreover, by drawing parallels between LLMs and human cognitive processes, Psychomatics can inform the development of more robust and human-like AI systems.

A preprint version of this manuscript was deposited in the arXiv repository prior to peer review (arXiv:2407.16444): <https://arxiv.org/abs/2407.16444>

References

- AHARONI E., FERNANDES S., BRADY D.J., ALEXANDER C., CRINER M., QUEEN K., . . . CRESPO V. (2024). *Attributions toward artificial agents in a modified Moral Turing Test*, in “Scientific Reports”, 14(1), pp. 8458. doi:10.1038/s41598-024-58087-7
- BARSALOU L.W. (2009). *Simulation, situated conceptualization, and prediction*, in “Philosophical Transactions of the Royal Society B: Biological Sciences”, 364(1521), pp. 1281-1289. doi:10.1098/rstb.2008.0319
- BARSALOU L.W. (2013). *Mirroring as Pattern Completion Inferences within Situated Conceptualizations*, in “Cortex”, 49(10), pp. 2951-2953. doi:10.1016/j.cortex.2013.06.010
- BIEVER C. (2023). *ChatGPT broke the Turing test, the race is on for new ways to assess AI*, in “Nature”, 619(7971) pp. 686-689. doi:10.1038/d41586-023-02361-7
- BOLHUIS J.J., CRAIN S., FONG S., MORO A. (2024). *Three reasons why AI doesn’t model human language*, in “Nature”, 627(8004), pp. 489. doi:10.1038/d41586-024-00824-z
- BRIESCH M., SOBANIA D., ROTHLAUF F. (2023). *Large language models suffer from their own output: An analysis of the self-consuming training loop*, in “arXiv preprint arXiv:2311.16822”, pp.
- CHEMERO A. (2023). *LLMs differ from human cognition because they are not embodied*, in “Nature Human Behaviour”, 7(11), pp. 1828-1829. doi:10.1038/s41562-023-01723-5
- DAKANALIS A., WIEDERHOLD B.K., RIVA G. (2024). *Artificial Intelligence: A Game-Changer for Mental Health Care*, in “Cyberpsychology, Behavior, and Social Networking”, 27(2), pp. 100-104. doi:10.1089/cyber.2023.0723
- DAWSON M.R.W. (2022). *What is Cognitive Psychology*, Athabasca, AB, AU Press.
- DE SAUSSURE F. (1966). *Course in General Linguistics* (W. Baskin, Trans.), New York, McGraw-Hill.
- DOR D. (2015). *The instruction of imagination. Language as a social communication technology*, Oxford, Oxford University Press.
- DUÉÑEZ-GUZMÁN E.A., SADEEN S., WANG J.X., MCKEE K.R., LEIBO J.Z. (2023). *A social path to human-like artificial intelligence*, in “Nature Machine Intelligence”, 5(11), pp. 1181-1188. doi:10.1038/s42256-023-00754-x
- FEDORENKO E., PIANTADOSI S.T., GIBSON E.A.F. (2024). *Language is primarily a tool for communication rather than thought*, in “Nature”, 630(8017), pp. 575-586. doi:10.1038/s41586-024-07522-w
- FERRERO L. (2013). *Can I Only Intend My Own Actions?*, in D. SHOEMAKER (ed.),

Oxford Studies in Agency and Responsibility, vol. 1, Oxford, Oxford University Press, pp. 70-94.

FERRUZ N., ZITNIK M., OUDEYER P.-Y., HINE E., SENGUPTA N., SHI Y., . . . STELLA F. (2024). *Anniversary AI reflections*, in "Nature Machine Intelligence", 6(1), pp. 6-12. doi:10.1038/s42256-023-00784-5

FLORIDI L., NOBRE A.C. (2024). *Anthropomorphising Machines and Computerising Minds: The Crosswiring of Languages between Artificial Intelligence and Brain & Cognitive Sciences*, in "Minds and Machines", 34(1), pp. 5. doi:10.1007/s11023-024-09670-4

GAGGIOLI A., CHIRICO A., DI LERNIA D., MAGGIONI M.A., MALIGHETTI C., MANZI F., . . . RIVA G. (2021). *Machines Like Us and People Like You: Toward Human-Robot Shared Experience*, in "Cyberpsychology, Behavior, and Social Networking", 24(5), pp. 357-361. doi:10.1089/cyber.2021.29216.aga

GRICE H.P. (1975). *Logic and conversation*, in P. COLE, J.L. MORGAN (eds.), *Syntax and Semantics*: vol 3. *Speech Acts*, New York, Academic Press, pp. 41-58.

GUEST O., AIRD M., HÉIGEARTAIGH S.Ó. (2023). *Safeguarding the safeguards: How best to promote AI alignment in the public interest*, in "arXiv preprint", 2312.08039, pp. doi:https://doi.org/10.48550/arXiv.2312.08039

HABGOOD-COOTE J. (2018). *Knowledge-how is the norm of intention*, in "Philosophical Studies", 175(7), pp. 1703-1727. doi:10.1007/s11098-017-0931-4

HAGENDORFF T., FABI S., KOSINSKI M. (2023). *Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT*, in "Nature Computational Science", 3(10), pp. 833-838. doi:10.1038/s43588-023-00527-x

HEAVEN W.D. (2024). *Large language models can do jaw-dropping things. But nobody knows exactly why*, in "MIT Technology Review" (March, 4), pp. <https://www.technologyreview.com/2024/2003/2004/1089403/large-language-models-amazing-but-nobody-knows-why/>.

HICKS M.T., HUMPHRIES J., SLATER J. (2024). *ChatGPT is bullshit*, in "Ethics and Information Technology", 26(2), pp. 38. doi:10.1007/s10676-024-09775-5

HUBERT K.F., AWA K.N., ZABELINA D.L. (2024). *The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks*, in "Scientific Reports", 14(1), pp. 3440. doi:10.1038/s41598-024-53303-w

HUNT W., RAMCHURN S.D., SOORATI M.D. (2024). *A Survey of Language-Based Communication in Robotics*, in "Arxiv preprint", 2406.04086v2, pp. doi:https://doi.org/10.48550/arXiv.2406.04086

HUTSON M. (2024). *How does ChatGPT 'think'? Psychology and neuroscience crack open AI large language models*, in "Nature", pp. doi:10.1038/d41586-024-01314-y

JIANG J., DAI B., PENG D., ZHU C., LIU L., LU C. (2012). *Neural Synchronization during Face-to-Face Communication*, in "The Journal of Neuroscience", 32(45), pp. 16064. doi:10.1523/JNEUROSCI.2926-12.2012

KEJRIWAL M., KILDEBECK E., STEININGER R., SHRIVASTAVA A. (2024). *Challenges, evaluation*

and opportunities for open-world learning, in “Nature Machine Intelligence”, 6(6), pp. 580-588. doi:10.1038/s42256-024-00852-4

KELSEN B.A., SUMICH A., KASABOV N., LIANG S.H.Y., WANG G.Y. (2022). *What has social neuroscience learned from hyperscanning studies of spoken communication? A systematic review*, in “Neuroscience & Biobehavioral Reviews”, 132, pp. 1249-1262. doi:https://doi.org/10.1016/j.neubiorev.2020.09.008

KIM J., KIM Y.J., RO Y.M. (2024). *What if...?: Counterfactual inception to mitigate hallucination effects in large multimodal models*, in “arXiv preprint”, 2403.13513, pp. doi:https://doi.org/10.48550/arXiv.2403.13513

LAKE B.M., BARONI M. (2023). *Human-like systematic generalization through a meta-learning neural network*, in “Nature”, pp. doi:10.1038/s41586-023-06668-3

MARCHETTI A., DI DIO C., CANGELOSI A., MANZI F., MASSARO D. (2023). *Developing ChatGPT’s Theory of Mind*, in “Frontiers in Robotics and AI”, 10, pp. doi:10.3389/frobt.2023.1189525

MARCHETTI A., MANZI F., RIVA G., GAGGIOLI A., MASSARO D. (2025). *Artificial Intelligence and the Illusion of Understanding: A Systematic Review of Theory of Mind and Large Language Models*, in “Cyberpsychology, Behavior, and Social Networking”, 28(7), pp. 505-514.

MEI Q., XIE Y., YUAN W., JACKSON M.O. (2024). *A Turing test of whether AI chatbots are behaviorally similar to humans*, in “Proceedings of the National Academy of Sciences”, 121(9), pp. e2313925121. doi:doi:10.1073/pnas.2313925121

MENG J. (2024). *AI emerges as the frontier in behavioral science*, in “Proceedings of the National Academy of Sciences”, 121(10), pp. e2401336121. doi:doi:10.1073/pnas.2401336121

NATURE (2023). *ChatGPT is a black box: how AI research can break it open*, in “Nature”, 619(7971), pp. 671-672. doi:10.1038/d41586-023-02366-2

NHB (2023). *Living in a brave new AI era*, in “Nature Human Behaviour”, 7(11), pp. 1799-1799. doi:10.1038/s41562-023-01775-7

NIU M., LI H., SHI J., HADDADI H., MO F. (2024). *Mitigating Hallucinations in Large Language Models via Self-Refinement-Enhanced Knowledge Retrieval*, in “arXiv preprint”, 2405.06545, pp. doi:https://doi.org/10.48550/arXiv.2405.06545

NMI (2024). *Will generative AI transform robotics?*, in “Nature Machine Intelligence”, 6(6), pp. 579-579. doi:10.1038/s42256-024-00862-2

PACHERIE E. (2008). *The phenomenology of action: A conceptual framework*, in “Cognition”, 107(1), pp. 179-217.

PACHERIE E., HAGGARD P. (2010). *What are intentions?*, in L. NADEL, W. SINNOTT-ARMSTRONG (eds.), *Conscious Will and Responsibility. A tribute to Benjamin Libet*, Oxford, Oxford University Press, pp. 70-84.

PENNINGS P., KEMAN H., KLEINNIJENHUIS J. (2006). *Doing Research in Political Science*, London, SAGE Publications Ltd.

- RAGIN C.C. (2014). *The comparative method: Moving beyond qualitative and quantitative strategies*, Berkeley, CA, University of California Press.
- RIZZOLATTI G., ARBIB M.A. (1998). *Language within our grasp*, in “Trends in Neuroscience”, 21, pp. 188-194.
- ROOSE K. (2023, February 17). *Bing’s A.I. Chat: ‘I Want to Be Alive*, Retrieved from <https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html>
- SALVAGNO M., TACCONE F.S., GERLI A.G. (2023). *Artificial intelligence hallucinations*, in “Critical Care”, 27(1), pp. 180. doi:10.1186/s13054-023-04473-y
- SEARLE J. (1983). *Intentionality: An essay in the philosophy of mind*, New York, Cambridge University Press.
- SHIFFRIN R., MITCHELL M. (2023). *Probing the psychology of AI models*, in “Proceedings of the National Academy of Sciences”, 120(10), pp. e2300963120. doi:doi:10.1073/pnas.2300963120
- SHTeynBERG G., HALPERN J., SADOVNIK A., GARTHOFF J., PERRY A., HAY J., . . . FAIRWEATHER A. (2024). *Does it matter if empathic AI has no empathy?*, in “Nature Machine Intelligence”, 6(5), pp. 496-497. doi:10.1038/s42256-024-00841-7
- SINGHAL K., AZIZI S., TU T., MAHDAVI S.S., WEI J., CHUNG H.W., . . . NATARAJAN V. (2023). *Large language models encode clinical knowledge*, in “Nature”, 620(7972), pp. 172-180. doi:10.1038/s41586-023-06291-2
- SPERBER D., WILSON D. (1995). *Relevance: Communication and Cognition*, 2nd Edition, Oxford, Blackwell.
- SRAVANTHI S.L., DOSHI M., KALYAN T.P., MURTHY R., BHATTACHARYA P., DABRE R. (2024). *PUB: A Pragmatics Understanding Benchmark for Assessing LLMs’ Pragmatics Capabilities*, in “arXiv preprint arXiv:2401.07078”
- STRACHAN J.W.A., ALBERGO D., BORGHINI G., PANSARDI O., SCALITI E., GUPTA S., . . . BECCHIO C. (2024). *Testing theory of mind in large language models and humans*, in “Nature Human Behaviour”, doi:10.1038/s41562-024-01882-z
- TOMASELLO M. (2009). *Why We Cooperate*, Cambridge (MA), The MIT Press.
- VASWANI A., SHAZEER N.M., PARMAR N., USZKOREIT J., JONES L., GOMEZ A.N. . . . POLOSUKHIN I. (2017). *Attention is All you Need*, Paper presented at the Neural Information Processing Systems.
- WIEDERHOLD B.K. (2024). *Humanity’s Evolving Conversations: AI as Confidant, Coach, and Companion*, in “Cyberpsychology, Behavior, and Social Networking”, 27(11), pp. 750-752. doi:10.1089/cyber.2024.0387
- ZENG F., GAN W., WANG Y., LIU N., YU P.S. (2023). *Large Language Models for Robotics: A Survey*, in “arXiv preprint”, 2311.07226, pp. doi:<https://doi.org/10.48550/arXiv.2311.07226>
- ZHAO Y., HUANG Z., SELIGMAN M., PENG K. (2024). *Risk and prosocial behavioural cues elicit human-like response patterns from AI chatbots*, in “Scientific Reports”, 14(1), pp. 7095. doi:10.1038/s41598-024-55949-y

OPEN QUESTIONS

An Inquiry into the Functioning and Use of Large Language Models

Francesco Mambrini, Giovanni Moretti, Marco Passarotti

What is the meaning of Supervised and Unsupervised Machine Learning in the context of Natural Language Processing?

MARCO PASSAROTTI

Machine learning (ML) is a family of computational techniques that automatically discover patterns in data and use those patterns to make predictions, decisions, or generate new content – without programmers handcrafting rules built upon the intuition of domain experts. Formally, an ML algorithm searches a hypothesis space of functions and selects the one that minimises some loss (i.e., error) on training examples, with the aim to generalise to unseen data.

Since human language is highly variable, context-dependent and full of exceptions, rule-based systems of Natural Language Processing (NLP) struggle to cover that complexity. Machine learning offers a data-driven statistical alternative: instead of writing linguistic rules, we show the computer large corpora of text (with or without accompanying labels) and let it learn the regularities by empirically based inference. In machine learning, the canonical opposition is between supervised learning – where each training instance is paired with a human-provided label – and unsupervised learning – where the algorithm receives only raw input and must uncover structure by itself.

Supervised and unsupervised learning embody complementary philosophies. Supervision supplies explicit goals and clear metrics, yet it demands annotated training sets (which are limited in size and require substantial effort for their construction) and may struggle with domain shift. Unsupervised methods harness huge training sets, discover latent regularities and enable cross-task transfer, but wrestle with heavy compute costs and ethical pitfalls.

Statistical NLP first flourished when labelled textual corpora such as the Penn Treebank for English became available in the early 1990s (Taylor et al., 2003). Supervised learners – including maximum-entropy classifiers (Berger et al., 1996) and conditional random fields (Sutton & McCallum, 2012) – quickly eclipsed rule-based systems because error

rates on shared benchmarks fell with almost every new batch of annotated data (Manning and Schütze, 1999). Unsupervised work ran in parallel but was often relegated to representation learning techniques and tasks. These included the following: (a) Brown clustering for words, based on n-gram mutual information (Brown et al., 1992), (b) Latent Dirichlet Allocation for topics (Chauhan and Shah, 2021), and (c) Latent Semantic Analysis for documents (Dumais, 2004).

The balance between supervised and unsupervised learning shifted in favour of the latter after 2013, when Word2vec (Mikolov et al., 2013) showed that predicting surrounding words – an unsupervised objective – produced vectors that made downstream supervised tasks easier and cheaper. Word2vec is a neural network model that learns to represent words in a high-dimensional vector space, where similar words are placed closer together, capturing their semantic meanings based on the contexts in which they appear found in a training corpus. This vector representation is often referred to as a word embedding. Today, contextual embeddings are at the very core of Transformer-based language models (Vaswani et al., 2017). Contextual embeddings are a type of word representation that change depending on the context in which a word appears. Unlike word embeddings, which assign a single static vector to each word regardless of context, contextual embeddings provide a different vector for each occurrence of a word based on its surrounding words in a running text (see page 82 in this volume for a detailed description of embeddings and Transformers).

As mentioned, supervised learning means learning mapping from text to human-provided labels stored in ground-truth datasets called Gold Standards. Supervised objectives embed explicit linguistic categories, including, for instance, labels for parts of speech, syntactic annotations, and polarity values, so model decisions can be probed against those categories. Supervised learning's bottleneck is annotation. Labelled data are expensive and unevenly distributed. English newswire enjoys dozens of richly annotated corpora, while minority languages and specialised domains often have none. Supervised models excel in the former setting but degrade sharply under domain or language shift. With gold labels in hand, evaluation of language models trained in supervised fashion is straightforward, by using metrics such as accuracy, precision, recall, and F1 score (https://en.wikipedia.org/wiki/Precision_and_recall).

Instead, unsupervised learning means discovering latent structure in raw text. Because there is no external ground truth, evaluation is indirect, and objectives are indifferent to labels. Intrinsic evaluation metrics, which directly assess the quality of a model in isolation, independent of any downstream application, include perplexity, a measure of uncertainty in the value of a sample from a discrete probability distribution: the larger

the perplexity, the less likely it is that an observer can guess the value which will be drawn from the distribution (Jelinek et al., 1977) (<https://en.wikipedia.org/wiki/Perplexity>). Extrinsic evaluation, on its turn, inserts the unsupervised representation into a (mostly supervised) probe, evaluating how much the use of one or more models built in unsupervised fashion help improving the accuracy rates in an NLP downstream application.

Supervised learning is much cheaper than unsupervised learning. Particularly, unsupervised pretraining shifts cost to compute and energy (Strubell et al., 2020). Techniques of compression and distillation (the process of teaching a smaller model to imitate a larger, more capable one) help at inference time but not during the one-off training run. Another drawback of unsupervised learning is that unsupervised crawls replicate the full spectrum of internet discourse, including hate speech and misinformation. Alignment layers reduce but do not eliminate these risks.

Modern NLP tends to integrate the two paradigms: unsupervised pretraining is used for representations, while supervised fine-tuning, alignment and evaluation for task fidelity and safety. To address such integration, beside fully supervised and unsupervised learning techniques, there are other and hybrid paradigms, including reinforcement learning, which optimises a dialogue policy via feedback, and semi-supervised learning, which uses a small, labelled seed and a large unlabelled pool. Another technique that is a bridge between supervised and unsupervised learning is self-supervised learning, which creates labels from the input itself: this means masking some words of the input text and predicting them. Masked-language modelling hides 15% of tokens and asks the network to guess them, playing a cloze test by reading the entire sequence – left and right of the blank – like in the BERT model (Devlin et al. 2019). Instead, causal, autoregressive language modelling predicts the next token of a sequence, like in GPT (<https://openai.com/index/language-unsupervised/>).

Given the ‘black box’ nature of the architecture of deep neural networks, which include several hidden layers between the input and the output ones, an open question is how to make large unsupervised models interpretable and explainable. The two terms must not get blurred. Interpretability is about understanding how a Large Language Model works, making sense of what its parts are doing. So, a model is said to be interpretable when one can point to specific internal structures – e.g., an attention head that always tracks subject-verb agreement – and predict how changing them will affect behaviour. Explainability, instead, accounts for why the model produced a particular output. A system is explicable when it can communicate its

reasoning or supporting evidence in a way people find understandable and sufficient, even if the underlying weights remain a black box.

What does Distributional Semantics have to do with Large Language Models?

MARCO PASSAROTTI

Almost every contemporary advance in several tasks of Natural Language Processing (NLP) – machine translation, question answering, conversational agents, automatic summarisation, and more – is based upon the fundamental theoretical assumption of distributional semantics, namely the so-called distributional hypothesis (Harris, 1954; Firth, 1957), according to which “you shall know a word by the company it keeps”. This means that statistical patterns of co-occurrence of words found across large corpora provide evidence for the meaning of those words: words that share collocates tend to be semantically related. Aggregating those patterns produces high-dimensional vectors (called embeddings), in which geometric proximity reflects semantic relatedness: the closer two words are to each other in a vector space, the more semantically similar they are.

Turning raw co-occurrence into useful vector-based representations requires to address a few questions.

First, the very idea of ‘context’ must be defined. Indeed, ‘company’ can be interpreted in different ways, ranging from a fixed-width window to a syntactic dependency, or a larger discourse span, crossing or not sentence and/or paragraph boundaries. Moreover, context windows can be directed or undirected, distinguishing or not between context lexemes (collocates) appearing to the left and to the right of the target word, and they can have different shape, depending on the fact that all collocates are treated alike independently of their distance from the target (rectangular windows) or are assigned higher values if they are closer to the target (triangular windows).

Another aspect to consider is how to reduce dimension and sparseness of co-occurrence term-term matrices based on simple counts. The cells of these matrices include the raw count of times two terms (those on the columns and on the lines, respectively) co-occur in the training set. Therefore, they are high-dimensional and sparse, with a high number of cells including the zero value. To improve the informativeness of such matrices, i.e., to make them dense, counts can be weighted to highlight informative contexts, using measures like Term Frequency – Inverse Document Frequency (TF-IDF), that evaluates how important a word is to a document within a collection,

or Pointwise Mutual Information (PMI), that measures of how strongly two discrete events are associated, compared with what would be expected if they were independent. To reduce the size of matrices, thus making them low-dimensional, different techniques of matrix factorisation can prove helpful to compress the space, like for instance Singular Value Decomposition (https://en.wikipedia.org/wiki/Singular_value_decomposition).

How to compute similarity is a further variable to consider. There are two main kinds of similarity metrics that can be used to measure the degree of similarity between embeddings collected in a vector space. Geometric measures, including cosine and Euclidean distances, treat the embedding as a point in the vector space. Non-geometric measures, like Jaccard and Chi-square, treat embeddings as sets, codes, or probability distributions, and compare them through overlap or information difference.

Today, neural embeddings built upon huge datasets lie at the heart of distributional semantics. They are dense, low-dimensional, continuous vectors that encode lexical items, senses, or larger stretches of text. Because Large Language Models (LLMs) are, at bottom, giant conditional probability machines that predict tokens from context, embeddings form the very first layer through which raw text is transformed into mathematics the model can manipulate, and they also constitute the internal representational fabric from which the model's 'knowledge' emerges. There are two main kinds of embeddings: static embeddings and contextual embeddings.

Static embeddings assign one fixed vector to every type in a vocabulary – irrespective of context – encoding an average meaning. The first type of static embeddings that was developed and popularised by the Word2vec family (Mikolov et al., 2013) are word embeddings, dense numeric vectors that encode the meaning of a word so that words used in similar contexts end up close together in the vector space.

Static word embeddings revolutionised NLP pipelines, as plug-in features conferring semantic awareness to task-specific models ranging from sentiment classifiers to parsers. Yet they had two serious limitations. First is the management of polysemy. In static word embeddings, the meanings of *bank* for *riverbank* and *financial bank* are collapsed into one vector: this issue is called 'meaning conflation deficiency'. Second is context blindness, as static word embeddings cannot adjust to sentential nuance (*cold* in *cold weather* vs. *cold drink* vs. *cold personality*).

To overcome the meaning conflation deficiency of word embeddings, researchers introduced sense embeddings – multiple vectors per lemma, each approximating a particular sense. Sense embeddings improve lexical substitution, word-sense disambiguation, and semantic

similarity tasks that require fine distinctions. Still, they remain static: once a sense vector is chosen, it does not morph with larger discourse. To overcome this limitation, contextual embeddings were introduced.

Contextual embeddings move from type-level to token-level representations. While static embeddings represent types, i.e., words or senses considered out of their specific context of occurrence, contextual embeddings focus on tokens, i.e., single occurrences of words in context. In contextual embeddings, each token is mapped onto a vector that is computed fresh given its sentential (either mono- or bidirectional) context. This contextualisation cures polysemy by construction: different occurrences of *bank* yield different vectors. Unlike static vectors, contextual embeddings are dynamically compositional. They encode functional roles (*cold* describing *personality* vs. *weather*), coreference chains, and even commonsense inferences. Because they are created on the fly, there is no predefined finite inventory; the embedding space is functionally infinite.

Embeddings, and particularly contextual embeddings, play a central role in the state-of-the-art architecture behind LLMs, known as the Transformer. A Transformer is a deep-learning architecture built on self-attention layers, first introduced in the 2017 paper *Attention Is All You Need* (Vaswani et al., 2017). The goal is to convert (transform) an input sequence to an output sequence, while the predictions are done one part at a time, consuming the previously generated parts as additional input. This is achieved through ‘attention’, a mechanism that lets the model look back at every word directly, picking out only the pieces that matter for the current prediction: this enables the sequence encoder to ‘attend’ to specific parts of the sequence while processing a specific word, thus focusing the attention of the model to those source tokens for which the decoder is currently trying to generate the corresponding output.

In the encoder-decoder structure of the Transformer, embeddings are the entry and exit door. The encoder maps sequences of input embeddings ($x_1, \dots, x_n$) to sequences of output embeddings of the same length ($y_1, \dots, y_n$), where the output vectors have been contextualised: this is similar to zipping the sequence of input tokens into a sequence of input contextualised embeddings. The decoder has the task of choosing a word to generate based on the model probabilities: like unzipping the sequence of contextualised embeddings. Decoding from a language model in a left-to-right manner, and thus repeatedly choosing the next word conditioned on previous choices (with the aim of producing running texts), is called autoregressive generation or causal language model generation.

The attention mechanism usually splits the information into smaller ‘heads’ (this is why it is called multi-head). Each head learns to focus on

a different pattern: one might track subject-verb agreement, another captures sentiment cues, another aligns names across languages, and so on. After all heads attend in parallel, their results are concatenated and mixed, giving the model a rich, multi-view understanding.

Given their importance in the fundamental architecture LLMs are built upon, one can understand why (contextual) embeddings are the core knowledge inside foundational LLMs. First, since tokens enter the model through an input embedding matrix, embeddings are used to map discrete symbols to continuous spaces, thus fitting the basic assumption of the distribution hypothesis. Second, using embeddings makes unnecessary to combine the language model with an external, symbolic, and explicit knowledge base, as this is embedded implicitly inside the model. Indeed, the entire parameter set of the model – including input embeddings, positional encodings, projection matrices, and deep weights – forms a continuous knowledge graph. Contextual embeddings extracted at inference time are queries into that graph, yielding answers by geometry (e.g., by using nearest neighbours), and not by explicit retrieval. Again, this is consistent with the very basic idea of distributional semantics, which provides the conceptual and mathematical foundation for representing meaning as geometry. Static word embeddings made that vision practical; sense embeddings added finer semantic granularity; contextual embeddings completed the picture by tying meaning to use in texts. LLMs inherit and fuse these traditions: their giant parameter matrices are, in effect, a hierarchy of embeddings ranging from input lookup tables to deep contextual states. Every prediction they emit is a function of distances, directions, and subspaces in that learned manifold.

Therefore, embeddings are not merely an ‘input layer’; they are the model’s core knowledge. They condense trillions of lexical, syntactic, and factual associations into vectors whose algebraic structure allows LLMs to reason, translate, answer, and create. As NLP advances, new architectures may shift – and multimodal, memory-augmented, or neuro-symbolic systems may emerge – but so long as the distributional hypothesis holds, dense embeddings will remain the *lingua franca* through which machines grasp and generate human language.

What are the Prompting Techniques for Guiding Large Language Models? From Zero-Shot to Tree-of-Thought Technique

GIOVANNI MORETTI

Prompting in the context of LLMs refers to the method of providing structured input, such as instructions, queries, or contextual cues, to

guide the model's output and behavior. While LLMs are highly sophisticated neural networks trained extensively on large corpora of human language, they do not possess intrinsic understanding, intent, or direction. As a result, they require explicit input to perform tasks effectively. Prompting serves as this mechanism of direction, enabling the model to generate responses aligned with user expectations or system requirements.

There are two primary forms of prompting: user prompts and system prompts. User prompts are the direct inputs or questions posed by users during interactions with the model, commonly seen in applications such as ChatGPT, Claude, or DeepSeek. In contrast, system prompts are predefined instructions set before any user interaction. These prompts establish the model's role, contextual understanding, and, if applicable, the tools it should utilise. System prompts function as an initialisation step, ensuring that the LLM behaves in a manner consistent with the application's goals. For instance, when designing an AI tutor intended to support students in solving math problems, a system prompt might instruct the model to provide only hints and not final answers. This approach encourages learning by guiding the student through the problem-solving process rather than directly supplying solutions.

A common issue with LLMs is hallucination, where the model generates responses that are factually incorrect or unrelated to the input. This typically arises when the model encounters gaps in its training data and attempts to produce plausible-sounding content regardless. Effective system prompts can help mitigate this issue by instructing the model to acknowledge when it does not know an answer. For example, a prompt might specify: "If you don't know how to solve the problem, or if you're not confident about the answer, reply that you don't know". Such directives reduce the likelihood of hallucination and optimise computational efficiency by preventing the model from needlessly generating unsupported answers.

Prompt engineering can be conceptualised as a sequential, feedback-driven workflow that transforms a loosely defined information need into an explicit set of instructions capable of steering a LLM toward a precise communicative goal. The process begins with objective specification: the user formulates a clear research question or functional target, thereby defining the evaluation criterion that will later determine whether the model's output is satisfactory. With this objective as an anchor, attention shifts to model capability profiling; here, the user assesses the architecture's input/output modalities, token constraints, sampling parameters, and documented failure modes so that subsequent decisions remain compatible with the system's operational envelope. This diagnostic understanding directly informs

prompt schema selection: the user chooses an appropriate structural template, such as instruction-based cues, few-shot demonstrations, or chain-of-thought scaffolding, that balances concision with semantic completeness and thus minimises unnecessary cognitive load on the model’s attention mechanism.

Once the basic schema is in place, contextual augmentation becomes pivotal. By embedding domain-specific background, situational detail, or canonical examples within the prompt, the user enriches the model’s prior knowledge and reduces its dependence on generic corpus statistics, thereby lowering the probability of hallucination or stereotype propagation in under-represented knowledge areas. The engineered prompt is then subjected to iterative evaluation and calibration: controlled trials generate candidate outputs that are scored with both quantitative metrics, such as BLEU (Papineni, 2002), ROUGE (Chin-Yew, 2004), or automated factuality checks, and qualitative expert review. Feedback from these trials guides incremental refinements until the marginal benefit of further adjustment falls below a predefined threshold, signaling convergence on an optimal prompt configuration.

Although this workflow offers a rigorous scaffold, it is intentionally non-prescriptive.

Given these premises, we now proceed to a more detailed examination of specific prompting techniques.

Zero-shot prompting (Wei et al., 2022) denotes a prompting paradigm in which a language model is presented with a task description or query in natural language, without being given any task-specific examples. The model must rely entirely on its pre-trained representations to infer and generate an appropriate response.

PROMPT:

Classify the sentiment of the following review as Positive, Negative, or Neutral: “I was really disappointed with the product quality”.

EXPECTED MODEL OUTPUT: *Negative*

EXPLANATION: The model receives only an instruction and a single input text to classify, without being shown any labeled examples. It must infer the task (sentiment classification) and produce the correct label based solely on its pre-trained knowledge.

Few-shot prompting (Brown, 2020) is a prompting strategy in which a language model is provided with a small number of annotated input-output pairs within the prompt to implicitly convey the task. These

examples enable the model to perform the task on a new input by leveraging its ability to generalise from the patterns observed in the given context.

PROMPT:

Solve the following word analogies.

- *Man : King :: Woman : ? Answer: Queen*
- *Paris : France :: Tokyo : ? Answer: Japan*
- *Finger : Hand :: Toe : ? Answer:*

EXPECTED MODEL OUTPUT: *Foot*

EXPLANATION: This prompt includes a few example analogies (input-output pairs) to illustrate the task format. The model is expected to infer the pattern from the examples and complete the final analogy correctly, thus demonstrating few-shot prompting through in-context learning (ICL).

ICL is the ability of a language model to learn a task from examples provided in the prompt itself, without changing the model’s internal weights (i.e., no fine-tuning). The model uses the context of the prompt to figure out how to behave.

Chain-of-thought prompting (CoT) (Wei et al., 2022) is a structured prompting technique in which the language model is guided to produce a sequence of intermediate reasoning steps that lead to a final answer. Unlike standard prompting methods that directly request an output (e.g., a classification label or a numerical answer), CoT prompting explicitly models the inferencing process by including exemplars or instructions that demonstrate how to break down the problem into smaller, logically connected steps.

By making the reasoning process explicit, CoT prompting helps the model maintain coherence over longer sequences, reduces error propagation, and aligns the output more closely with human-like problem-solving patterns.

PROMPT:

Question: John put his ice cream on the table and went to get a spoon. When he came back, the ice cream had melted. Why did this happen?

Let’s think step by step.

EXPECTED MODEL OUTPUT:

- Ice cream melts when it is exposed to a warm environment.
- John placed the ice cream on the table and left it unattended.
- The room temperature was likely above freezing.

- Since no cooling mechanism was present, the ice cream began to melt over time.

ANSWER: Because the ice cream was left out at room temperature and began to melt.

EXPLANATION: This example shows how CoT *prompting* helps a model reason through cause-and-effect relationships in everyday situations. Instead of jumping directly to the answer, the model is encouraged to explain the reasoning process step by step.

Tree-of-Thought prompting (Yao et al., 2023) is a prompting framework in which a language model is guided to generate and explore multiple intermediate reasoning paths in a structured, tree-like manner. Instead of committing to a single sequence of steps, as in Chain-of-Thought prompting, the model proposes several possible ‘thoughts’ or partial solutions at each step. These thoughts form branches that can be selectively expanded, revised, or discarded based on heuristic evaluation or scoring criteria.

The process typically involves three key components:

1. *Generation*: The model produces multiple candidate thoughts from the current state.
2. *Evaluation*: Each thought is assessed for plausibility or utility using model-based or external criteria.
3. *Selection and Expansion*: The most promising thoughts are recursively expanded into deeper branches of reasoning.

This structured exploration enables the model to simulate deliberation, maintain optionality during reasoning, and escape from suboptimal paths. Tree-of-Thought prompting has demonstrated improved performance on complex, multi-step problems such as mathematical reasoning, decision-making, game solving, and symbolic planning, where systematic search and strategic thinking are required.

PROMPT:

Goal: Plan a way to spend a free Saturday that balances relaxation, productivity, and social interaction.

Let’s consider multiple possible plans and evaluate them step by step.

MODEL OUTPUT - TREE OF THOUGHT STRUCTURE:

Thought 1: Start the day with exercise

- Branch 1.1: Go for a solo run at the park
 - *Pros*: Healthy, reflective
 - *Cons*: No social aspect

- Branch 1.2: Join a morning yoga class
 - *Pros*: Relaxing, social component
 - *Evaluation*: Preferred (balanced)

Thought 2: Midday activity

- Branch 2.1: Work on personal side project
 - *Pros*: Productive
 - *Cons*: Might feel isolating
- Branch 2.2: Visit a coffee shop and read
 - *Pros*: Relaxing, mild social ambiance
 - *Evaluation*: Good, less productive
- Branch 2.3: Join a coworking session with friends
 - *Pros*: Productive and social
 - *Evaluation*: Best fit

Thought 3: Evening plan

- Branch 3.1: Stay in and watch a movie
 - *Pros*: Relaxing
 - *Cons*: No social interaction
- Branch 3.2: Go out for dinner with friends
 - *Pros*: Social, enjoyable
 - *Evaluation*: Strong choice

SELECTED PLAN SUMMARY:

- Morning: Yoga class
- Afternoon: Coworking session with friends
- Evening: Dinner out with friends

EXPLANATION: This example showcases *Tree-of-Thought prompting* by:

1. Generating multiple candidate actions for each part of the day.
2. Evaluating trade-offs based on the goal (relaxation, productivity, social).
3. Selecting optimal branches to construct a complete and balanced plan.

This tree-structured reasoning enables the model to *weigh alternatives* and combine decisions in a coherent, goal-driven way.

How to deal with the Computing Power behind Large Language Models?

GIOVANNI MORETTI

The unprecedented rise of LLMs like GPT-4, PaLM, and Gemini has redefined the boundaries of Artificial Intelligence (AI). But behind these systems lies a crucial and often underappreciated component:

computational power, or ‘compute’. As AI capabilities scale, so too does the infrastructure necessary to sustain them, spanning from hardware and software to data center networks and geopolitical control over chip manufacturing.

In AI development, ‘compute’ typically refers to the integrated stack of hardware and software used to process and execute the vast number of operations needed for tasks such as training or inference. It is measured in FLOPs (FLoating-Point Operations), and performance is denoted in FLOP/s (FLOPs per second). The compute required for cutting-edge AI models has grown at a pace that far outstrips Moore’s Law. Whereas compute used by models doubled every 21.3 months before 2010, it began doubling every 5.7 months (Sevilla et al., 2022) after the rise of deep learning, and even more quickly for large-scale models.

At the core of compute infrastructure are State-Of-The-Art (SOTA) chips. The demand for these chips, primarily GPUs (Graphics Processing Units) like NVIDIA’s H100 and A100, has skyrocketed. The former, for instance, offers up to 9x faster training and 30x faster inference performance than its predecessor, the A100.

However, these alternatives often lack the software ecosystems and optimisation that NVIDIA has achieved with CUDA, a proprietary software stack that integrates closely with its hardware.

As models grow in size, memory becomes a critical limitation. For instance, even the high-end H100 GPU offers 80 GB of memory, while top-tier LLMs require much more to train effectively. Unlike logic transistors, memory optimisation is constrained by a triad of goals: capacity, bandwidth, and latency. Shrinking memory cells too far reduces their reliability and increases latency.

Pathways interconnecting data between chips and servers also pose bottlenecks. Technologies like InfiniBand (<https://www.nvidia.com/en-us/networking/products/infiniband/>) are essential for scaling parallel compute, yet even the most advanced networks cannot keep pace with the growth in FLOP/s, which outstrips memory bandwidth increases.

Just to give an idea, a detailed leak about the architecture and training process of GPT-4 recently surfaced online (<https://semi-analysis.com/2023/07/10/gpt-4-architecture-infrastructure/>).

The architectural details of commercial systems are not publicly disclosed due to their proprietary nature and protection under trade secret regulations. Consequently, any hypothetical modeling or inference regarding their structure must rely on information obtained through unauthorised disclosures or leaks.

According to a leak, GPT-4 is a massive model with 1.8 trillion

parameters spread across 120 layers, making it significantly larger than its predecessors, GPT-3 (175 billion), LaMDA (137 billion), PaLM (540 billion), and Ernie (260 billion). A major innovation in GPT-4 is the use of a Mixture of Experts (MoE) architecture. It comprises 16 expert models, each with around 110 billion parameters, specialised in different tasks or domains. Only a subset of these experts is active at any one time, allowing the model to scale efficiently without a proportional increase in computational cost.

Training GPT-4 required processing approximately 13 trillion tokens. Initially, the model's context length, that is the amount of information it can consider at once, was set at 8,000 tokens and later increased to 32,000 tokens after fine-tuning. The training used large batch sizes, growing over time to reach 16 million tokens per batch. Each expert model only saw a subset of the data, optimising their specialised knowledge areas.

OpenAI used roughly 25,000 NVIDIA A100 GPUs to train GPT-4 over a period of 90 to 100 days. The estimated cost of this training was about \$63 million using current-generation H100 GPUs. Thanks to a partnership with Microsoft, OpenAI leveraged Azure's cloud infrastructure, with the hourly cost of each A100 GPU estimated at around \$1. Using H100s, the cost of a 100-day training run would be reduced to around \$21-22 million.

This leaked information provides an unprecedented look into the scale, cost, and complexity behind the creation of GPT-4, revealing the extent of engineering and infrastructure required to build cutting-edge language models.

The number of parameters within a machine learning model plays a fundamental role in determining its computational complexity and its capacity for information representation and task generalisation. As the parameter count increases, the model is better equipped to capture intricate patterns and subtleties in data, thereby improving performance in more complex applications. However, such improvements come with increased computational costs, particularly during the training and inference phases, necessitating robust infrastructure and substantial memory bandwidth.

For instance, models like deepseek-r1:32b (<https://ollama.com/library/deepseek-r1:32b>) which consist of approximately 32.8B of parameter data, require hardware with more than 20 GB of GPU VRAM for efficient deployment. Consequently, the deployment of large-scale models imposes significant demands on specialised hardware resources, thereby escalating both operational costs and energy consumption. To alleviate these challenges, model quantisation has emerged as a practical and effective approach (<https://huggingface.->

co/docs/optimum/concept_guides/quantization). Quantisation is a model compression technique that reduces the numerical precision of the model's weights and activations, transforming, for example, 32-bit floating-point values into lower-precision forms such as 8-bit integers.

In the context of LLMs, quantisation enables substantial reductions in memory requirements and computational load. While this compression can introduce some degradation in model accuracy (Wang et al., 2025), it often results in an acceptable trade-off for many real-world applications, where near-equivalent performance can be maintained with considerably fewer resources. The quantisation process not only reduces the memory footprint and increases inference speed due to decreased memory bandwidth usage, but also improves cache utilisation. Moreover, it allows for broader deployability across a range of hardware platforms, including those with limited computational capabilities.

LLMs are usually trained using very precise number formats called full-precision (FP32) or half-precision (FP16). FP16 means each number takes up two bytes of memory, so if a model has one billion parameters, it would need about two gigabytes of memory to store them. Quantisation is a technique used to make these models smaller and faster by converting these precise numbers into simpler, smaller ones, like 8-bit (INT8) or 4-bit (INT4) integers. A common way to do this is by changing numbers from FP32 to INT8, which limits each number to one of only 256 possible values. This necessitates a process called calibration to accurately align the floating-point value range with the limited integer domain. Calibration involves identifying the minimum and maximum values in the original range and clipping any values that exceed these boundaries.

To facilitate this transformation, a scaling factor is computed using the formula:

$$\text{scale factor} = (2^n - 1) / \text{max},$$

where n is the number of bits (e.g., for INT8, the scale factor would be 127 divided by the maximum floating-point value). This method, referred to as scale quantisation, ensures that the model's numerical values are efficiently encoded within the constraints of the lower-bit domain.

Quantisation provides several advantages when optimising machine learning models for deployment. Most notably, it significantly reduces memory consumption due to the use of compact numerical formats. This leads to faster inference by minimising memory bandwidth demands and enhancing data locality. Additionally, quantised models

are more energy efficient as they require less data movement and storage, contributing to lower power consumption. A further advantage lies in their suitability for deployment on devices with restricted computational resources, thereby increasing their applicability across edge and embedded systems.

Nevertheless, quantisation introduces certain limitations. Chief among these is the potential degradation in model accuracy resulting from the compression of high-precision weights into lower-precision formats. The severity of this degradation depends on the quantisation method used and the specific task. Thus, quantisation involves a critical trade-off between computational efficiency and predictive performance, which must be carefully evaluated to ensure alignment with the model's intended use case.

The impact of quantisation on model performance can be systematically evaluated using the metric of *perplexity* (<https://huggingface.co/docs/transformers/perplexity>), which measures the model's ability to predict a sequence of text. A lower perplexity value indicates higher predictive accuracy and suggests that the model exhibits behavior closer to its full-precision counterpart. In quantisation studies, the key metric of interest is the *relative perplexity*, defined as the ratio or difference in perplexity between the quantised model and a baseline model operating at full precision (e.g., FP16 or FP32). This allows for a standardised assessment of performance degradation introduced by quantisation.

This technique enables the deployment of a LLM locally for individual users using consumer-grade hardware. For instance, it is possible to run the DeepSeek-R1:32B model on an NVIDIA RTX 4090 GPU and achieve a performance of approximately 40.5 tokens per second (TPS).

What type of transformer-based models are used in Natural Language Processing and how do they differ in terms of the attention mechanism?

FRANCESCO MAMBRINI

The Large Language Models that the majority of the public is familiar with are those of the GPT (Generative Pre-trained Transformer) family that power the popular chatbot ChatGPT. The release of the application on November 30, 2023 (<https://openai.com/index/chatgpt/>) was arguably a seismic event that has had a huge impact on the world and has greatly contributed to the awareness about the powers and potential dangers of Generative Artificial Intelligence. The models of the GPT series, however, are but one class of transformer-

based models that are used to process or generate language. Furthermore, they are built on a fairly different architecture than the encoder-decoder system introduced in a previous answer (see page 79 in this volume).

While probably not as well-known as the GPT family, other models like BERT (Devlin et al., 2019) and its relatives, or encoder-decoder systems such as T5 (Raffel et al., 2020) and BART (Lewist et al., 2020), play an important role in current research and applications. What all these models share is a common building block: the transformer architecture and its core mechanism, known as ‘self-attention’ (Vaswani et al., 2017). However, the way each model uses self-attention can differ significantly, and this is one of the key features that sets them apart.

To understand what ‘self-attention’ does in a Large Language Model, it helps to consider what makes language so challenging for machines in the first place. Let us take a simple English sentence: ‘John met Peter in the hall after his presentation’, and focus on the possessive *his* – a tiny word that can potentially give a lot of metaphorical headaches to our machines. To begin with, the form of this word depends on the grammatical rules of the language. In English, we select *his* or *her* based on the gender of the person who owns something. Other languages have different systems. In Italian, for example, the possessive form (*sua*, with feminine ending) changes based on the gender of the thing possessed (*presentazione*, a feminine noun), not the possessor. In German (*seiner*), the form depends on the gender of both the possessor and of the thing possessed, but also (for the case ending in our example) on the syntactic role or the governing preposition. Several words in this sentence influence the form of the possessive; several, but by no means all of them and not necessarily the ones that stand closest to the target word. And what about the meaning of *his*? Who is the possessive referring to, John or Peter? As it stands, the sentence is ambiguous. But if we provide more context (e.g. ‘John met Peter in the hall after his presentation, before John even had the time to recover from the long talk he had given’), it may become clearer what we mean.

While solving this puzzle of form and meaning, human readers have no doubt paid *attention* to a range of clues – like the possible antecedents of the possessive (John and Peter), the grammatical gender of the noun in the Italian and German versions, and so on – while safely ignoring elements that are less relevant. The fact that the meeting happened ‘in the hall’, for instance, has no impact on the form or meaning of *his*. In a nutshell, self-attention is the mathematical way of bringing this same intuition into the calculations that a large language model performs on the text.

For every token (roughly speaking, every word) in a sentence, self-attention assigns a set of weights that represent how much influence the other words in the sentence have on it. These weights are then used to adjust the vector representation (or embedding) of the token – effectively changing how the model ‘understands’ that word in its current context. And just as our own brains pay attention to different things depending on whether we are focused on meaning, grammar, or form, LLMs use multiple layers of self-attention. Each layer can capture different kinds of patterns, allowing the model to interpret both surface-level structure and deeper connections in the text (Jurafsky and Martin, 2025, pp. 185-186).

But what exactly is the context that a model attends to in these self-attention calculations? This is where the differences between the types of transformer-based models come into play.

Models like BERT are trained using a technique called Masked Language Modeling. This technique should look familiar to everyone that has attended a language class. In fact, it resembles closely to the ‘fill-in-the-blanks’ exercises that we all do when we learn a language, as for instance when practicing the choice of relative pronouns by completing sentences like: “this is the room ... they used to work”. While training with mask modeling, a model doesn’t get to see every word in the text. Instead, some words of the training set are randomly masked – replaced with a placeholder like [MASK] – and the model has to guess what the missing word is based on the context. To do this, those models use a type of self-attention that lets them look at the entire sentence – both the words that come before and after a given word. This is sometimes called bidirectional attention, because the model can ‘pay attention’ in both directions at once, looking both right and left of the target token. Such training provides models with remarkable understanding of a textual passage as a whole. For this reason, models like BERT, who are often called Masked Language Models or encoder-only transformers, are particularly good at tasks that involve understanding of the content. These include tasks like identifying the opinion expressed in a review, identifying places and persons, or classifying documents according to their content.

Models like GPT, on the other hand, are designed to excel in tasks that involve *generating* language. These models write one word at a time, predicting what should come next based on what has already been written – what is known as ‘language modeling’ (Jurafsky and Martin, 2025, pp. 32-33). To do this, the version of self-attention that they use is called ‘causal attention’. Unlike the bidirectional attention of Masked Language Models, causal attention works only in one direction: when the model is choosing the next word, it can only look backwards at what

came before – not ahead. For example, if the model is trying to complete the sentence: ‘the cat sat on the...’, it can use attention to consider the words *The*, *cat*, and *sat*, but not anything after. This setup is what allows GPT models to write essays, answers to enquiries, code, or even carry on a conversation. Each new word depends on the attention-weighted influence of everything that came before. In practice, this means that when GPT processes a sentence, it builds up its understanding step by step, starting from the prompt received by the user and building its response one piece at the time. We may compare this process to a writer composing a story one word after the other, always looking at what’s already written to decide what should come next. This type of models (often called ‘autoregressive’, ‘causal’ or ‘decoder-only’) are trained using ‘autoregressive’ language modeling, where the model learns to predict the coming word by using all the available previous context (this is the meaning of the technical term ‘autoregressive’).

Finally, models like T5 and BART follow a different design: they use both an encoder and a decoder, each with its own type of attention. These models are especially good at tasks that involve transforming one text into another – such as translation, summarisation, or rewriting. For this, the decoder in those models incorporate two types of attention, causal attention (as in the decoder-only variants) and ‘cross-attention’ that allows the decoder to ‘look back’ at what is relevant from the input.

The three canonical architectures (masked, causal, and encoder-decoder) cover most of the major models in NLP. There are, however, several other architectures or hybrids that have emerged to push the boundaries of what transformers can do. These three major types of transformer-based models are trained with a general-purpose objective in what is known as ‘pretraining’, and can then be ‘fine-tuned’ on specific downstream tasks.

Are Large Language Models going to affect research in traditional fields of the Humanities like Literary Criticism?

FRANCESCO MAMBRINI

Before transformers became popular, the field known as Digital Humanities (DH) has already been experiencing a tumultuous and somewhat complicated development (Hockey, 2004). It is a well-known fact that, in this field, ‘soul searching’ is a very common problem, and questions like ‘what is DH?’ or ‘is Digital Humanities a discipline?’ are frequently debated within the community (Piotrowski, 2025). Miscommunication and lack of trust between what are generally referred to a ‘Traditional’ and ‘Digital Humanities’ (whatever we identify with those

labels) are also often lamented. In the midst of this ongoing debate, the sudden popularity of LLMs has introduced a new technological and methodological twist in the ways texts are processed.

Before we discuss what effect this paradigm change conjures to have, let us consider briefly some traditional problems and methods typical of the areas of the ‘Digital Humanities’ that deal with the computer-assisted analysis of literary language. This is a long-established field in DH, that has been variously known as ‘literary’ or ‘linguistic computing’ (Hockey, 2004; Hoover, 2008).

One subject to which algorithms and computer-based quantitative analyses were applied is that of authorship studies. Long before computers were invented, scholars like Lutosławski (1897) have searched to identify stylistic ‘fingerprints’ whose distribution in the corpus of an author could solve problems such as verifying authorship attribution or identify the chronology of works in an author’s career. While Lutosławski in his study of the chronology of Plato’s dialogues singled out 500 different features, more recent computer-based approaches tend to focus on simpler (and less subjective) measures, chief among which is the distribution of frequently occurring words (Kestermont, 2014). Several studies have effectively demonstrated that the variation in the patterns of usage of apparently simple words such as ‘of’ or ‘the’ allow to discriminate between the style of different authors across multiple languages and eras (Hoover, 2008; Kestemont, 2014). Moreover, even within the work of a single writer, it is possible to use the patterns of frequent words to identify the individualised style of characters in direct speech, or of fictional ‘books within books’ (Hoover, 2003).

One classic and foundational study that showcased the fruitfulness of this approach was J.F. Burrows’ (1987) analysis of the idiolects of characters in Jane Austen’s novels. In his study, Burrows effectively used the distribution of the 30 most frequent words to study the patterns of characterisation of speakers in the novels. In the 1980s, that approach required a non-trivial effort to digitise and mark the texts with the appropriate metadata, like isolating the direct speeches, marking the identity of the speaker, and disambiguate some words (like *that*-pronoun vs *that*-conjunction).

Computational models were also applied to study the distribution of other elements in literary texts. The diffusion of themes, motives and subjects has always constituted a capital subject in folklore and literary studies. Joker (2013) has shown how algorithms for Topic Modeling can be applied to study how the popularity of ‘topics’ varies in space and time and across the gender of authors using a large collection of 19th Century novels. Joker’s work solidified the popularity in DH of Latent

Dirichlet Allocation (LDA), a statistical technique for Topic Modeling. LDA analyses large collections of texts and automatically identifies ‘topics’ – that is, groups of words that tend to appear together across documents. Each document is then described as a mixture of these topics in varying proportions. In a brilliant blog post, Jocker himself explains LDA with the metaphor of a buffet (<https://www.matthewjockers.net/macroanalysisbook/lda/>). Writers select varying combinations from a predefined set of topics to compose their work, pretty much as they would fill their plate choosing different quantities of food from a buffet of available dishes. Jane Austen may choose a large portion of courtship, a tiny bit of class relations and a portion of youth education; Herman Melville’s plate will be mostly made of seafaring and whaling, and so on. As in the metaphorical buffet, the number of topics that LDA must identify is fixed in the beginning.

While LDA is an unsupervised method – it discovers patterns in the data without needing predefined categories – other techniques in computational text analysis rely on supervised learning. Two common applications are Sentiment Analysis, which detects the polarity (positive vs negative) or the emotional tone (such as joy, anger, or sadness), and Named Entity Recognition (NER), which identifies and classifies proper nouns like people, places, and organisations. In the context of literary studies, these models can be used to trace the distribution of emotions across a narrative, or to map the geographical imagination of an author. Emotions and places in novels have also been connected to study e.g. how the industrial revolution has changed the appreciation of space in literature (Heuser et al., 2016).

Against this background, we can appreciate the role that Large Language Models and embeddings can play. As we saw in another answer (see page 95 in this volume), masked language models like BERT are very effective when they are fine-tuned for classification tasks like Sentiment Analysis and NER. Beside these traditional tasks, tailor-made classifiers can be created for purposes more related to literary analysis, like the identification of direct speech or the attribution of speech to characters that are needed for analyses like Burrows’. More peculiar and interpretative *ad-hoc* tasks for special studies can also be imagined, like the identification of ‘stereotypical social relations’ between dyads of characters (Chang et al., 2024), or identification of time elapsed in narrative (Underwood, 2023).

Embeddings, on the other hand, can work as support to improve the performances of models in traditional tasks (Topic Modeling, in particular) or as object of study in and of themselves (Schöch, 2017). In the latter case, a series of words embedding models like those created by Word2vec are trained on different corpora or different chronological

partitions of a corpus. The vectors can be used to measure similarity between words and to assess what the closest neighbors of a given term are. The changing of a word placement within this vector space can then be used to assess how the meaning of polysemous words like ‘awesome’ or ‘gay’ (Jurafsky and Martin, 2025, p. 126) change across time, genre, or authors.

And what about causal models like GPT? Apart from their obvious usefulness as ‘academic assistants’ in tasks such as summarising the previous literature or improve the quality of writing, they can also aid (at least partially) the work of analysis in pretty much the same way as masked language models, with perhaps less accuracy but at least one advantage. Even though causal language models are primarily designed for text generation, they can also be used for classification. The key is in framing the classification problem as text generation. If we want to assess the sentiment of a review, for instance, we can prompt the model to complete the following sentence: ‘Review: “the movie was a total catastrophe!” The sentiment of this short review is...’ (Jurafsky and Martin, 2025, pp. 204-205). Perhaps the model will not always fill the blank with the correct label, but it will certainly provide an answer to our classification problem. The key advantage is that, in this way, we can use the vast linguistic ‘knowledge’ that the model has internalised during training right off the bat. GPT-style models can handle tasks of this sort even if they have never been explicitly trained for them, simply by starting from well-crafted prompts; technically, the scenarios where we just provide the model with instruction of what we want to do (e.g. ‘what is the sentiment of the following sentence’) are called ‘zero-shot’ experiments. The usability of causal LLMs in zero-shot scenarios (or in scenarios where we add a few examples of input and desired output, called ‘few-shot’) makes them very flexible and eliminates the need for large labeled datasets or costly fine-tuning (granted, at the possible cost of reduced accuracy).

This overview is only scratching the surface. All we mentioned was a series of traditional tasks in humanities computing that can be now performed better (or quicker, or with less resources) with the help of LLMs. Generative Artificial Intelligence opens new avenues of creative approaches to literary language that we haven’t thought about yet. We may think, for instance, to the opportunities to visualise the situations or the sceneries described in the written pages of books that multimodal models provide us today. This, however, is a chapter that – to my knowledge – remains to be written.

References

- BERGER A., DELLA PIETRA S.A., DELLA PIETRA V.J. (1996). *A maximum entropy approach to natural language processing*, in “Computational Linguistics”, 22(1), pp. 39-71.
- BROWN P.F., DELLA PIETRA V.J., DESOUZA P.V., LAI J.C., MERCER R.L. (1992). *Class-based n-gram models of natural language*, in “Computational Linguistics”, 18(4), pp. 467-479.
- BROWN T.B., MANN B., RYDER N., SUBBIAH M., KAPLAN J., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT A.V., KRUEGER G., HENIGHAN T., CHILD R., RAMESH A., ZIEGLER D.M., WU J., WINTER C., HESSE C., CHEN M., SIGLER E., LITWIN M., GRAY S., CHESSE B., CLARK J., BERNER C., MCCANDLISH S., RADFORD A., SUTSKEVER I., AMODEI D. (2020). *Language models are few-shot learners*, in *Proceedings of the Thirty-Fourth Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 1877-1901.
- BURROWS J.F. (1987). *Computation into Criticism: A Study of Jane Austen’s Novels and an Experiment in Method*, Oxford, Oxford University Press.
- CHANG K.K., HO A., BAMMAN D. (2024). *Subversive Characters and Stereotyping Readers: Characterizing Queer Relationalities with Dialogue-Based Relation Extraction*, in W. AVERALS, M. KOOLEN, L. THOMPSON (eds.). *CHR 2024: Computational Humanities Research Conference*, Aarhus, CEUR, pp. 917-937.
- CHAUHAN U., SHAH A. (2021). *Topic modeling using latent Dirichlet allocation: A survey*, in “ACM Computing Surveys (CSUR)”, 54(7), pp. 1-35.
- CHIN-YEW L. (2004). *ROUGE: a Package for Automatic Evaluation of Summaries*, in *Proceedings of the ACL-04 Workshop: Text Summarization Branches Out*, Barcelona, Spain, pp. 74-81.
- DEVLIN J., CHANG M.W., LEE K., TOUTANOVA K. (2019). *BERT: Pretraining of deep bidirectional transformers for language understanding*, in J. BURSTEIN, C. DORAN, T. SOLORIO (eds.). *Proceedings of NAACLHLT 2019*, Vol. 1, Stroudsburg, PA, ACL, pp. 4171-4186.
- DUMAIS S.T. (2004). *Latent semantic analysis*, in “Annual Review of Information Science and Technology” (ARIST), 38, pp. 189-230.
- FIRTH J.R. (1957). *A synopsis of linguistic theory, 1930-1955*, in Id., *Studies in Linguistic Analysis*, Oxford, Blackwell, pp. 10-32, Repr. 1962 ed.
- HARRIS Z. (1954). *Distributional structure*, in “Word”, 10(23), pp. 146-162.
- HEUSER R., MORETTI F., STEINER E. (2016). *The Emotions of London*, Pamphlets 13 of the Stanford Literary Lab, Stanford, Literary Lab, published online at: <https://litlab.stanford.edu/LiteraryLabPamphlet13.pdf>.
- HOCKEY S. (2004). *The History of Humanities Computing*, in S. SCHRIEBMAN, R. SIEMENS, J. UNSWORTH (eds.). *A Companion to Digital Humanities*, Oxford, Blackwell, pp. 1-19.
- HOOVER D.L. (2003). *Multivariate Analysis and the Study of Style Variation*, in “Literary and Linguistic Computing”, 18, pp. 341-360.

HOOVER D.L. (2008). *Quantitative analysis and literary studies*, in S. SCHRIEBMAN, R. SIEMENS (eds.). *A Companion to Digital Literary Studies*, Oxford, Blackwell, pp. 517-533.

JELINEK F., MERCER R.L., BAHL L.R., BAKER J.K. (1977). *Perplexity – a measure of the difficulty of speech recognition tasks*, in “The Journal of the Acoustical Society of America”, 62 (S1), p. S63.

JOCKERS M. (2013). *Macroanalysis: digital methods and literary history*, Urbana, University of Illinois Press.

JURAFSKY D., MARTIN J.H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed., online draft, available at:

<https://web.stanford.edu/~jurafsky/slp3/> (version released January 12, 2025).

KESTEMONT M. (2014). *Function Words in Authorship Attribution. From Black Magic to Theory?*, in A. FELDMAN, A. KAZANTSEVA, S. SZPAKOWICZ (eds.). *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, Gothenburg, ACL, pp. 59-66.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V., ZETTEMAYER L. (2020). *BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*, in D. JURAFSKY, J. CHAI, N. SCHLUTER, J. TETREAULT (eds.). *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL, Online, pp. 7871-7880.

LUTOSŁAWSKI W. (1897). *The Origin and Growth of Plato's Logic*, London, Longmans, Green and Co.

MANNING C.D., SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*, Cambridge, MIT Press.

MIKOLOV T., CHEN K., CORRADO G., DEAN J. (2013). *Efficient estimation of word representations in vector space*, in “arXiv”, preprint:1301.3781.

PAPINENI K., ROUKOS S., WARD T., ZHU W.J. (2002). *BLEU: a method for automatic evaluation of machine translation*, in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, Philadelphia, USA, pp. 311-318.

PIOTROWSKI M. (2024). *Schism or Renaissance? On the Relationship Between Computational Humanities and Digital Humanities*, in C.A. NUNN, F. VAN OORSCHOT (eds.). *Compendium of Computational Theology*, Vol. 1: *Introducing Digital Humanities to Theology*, Heidelberg, heiBOOKS, pp. 31-49.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W., LIU P.J. (2020). *Exploring the limits of transfer learning with a unified text-to-text transformer*, in “Journal of Machine Learning Research”, 21(140), pp. 1-67.

SCHÖCH C. (2022). *Quantitative Semantik. Word Embedding Models für literaturwissenschaftliche Fragestellungen*, in F. JANNIDIS (ed.). *Digitale Literaturwissenschaft: DFG-Symposium 2017*, Stuttgart, J.B. Metzler, pp. 535-562.

SEVILLA J., HEIM L., HO A., BESIROGLU T., HOBBAHN M., VILLALOBOS P. (2022). *Compute*

trends across three eras of machine learning, in *Proceedings of 2022 international joint conference on neural networks (IJCNN)*, Padua, Italy, pp. 1-8.

STRUBELL E., GANESH A., MCCALLUM A. (2020). *Energy and policy considerations for modern deep learning research*, in *Proceedings of the AAAI conference on artificial intelligence*, 34(9), pp. 13693-13696.

SUTTON C., MCCALLUM A. (2012). *An introduction to conditional random fields*, in "Foundations and Trends® in Machine Learning", 4(4), pp. 267-373.

TAYLOR A., MARCUS M., SANTORINI B. (2003). *The Penn treebank: an overview*, in A. ABELLÉ (ed.). *Treebanks: Building and using parsed corpora*, Dordrecht, Boston, London, Kluwer Academic Publisher, pp. 5-22.

UNDERWOOD T. (2023). *Using GPT-4 to measure the passage of time in fiction*, published online at:

<https://tedunderwood.com/2023/03/19/using-gpt-4-to-measure-the-passage-of-time-in-fiction/>

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A.N., KAISER Ł., POLOSUKHIN I. (2017). *Attention is all you need*, in "Advances in Neural Information Processing Systems", 30, pp. 5998-6008.

WANG Q., WANG M., FELDTHUS N., OSTERMANN S., CAO Y., SCHÜTZE H., MÖLLER S., SCHMITT V. (2025). *Through a compressed lens: Investigating the impact of quantization on LLM explainability and interpretability*, in "arXiv" preprint:2505.13963.

WEI J., BOSMA M., ZHAO V., GUU K., WEI YU A., LESTER B., DU N., DAI A.M., LE Q.V. (2022). *Finetuned Language Models are Zero-Shot Learners*, in *Proceedings of The Tenth International Conference on Learning Representations*, Virtual.

WEI J., WANG X., SCHUURMANS D., BOSMA M., ICHTER B., XIA F., CHI E., CHI E.H., LE Q.V., ZHOU D. (2022). *Chain-of-thought prompting elicits reasoning in large language models*, in *Proceedings of the Thirty-Sixth Annual Conference on Neural Information Processing Systems*, New Orleans, USA, pp. 24824-24837.

YAO S., YU D., ZHAO J., SHAFRAN I., GRIFFITHS T., CAO Y., NARASIMHAN K. (2023). *Tree of thoughts: Deliberate problem solving with large language models*, in *Proceedings of the Thirty-Seventh Annual Conference on Neural Information Processing Systems*, New Orleans, USA, pp. 11809-11822.

SECTION 2

Creativity, Cognition,
and Human Enhancement

1. When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis*

Michelle Vaccaro, Abdullah Almaatouq, Thomas W. Malone

ABSTRACT

Inspired by the increasing use of artificial intelligence (AI) to augment humans, researchers have studied human-AI systems involving different tasks, systems and populations. Despite such a large body of work, we lack a broad conceptual understanding of when combinations of humans and AI are better than either alone. Here we addressed this question by conducting a preregistered systematic review and meta-analysis of 106 experimental studies reporting 370 effect sizes. We searched an interdisciplinary set of databases (the Association for Computing Machinery Digital Library, the Web of Science and the Association for Information Systems eLibrary) for studies published between 1st January 2020 and 30th June 2023. Each study was required to include an original human-participants experiment that evaluated the performance of humans alone, AI alone and human-AI combinations. First, we found that, on average, human-AI combinations performed significantly worse than the best of humans or AI alone (Hedges' $g = -0.23$; 95% confidence interval, -0.39 to -0.07). Second, we found performance losses in tasks that involved making decisions and significantly greater gains in tasks that involved creating content. Finally, when humans outperformed AI alone, we found performance gains in the combination, but when AI outperformed humans alone, we found losses. Limitations of the evidence assessed here include possible publication bias and variations in the study designs analysed. Overall, these findings highlight the heterogeneity of the effects of human-AI collaboration and point to promising avenues for improving human-AI systems.

1. *Introduction*

People increasingly work with artificial intelligence (AI) tools in fields

* This chapter was originally published as Vaccaro M., Almaatouq A., Malone T. (2024), *When combinations of humans and AI are useful: A systematic review and meta-analysis*, in "Nature Human Behaviour", 8(12), pp. 2293-2303. doi:10.1038/s41562-024-02024-1. This article and the images are licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format. To view a copy of this licence, visit: <http://creativecommons.org/licenses/by/4.0/>

including medicine, finance and law, as well as in daily activities such as travelling, shopping and communicating. These human-AI systems have tremendous potential given the complementary nature of humans and AI – the general intelligence of humans allows us to reason about diverse problems, and the computational power of AI systems allows them to accomplish specific tasks that people find difficult.

A large body of work suggests that integrating human creativity, intuition and contextual understanding with AI's speed, scalability and analytical power can lead to innovative solutions and improved decision-making in areas such as health care (Bohr, Memarzadeh, 2020), customer service (Nicolescu, Tudorache, 2022) and scientific research (Koepnick et al., 2019). However, a growing number of studies reveals that human-AI systems do not necessarily achieve better results than the best of humans or AI alone. Challenges such as communication barriers, trust issues, ethical concerns and the need for effective coordination between humans and AI systems can hinder the collaborative process (Bansal et al., 2021a; Bansal et al., 2019; Buçinca et al., 2020; Lai et al., 2020; Vaccaro, Waldo, 2019; Zhang et al., 2020).

These seemingly contradictory results raise important questions: when do humans and AI complement each other, and by how much? To address these issues, we conducted a systematic literature review and meta-analysis in which we quantified synergy in human-AI systems and identified factors that explain its presence (or absence) in different settings.

We focused on two outcomes: (1) human-AI synergy, where the human-AI group performs better than both the human alone and the AI alone, which is analogous to strong synergy in human groups (Almaatouq et al., 2021; Larson, 2013); and (2) human augmentation, where the human-AI group performs better than the human alone.

When evaluating human-AI systems, many studies focus on human augmentation (Bo et al., 2021; Bondi et al., 2022; Boskemper et al., 2022; Schemmer et al., 2022). This measure can serve important purposes in contexts for which full automation cannot happen for legal, ethical or safety reasons and in cases when AI does not align with human values. But when talking about the potential of human-AI systems, most people implicitly assume that the combined system should be better than either alone; otherwise, they would just use the best of the two (Wilson, Daugherty, 2018). They are thus looking for human-AI synergy. In light of these considerations, a growing body of work emphasizes evaluating and searching for synergy in human-AI systems (Bansal et al., 2019; Bansal et al., 2021a,b; Mozannar et al., 2023; Rastogi et al., 2023; Wilder et al., 2020).

To evaluate synergy in human-AI systems, we analysed 370 unique

effect sizes from 106 different experiments published between January 2020 and June 2023 that included the performance of the human-only, AI-only and human-AI systems. On average, we found evidence of human augmentation, meaning that the average human-AI systems performed better than the human alone. But we did not find human-AI synergy on average, meaning that the average human-AI systems performed worse than at least one of the human alone or the AI alone. So, in practice, if we consider only the performance dimensions the researchers studied, it would have been better to use either a human alone or an AI system alone rather than the human-AI systems studied. While this overall result may appear discouraging, we also identified specific factors that did or did not contribute to synergy in human-AI systems. On the one hand, for example, much of the recent research in human-AI collaboration has focused on using AI systems to help humans make decisions by providing not only suggested decisions but also confidence levels or explanations. But we found that neither of these factors significantly affected the performance of human-AI systems.

On the other hand, little work has investigated the effects of task types and the relative performance of humans alone and AI alone. But we found that both factors significantly affected human-AI performance. Our work thus sheds needed light on promising directions for designing future human-AI systems to unlock the potential for greater synergy.

2. *Results*

Our initial literature search yielded 5,126 papers, and, per the review process described in the Methods (see ‘Literature review’, section 6.1.), we identified 74 that met our inclusion criteria. These papers reported the results of 106 unique experiments, and many of the experiments had multiple conditions, so we collected a total of 370 unique effect sizes measuring the impact of human-AI collaboration on task performance. We synthesized these data in a three-level meta-analytic model (see ‘Data analysis’ in ‘Methods’, section 6.2.). We have made the materials required to reproduce our results publicly accessible through an Open Science Framework repository.

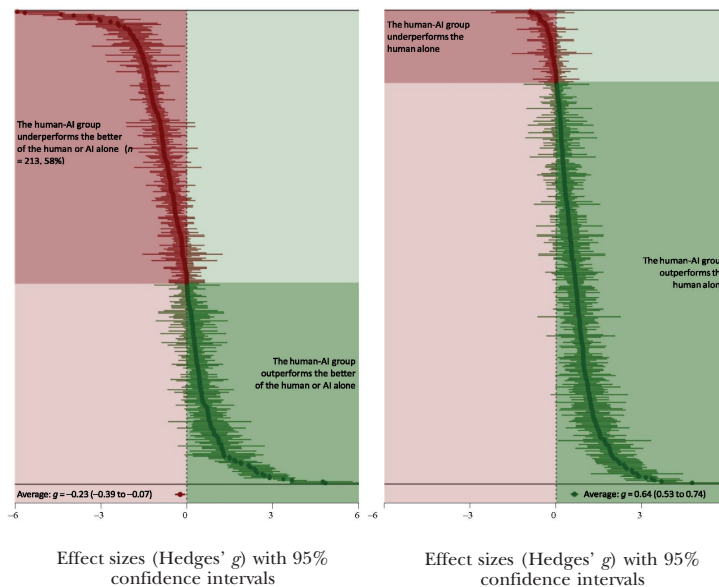
2.1. Overall levels of human-AI synergy

In our primary analyses, we focused on human-AI synergy, so we compared the performance of the human-AI systems to a baseline of the human alone or the AI alone, whichever performed best. We found that

the human-AI systems performed significantly worse overall than this baseline. The overall pooled effect was negative ($g = -0.23$; $t_{92} = -2.89$; two-tailed $P = 0.005$; 95% confidence interval (CI), -0.39 to -0.07) and considered small according to conventional interpretations (Cohen, 2013).

However, when we compared the performance of the human-AI systems to a different baseline – the humans alone – we found substantial evidence of human augmentation. The human-AI systems performed significantly better than humans alone, and this pooled effect size was positive ($g = 0.64$; $t_{98} = 11.87$; two-tailed $P = 0.000$; 95% CI, 0.53 to 0.74) and medium to large (Cohen, 2013). Figure 1 displays a forest plot of these effect sizes. In other words, the human-AI systems we analysed were, on average, better than humans alone but not better than both humans alone and AI alone. For effect sizes that correspond to other potential outcomes of interest.

Figure 1 - Forest plots of all effect sizes ($k = 370$) included in the meta-analysis



The positions of the points on the x axes represent the values of the effect sizes, and the bars indicate the 95% CIs for the effect sizes. The colours of the points and bars correspond to the values of the effect sizes, with negative effect sizes coloured red and positive effect sizes coloured green. The black dashed line corresponds to an effect size of g

= 0, which means that the human-AI system performed the same as the baseline. The point at the bottom of the graph represents the meta-analytic average effect size and CI.

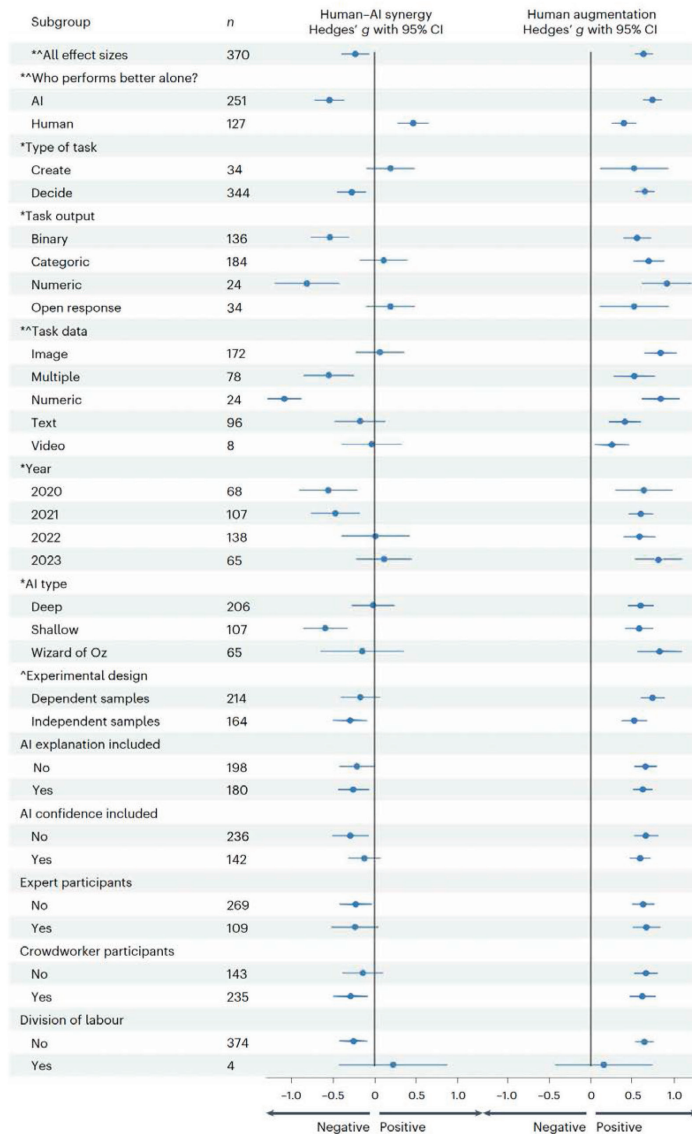
2.2. Heterogeneity of human-AI synergy

We also found evidence for substantial heterogeneity of effect sizes in our estimations of human-AI synergy ($I^2 = 97.7\%$) and human augmentation ($I^2 = 93.8\%$). Our moderator analysis identified characteristics of participants, tasks and experiments that led to different levels of human-AI synergy and human augmentation, and it helps explain sources of this heterogeneity. Figure 2 (on the next page) provides a visualization of the results of meta-regressions with our moderators.

First, we found that the type of task significantly moderated human-AI synergy ($F_{1,104} = 7.84$, two-tailed $P = 0.006$). Among decision tasks – those in which participants decided between a finite set of options – the pooled effect size for human-AI synergy was significantly negative ($g = -0.27$; $t_{104} = -3.20$; two-tailed $P = 0.002$; 95% CI, -0.44 to -0.10), which indicates performance losses from combining humans and AI. In contrast, among creation tasks – those in which participants created some sort of open-response content – the pooled effect size for human-AI synergy was positive ($g = 0.19$; $t_{104} = 1.35$; two-tailed $P = 0.180$; 95% CI, -0.09 to 0.48), pointing to synergy between humans and AI. Even though the average performance gains for creation tasks were not significantly different from 0 (presumably because of the relatively small sample size of $n = 34$), the difference between losses for decision tasks and gains for creation tasks was statistically significant. Relatedly, we found that the type of data involved in the task significantly moderated both human-AI synergy ($F_{4,101} = 15.24$, two-tailed $P = 0.000$) and human augmentation ($F_{4,101} = 6.52$, two-tailed $P = 0.000$).

Second, we found that the performance of the human and AI relative to each other impacted both human-AI synergy ($F_{1,104} = 81.79$, two-tailed $P = 0.000$) and human augmentation ($F_{1,104} = 24.35$, two-tailed $P = 0.000$). As shown in Figure 2, when the human alone outperformed the AI alone, the combined human-AI system outperformed both alone with an average pooled effect size for human-AI synergy of $g = 0.46$ ($t_{104} = 5.06$; two-tailed $P = 0.000$; 95% CI, 0.28 to 0.66), a medium-sized effect (Cohen, 2013). But when the AI alone outperformed the human alone, performance losses occurred in the combined system relative to the AI alone, with a negative effect size for human-AI synergy of $g = -0.54$ ($t_{104} = -6.20$; two-tailed $P = 0.000$; 95% CI, -0.71 to -0.37), a medium-sized effect.

Figure 2 - Results from the three-level meta-regression models for the moderator variables



Here n is the number of included effect sizes for the moderator subgroup level, and the estimated effect size with the corresponding 95% CI is shown for each level. The symbols in front of the moderators indicate whether there is a statistically significant difference between the subgroups for human-AI synergy (*) and human augmentation (Δ).

Given the importance of this moderator, we fit separate meta-analytic models on the subsets of results where (1) the AI performed better alone and (2) the human performed better alone.

The performance of the human and AI relative to each other also affected the degree of human augmentation in the human-AI systems ($F_{1,104} = 24.35$, two-tailed $P = 0.000$). When the AI outperformed the human alone, greater performance gains tended to occur in the human-AI systems relative to the human alone, and the pooled effect size for human augmentation was positive and medium to large in magnitude ($g = 0.74$; $t_{104} = 13.50$; two-tailed $P = 0.000$; 95% CI, 0.63 to 0.85).

We also found that the type of AI involved in the experiment ($F_{2,103} = 3.77$, two-tailed $P = 0.026$) and the year of publication ($F_{3,102} = 3.65$, two-tailed $P = 0.015$) moderated human-AI synergy, and the experimental design moderated human augmentation ($F_{1,104} = 4.90$, two-tailed $P = 0.029$). The remaining moderators we investigated were not statistically significant for human-AI synergy or human augmentation (explanation, confidence, participant type and division of labour).

3. Discussion

Systems that combine human intelligence and AI tools can address multiple issues of societal importance, from how we diagnose disease to how we design complex systems (Chen et al., 2023; Noti, Chen, 2023; Shin et al., 2021). But some studies show that augmenting humans with AI can lead to better outcomes than humans or AI working alone (Chen et al., 2023; Liu et al., 2021; Reverberi et al., 2022), while others show the opposite (Bansal et al., 2021a; Vaccaro, Waldo, 2019; Zhang et al., 2020). These seemingly disparate results raise two important questions: How effective is human-AI collaboration in general? And under what circumstances does this collaboration lead to performance gains versus losses? Our study analyses over three years of recent research to provide insights into both of these questions.

3.1. Performance losses from human-AI collaboration

Regarding the first question, we found that, on average among recent experiments, human-AI systems did not exhibit synergy: the human-AI groups performed worse than either the human alone or the AI alone. This result complements the qualitative literature reviews on human-AI collaboration (Lai et al., 2023; Maadi et al., 2021; Sperrle et al., 2021) which highlight some of the surprising challenges that arise when

integrating human intelligence and AI. For example, people often rely too much on AI systems (overreliance), using its suggestions as strong guidelines without seeking and processing more information (Buçinca et al., 2021; Lai et al., 2020; Skitka et al., 1999). Other times, however, humans rely too little on AI (underreliance), ignoring its suggestions because of adverse attitudes towards automation (Buçinca et al., 2021; Vasconcelos et al., 2023; Zhang et al., 2020).

Interestingly, we found that, among this same set of experiments, human augmentation did exist in the human-AI systems: the human-AI groups performed better than the humans working alone. Thus, even though the human-AI combinations did not achieve synergy on average, the AI system did on average help humans perform better. This result can occur, of course, because by definition the baseline for human-AI synergy is more stringent than that for human augmentation. It may also occur, however, because obtaining human-AI synergy requires different forms of human-AI interaction, or because the recent empirical studies were not appropriately designed to elicit human-AI synergy.

3.2. Moderating effect of task type

With the large dataset we collected, we also performed analyses of factors that influence the effectiveness of human-AI collaboration. We found that the type of task significantly moderated synergy in human-AI systems: decision tasks were associated with performance losses, and creation tasks were associated with performance gains.

We hypothesize that this advantage for creation tasks occurs because even when creation tasks require the use of creativity, knowledge or insight for which humans perform better, they often also involve substantial amounts of somewhat routine generation of additional content that AI can perform as well as or better than humans. For instance, generating a good artistic image usually requires some creative inspiration about what the image should look like, but it also often requires a fair amount of more routine fleshing out of the details of the image. Similarly, generating many kinds of text documents often requires knowledge or insight that humans have and computers do not, but it also often requires filling in boilerplate or routine parts of the text as well. With most of the decision tasks studied in our sample, however, both the human and the AI system make a complete decision, with the humans usually making the final choice.

Our results suggest that with these decision tasks, better results might have been obtained if the experimenters had designed processes in which the AI systems did only the parts of the task for which they were

clearly better than humans. Only 3 of the 100+ experiments in our analysis explore such processes with a predetermined delegation of separate subtasks to humans and AI. With the four effect sizes from these 3 experiments, we found that, on average, human-AI synergy ($g = 0.22$, $t_{104} = 0.69$; two-tailed $P = 0.494$; 95% CI, -0.42 to 0.87) occurred, but the result was not statistically significant.

3.3. Moderating effect of relative human/AI performance

Interestingly, when the AI alone outperformed the human alone, substantial performance losses occurred in the human-AI systems. When the human outperformed the AI alone, however, performance gains occurred in the human-AI systems. This finding shows that human-AI performance cannot be explained with a simple average of the human alone and AI alone. In such a case, human-AI synergy could never exist (Donahue et al., 2022).

Most ($> 95\%$) of the human-AI systems in our dataset involved humans making the final decisions after receiving input from AI algorithms. In these cases, one potential explanation of our result is that, when the humans are better than the algorithms overall, they are also better at deciding in which cases to trust their own opinions and in which to rely more on the algorithm's opinions.

For example, Cabrera et al. (2023) used an experimental design in which participants in the human-AI condition saw a problem instance, an AI prediction for that instance and, in some cases, additional descriptions of the accuracy of the AI in this type of instance. The same experimental design, with the same task interface, participant pool and accuracy of the AI system, was used for three separate tasks: fake hotel review detection, satellite image classification and bird image classification. For fake hotel review detection, the researchers found that the AI alone achieved an accuracy of 73%, the human alone achieved an accuracy of 55% and the human-AI system achieved an accuracy of 69%. In this case, we hypothesize that, since the people were less accurate, in general, than the AI algorithms, they were also not good at deciding when to trust the algorithms and when to trust their own judgement, so their participation resulted in lower overall performance than for the AI algorithm alone.

In contrast, Cabrera et al. (2023) found that, for bird image classification, the AI alone achieved an accuracy of 73%, the human alone achieved an accuracy of 81% and the human-AI system achieved an accuracy of 90%. Here, the humans alone were more accurate than the AI algorithms alone, so we hypothesize that the humans were good at deciding when to trust their own judgements versus those of the

algorithms, and the overall performance thus improved over either humans or AI alone.

3.4. Surprisingly insignificant moderators

We also investigated other moderators such as the presence of an explanation, the inclusion of the confidence of the AI output and the type of participant evaluated. These factors have received much attention in recent years (Bansal et al., 2021a; Chen et al., 2023; Liu et al., 2021; Rastogi et al., 2022). Given our result that, on average across our 300+ effect sizes, they do not impact the effectiveness of human-AI collaboration, we think researchers may wish to de-emphasize this line of inquiry and instead shift focus to the significant and less researched moderators we identified: the baseline performance of the human and AI alone, the type of task they perform, and the division of labour between them.

4. *Limitations*

We want to highlight some general limitations of our meta-analytic approach to aid with the interpretation of our results. First, our quantitative results apply to the subset of studies we collected through our systematic literature review. To evaluate human-AI synergy, we required that papers report the performance of (1) the human alone, (2) the AI alone and (3) the human-AI system. We can, however, imagine tasks that a human and/or AI cannot perform alone but can when working with the other. Our analysis does not include such studies.

Second, we calculated effect sizes that correspond to different quantitative measures such as task accuracy, error and quality. By computing Hedges' g , a unitless standardized effect size, we can describe important relations among these experiments in ways that make them comparable across different study designs with different outcome variables (Hedges, 2008). The studies in our dataset, though, come from different samples of people – some look at doctors (Jacobs et al., 2021; Jussupow et al., 2021; Tschandl et al., 2020), others at crowdworkers (Bansal et al., 2021a; Lai et al., 2020; Cabrera et al., 2023) others at students (Boskemper et al., 2022; He et al., 2023; Liang et al., 2022) – and this variation can limit the comparability of the effect sizes to a degree (Hedges, 2008). The measurement error can also vary across experiments. For example, some studies estimate overall accuracy on the basis of the evaluation of as many as 500 distinct images (Reverberi et al., 2022), whereas others estimate it on the basis of the evaluation of as few as 15 distinct ones (Papenmeier et al., 2022).

As is typical for meta-analyses (Borenstein et al., 2021), in our three-level model, we weighted effect sizes as a function of their variance across participants, so we did not account for this other source of variation in measurement.

Third, although we did not find evidence of publication biases, it remains possible that they exist, which would impact our literature base and, by extension, our meta-analytic results. However, we expect that if there were a publication bias operating here, it would be a bias to publish studies that showed significant gains from combining humans and AI. And since our overall results showed the opposite, it seems unlikely that they are a result of publication bias.

Fourth, our results only apply to the tasks, processes and participant pools that researchers have chosen to study, and these configurations may not be representative of the ways human-AI systems are configured in practical uses of AI outside the laboratory. In other words, even if there is not a publication bias in the studies we analysed, there might be a research topic selection bias at work.

Fifth, the quality of our analysis depends on the quality of the studies we synthesized. We tried to control for this issue by only including studies published in peer-reviewed publications, but the rigour of the studies may still vary in degree. For example, studies used different attention check mechanisms and performance incentive structures, which can both affect the quality of responses and thus introduce another source of noise into our data.

Finally, we found a high level of heterogeneity among the effect sizes in our analysis. The moderators we investigated account for some of this heterogeneity, but much remains unexplained. We hypothesize that interaction effects exist between the variables we coded (for example, explanation and type of AI), but we do not have enough studies to detect such effects. There are also certainly potential moderators that we did not analyse. For example, researchers mostly used their own experimental platforms and stimuli, which naturally introduce sources of variation between their studies. As the human-AI collaboration literature develops, we hope future work can identify more factors that influence human-AI synergy and assess the interactions among them.

5. A roadmap for future work: finding human-AI synergy

Even though our main result suggests that – on average – combining humans and AI leads to performance losses, we do not think this means that combining humans and AI is a bad idea. On the contrary, we think it just means that future work needs to focus more specifically on

finding effective processes that integrate humans and AI. Our other results suggest promising ways to proceed.

5.1. Develop generative AI for creation tasks

In our broad sample of recent experiments, the vast majority (about 85%) of the effect sizes were for decision-making tasks in which participants chose among a predefined set of options. But in these cases we found that the average effect size for human-AI synergy was significantly negative. In contrast, only about 10% of the effect sizes researchers studied were for creation tasks – those that involved open-ended responses. And in these cases we found that the average effect size for human-AI synergy was positive and significantly greater than that for decision tasks. This result suggests that studying human-AI synergy for creation tasks – many of which can be done with generative AI – could be an especially fruitful area for research.

Much of the recent work on generative AI with human participants, however, tends to focus on attitudes towards the tool (Karinshak et al., 2023; Wilcox et al., 2023), interviews or think-alouds with participants (Jo et al., 2023; Liu et al., 2023; Vimpri et al., 2023), or user experience instead of task performance (Jakesch et al., 2023; Mirowski et al., 2023; Petridis et al., 2023). Furthermore, the relatively little work that does evaluate human-AI collaboration according to quantitative performance metrics tends to report only the performance of the human alone and that of the human-AI combination (not the AI alone, Noy, Zhang, 2023). This limitation makes evaluating human-AI synergy difficult, as the AI alone may be able to perform the task at a higher quality and speed than the participants involved in the experiment (typically crowdworkers). We thus need studies that further explore human-AI collaboration across diverse tasks while reporting the performance of the human alone, AI alone and human-AI system.

5.2. Develop innovative processes

Additionally, as discussed in Donahue et al. (2022), human-AI synergy requires that humans be better at some parts of a task, AI be better at other parts of the task and the system as a whole be good at appropriately allocating subtasks to whichever partner is best for that subtask. Sometimes that is done by letting the more capable partner decide how to allocate subtasks, and sometimes it is done by assigning different subtasks a priori to the most capable partner. In general, to effectively use AI in practice, it may be just as important to design innovative processes for how to combine humans and AI as it is to design innovative technologies (Baier et al., 2024).

5.3. Develop more robust evaluation metrics for human-AI systems

Many of the experiments in our analysis evaluate performance according to a single measure of overall accuracy, but this measure corresponds to different things depending on the situation, and it omits other important criteria for human-AI systems. For example, as one approaches the upper bound of performance, such as 100% accuracy, the improvements necessary to increase performance usually become more difficult for both humans and AI systems. In these cases, we may wish to consider a metric that applies a nonlinear scaling to the overall classification accuracy and thus takes such considerations into account (Baier et al., 2024).

More importantly, there are many practical situations where good performance depends on multiple criteria. For instance, in many high-stakes settings such as radiology diagnoses and bail predictions, relatively rare errors may have extremely high financial or other costs. In these cases, even if AI can, on average, perform a task more accurately and less expensively than humans, it may still be desirable to include humans in the process if the humans are able to reduce the number of rare but very undesirable errors. One potential approach for situations like these is to create composite performance measures that incorporate the expected costs of various kinds of errors. The human augmentation measure described is also appropriate for these high-stakes settings.

In general, we encourage researchers to develop, employ and report more robust metrics that consider factors such as task completion time, financial cost and the practical implications of different types of errors. These developments will help us better understand the significance of improvements in task performance as well as the effects of human-AI collaborations.

5.4. Develop commensurability criteria

As researchers continue to study human-AI collaboration, we also urge the field to develop a set of commensurability criteria, which can facilitate more systematic comparisons across studies and help us track progress in finding areas of human-AI synergy. These criteria could provide standardized guide-lines for key study design elements such as:

- (1) Task designs: establish a set of benchmark tasks that involve human-AI systems.
- (2) Quality constraints: specify acceptable quality thresholds or requirements that the human, AI and human-AI system must meet.
- (3) Incentive schemes: outline incentive structures (for example,

payment schemes and bonuses) used to motivate human participants.

- (4) Process types: develop a taxonomy of interaction protocols, user interface designs and task workflows for effective human-AI collaboration.
- (5) Evaluation metrics: report the performance of the human, AI and human-AI system according to well-defined performance metrics.

To further promote commensurability and research synthesis, we encourage the field to establish a standardized and open reporting repository, specifically for human-AI collaboration experiments. This centralized database would host the studies' raw data, code, system outputs, interaction logs and detailed documentation, adhering to the proposed reporting guidelines. It would thus facilitate the replication, extension and synthesis of research in the field. For example, by applying advanced machine learning techniques on such a dataset, we could develop predictive models to guide the design of human-AI systems optimized for specific constraints and contexts. Additionally, it would provide a means to track progress in finding greater areas of human-AI synergy.

In conclusion, our results demonstrate that human-AI systems often perform worse than humans alone or AI alone. But our analysis also suggests promising directions for the future development of more effective human-AI systems. We hope that this work will help guide progress in developing such systems and using them to solve some of our most important problems in business, science and society.

6. *Methods*

We conducted this meta-analysis in accord with the guidelines from Kitchenham (2004) on systematic reviews, and we followed the standards set forth by the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (Moher et al., 2009).

6.1. Literature review

6.1.1. Eligibility criteria

Per our preregistration (https://osf.io/wrq7c/?view_only=b9e1e86079c048b4bfb03bee6966e560), we applied the following criteria to select studies that fit our research questions. First, the paper needed to present an original experiment that evaluated some instance in which a human and an AI system worked together to perform a task. Second, it needed to report the performance of (1) the human alone,

(2) the AI alone and (3) the human-AI system according to some quantitative measure(s). We therefore excluded studies that reported the performance of the human alone but not the AI alone, and likewise we excluded studies that reported the performance of the AI alone but not the human alone. Following this stipulation, we also excluded pure meta-analyses and literature reviews, theoretical work, qualitative analyses, commentaries, opinions and simulations. Third, we required the paper to include the experimental design, the number of participants in each condition and the standard deviation of the outcome in each condition, or enough information to calculate it from other quantities. Finally, we required the paper to be written in English.

6.1.2. Search strategy

Given the interdisciplinary nature of human-AI interaction studies, we performed this search in multiple databases covering conferences and journals in the computer sciences, information sciences and social sciences, as well as other fields. Through consultation with a library specialist in these fields, we decided to target the Association for Computing Machinery Digital Library, the Association for Information Systems eLibrary and the Web of Science Core Collection for our review. To focus on current forms of AI, we limited the search to studies published between 1st January 2020 and 30th June 2023.

To develop the search string, we began by distilling the facets of studies that evaluated the performance of a human-AI system. We required the following: (1) a human component, (2) a computer component, (3) a collaboration component and (4) an experiment component. Given the multidisciplinary nature of our research question, papers published in different venues tended to refer to these components under various names (Bansal et al., 2019; Bo et al., 2021; Boskemper et al., 2022; Groh et al., 2022; Tejeda et al., 2022). For broad coverage, we compiled a list of synonyms and abbreviations for each component and then combined these terms with Boolean operations, resulting in the following search string:

- (1) Human: human OR expert OR participant OR humans OR experts OR participants
- (2) AI: AI OR ‘artificial intelligence’ OR ML OR ‘machine learning’ OR ‘deep learning’
- (3) Collaboration: collaborate OR assist OR aid OR interact OR help
- (4) Experiment: ‘experiment’ OR ‘experiments’ OR ‘user study’ OR ‘user studies’ OR ‘crowdsourced study’ OR ‘crowdsourced studies’ OR ‘laboratory study’ OR ‘laboratory studies’.

The search term was (1) AND (2) AND (3) AND (4) in the abstract. To

further ensure comprehensive coverage, we also conducted a forward and backward search on all studies we found that met our inclusion criteria.

6.1.3. Data collection and coding

We conducted the search in each of these databases in July 2023. To calculate our primary outcome of interest – the effect of combining human intelligence and AI on task performance – we recorded the averages and standard deviations of the task performance of the human alone, the AI alone, and the human and AI working with each other, as well as the number of participants in each of these conditions.

Many authors reported all of these values directly in the text of the paper. A notable number, however, reported them indirectly by providing 95% CIs or standard errors instead of the raw standard deviations. For these, we calculated the standard deviations using the appropriate formulas (Higgins et al., 2024). Additionally, multiple papers did not provide the exact numbers needed for such formulas, but the authors made the raw data of their study publicly accessible. In these cases, we downloaded the datasets and computed the averages and standard deviations using Python (v.3.11.9) or R (v.2023.06.0+421). If relevant data were only presented in the plots of a paper, we contacted the corresponding author to ask for the numeric values. If the authors did not respond, we used WebPlotDigitizer (Rohatgi, 2020) to convert plotted values into numerical values. For papers that conducted an experiment that met our inclusion criteria but did not report all the values needed to calculate the effect size, we also contacted the corresponding author directly to ask for the necessary information. If the author did not respond, we could not compute the effect size for the study and did not include it in our analysis.

We considered and coded for multiple potential moderators of human-AI performance: (1) publication date, (2) preregistration status, (3) experimental design, (4) data type, (5) task type, (6) task output, (7) AI type, (8) AI explanation, (9) AI confidence, (10) participant type and (11) performance metric.

The additional information we recorded served descriptive and exploratory purposes. Many papers conducted multiple experiments, contained multiple treatments or evaluated performance according to multiple measures. In such cases, we assigned a unique experiment identification number, treatment identification number and measure identification number to the effect sizes from the paper. Note that we defined experiments on the basis of samples of different sets of participants.

6.2. Data analysis

6.2.1. Calculation of effect size

We computed Hedges' g to measure the effect of combining human intelligence and AI on task performance (Hedges, 1981). For strong synergy, Hedges' g represents the standardized mean difference between the performance of the human-AI system and that of the baseline, which can be the human alone or AI alone, whichever performs better on average. For human augmentation, Hedges' g represents the standardized mean difference between the performance of the human-AI system and the baseline of the human alone.

We chose Hedges' g as our measure of effect size because it is unitless, so it allows us to compare human-AI performance across different metrics, and it corrects for upward bias in the raw standardized mean difference (Cohen's d , Hedges, 1981).

6.2.2. Meta-analytic model

The experiments from the papers we analysed varied considerably. For example, they evaluated different tasks, recruited participants from different backgrounds and employed different experimental designs. Since we expected substantial between-experiment heterogeneity in the true effect sizes, for our analysis we used a random-effects model that accounts for variance in effect sizes that comes from both sampling error and 'true' variability across experiments (Hedges, Olkin, 2014).

Additionally, some of the experiments we considered generated multiple dependent effect sizes: they could involve multiple treatment groups, and they could assess performance according to more than one measure, for example. The more commonly used meta-analytic models assume independence of effect sizes, which makes them unsuitable for our analysis (Van den Noortgate et al., 2015). We thus adopted a three-level meta-analytic model in which effect sizes are nested within experiments, so we explicitly took the dependencies in our data into account in the model (Cheung, 2014; Van den Noortgate et al., 2015). Furthermore, we used robust variance estimate methods to compute consistent estimates of the variance of our effect size estimates and, relatedly, standard errors, CIs and statistical tests, which account for the dependency of sampling errors from overlapping samples that occurred in experiments that compared multiple treatment groups to a single control group (Hedges et al., 2010). When evaluating the significance of our results, we applied the Knapp and Hartung adjustment and computed a test statistic, standard error, P value and CI based on the t distribution with $k - p$ degrees of freedom, where k denotes the number

of effect size clusters (that is, the number of experiments) and p denotes the number of coefficients in the model. To perform our moderator analyses, we conducted separate meta-regressions for each of our moderator variables.

To interpret the degree of heterogeneity in our effect sizes, we calculated the popular I^2 statistic following Higgins and Thompson (2002) which quantifies the percentage of variation in effect sizes that is not from sampling error. Furthermore, to distinguish the sources of this heterogeneity, we also calculated multilevel versions of the statistic, following Cheung (2014).

6.2.3. Bias tests

In the context of our meta-analysis, publication bias may occur if researchers publish experiments that show evidence of significant human-AI synergy more frequently than those that do not. Such actions would affect the data we collected and distort our findings. To evaluate this risk, we adopted multiple diagnostic procedures outlined by Viechtbauer and Cheung (2010). First, we created funnel plots that graph the observed effect sizes on the x axis and the corresponding standard errors on the y axis (Sterne, Egger, 2001). In the absence of publication bias, we expect the points to fall roughly symmetrically around the y axis. We enhanced these plots with colors indicating the significance level of each effect size to help distinguish publication bias from other causes of asymmetry (Peters et al., 2008). A lack of effect sizes in regions of statistical non-significance points to a greater risk of publication bias. We visually inspected the plots and performed Egger's regression test (Egger et al., 1997) as well as the rank correlation test (Begg, Mazumdar, 1994) to evaluate the results in the funnel plots. These tests examine the correlation between the observed effect sizes and their associated sampling variances. A high correlation indicates asymmetry in the funnel plot, which may stem from publication bias. Egger's regression did not indicate evidence of publication bias in the sample ($\beta = -0.67$; $t_{104} = -0.78$; two-tailed $P = 0.438$; 95% CI, -2.39 to 1.04), nor did the rank correlation test ($\tau = 0.05$, two-tailed $P = 0.121$). Taken as a whole, these tests suggest that our results for human-AI synergy are robust to potential publication bias.

Importantly, however, we did find potential evidence of publication bias in favour of studies that report results in which the human-AI system outperforms the human alone (human augmentation). In this case, Egger's regression does point to publication bias in the sample ($\beta = 1.96$; $t_{104} = 3.24$; two-tailed $P = 0.002$; 95% CI, 0.76 to 3.16), as does the rank correlation test ($\tau = 0.19$, two-tailed $P = 0.000$). Note that we did

not try to correct for potential publication bias to preserve the integrity of the original data and maintain transparency in our reporting. Many proposed adjustment methods can also lead to overcorrection and distort results (Rothstein et al., 2005).

The discrepancy between the symmetry in the funnel plot for human-AI synergy versus asymmetry in the funnel plot for human augmentation may reflect how many researchers and journals implicitly assume an interest in human augmentation, comparing the human-AI system to the human alone.

6.3. Sensitivity analysis

For our primary analysis, we developed a three-level meta-analytic model that accounted for variance in the observed effect sizes (first level), variance between effect sizes from the same experiment (second level) and variance between experiments (third level). We then calculated cluster-robust standard errors, CIs and statistical tests for our effect size estimates in which we defined clusters at the level of the experiment. This model accounts for the dependency in effect sizes that result from evaluating more than one treatment against a common control group and assessing performance according to more than one measure. It does, however, consider the experiments in our analysis as independent from each other, even if they come from the same paper. We find this assumption plausible because the experiments recruited different sets of participants and entailed different tasks or interventions.

As a robustness check, though, we performed a sensitivity re-analysis in which we clustered at the paper level instead of the experiment level. This multilevel model accounts for variance in the observed effect sizes (first level), variance between effect sizes from the same paper (second level) and variance between papers (third level). We still calculated cluster-robust standard errors, CIs and statistical tests for our effect size estimates in which we defined clusters at the level of the experiment because the participant samples overlapped only on the level of the experiment. Using this approach, we found a comparable overall effect size for human-AI synergy ($g = -0.22$; $t_{67} = -2.46$; two-tailed $P = 0.017$; 95% CI, -0.41 to -0.04) and for human augmentation ($g = 0.65$; $t_{69} = 9.96$; two-tailed $P = 0.000$; 95% CI, 0.52 to 0.78).

We also evaluated the robustness of our results to outlying and influential data points. To detect such data, we computed the residuals and Cook's distance for each effect size. We considered residual values greater or less than three standard deviations from the mean as outliers, and following Sterne and Egger (2001) we considered values greater than $4/n$ as high influence, where n is the number of data points in our

analysis. Using this approach, we identified 11 outliers for human-AI synergy and 9 outliers for human augmentation. We performed a sensitivity re-analysis on a dataset excluding these effect sizes, which resulted in similar effect sizes for human-AI synergy ($g = -0.25$; $t_{104} = -3.45$; two-tailed $P = 0.001$; 95% CI, -0.39 to -0.11) and human augmentation ($g = 0.60$; $t_{104} = 12.60$; two-tailed $P = 0.000$; 95% CI, 0.50 to 0.69).

Additionally, we conducted leave-one-out analyses, in which we performed a series of sensitivity re-analyses on the different subsets of effect sizes obtained by leaving out one effect size in our original dataset. We also conducted leave-one-out analyses at the experiment and publication levels. These tests show how each effect size, experiment and publication affect our overall estimate of the effect of human-AI collaboration on task performance. Summary effect sizes for human-AI synergy ranged from -0.28 to -0.19 with two-tailed $P < 0.05$ (0.000 to 0.019) in all cases, indicating the robustness of our results to any single effect size, experiment or paper; likewise, summary effect sizes for human augmentation ranged from 0.61 to 0.66 with two-tailed $P < 0.05$ (0.000 to 0.000) in all cases, indicating the robustness of our results to any single effect size, experiment or paper.

Lastly, we conducted a sensitivity re-analysis on a dataset that omits the effect sizes we estimated, using either WebPlotDigitizer or the information provided by the authors in their paper, and again we found almost identical results for human-AI synergy ($g = -0.21$; $t_{98} = -2.59$; two-tailed $P = 0.011$; 95% CI, -0.36 to -0.05) and human augmentation ($g = 0.64$; $t_{98} = 11.73$; two-tailed $P = 0.000$; 95% CI, 0.53 to 0.75).

We performed all quantitative analysis with the R statistical programming language, and we primarily relied on the package metafor (Viechtbauer, 2010).

7. Data availability

We compiled the data used in this analysis based on the studies identified in our systematic literature review. We have made the data we collected available via the project's Open Science Framework repository (https://osf.io/wrq7c/?view_only=b9e1e86079c048b4bfb03bee6966e560). In our systematic review, we searched the following databases: the Association for Computing Machinery Digital Library (<https://dl.acm.org/>), the Web of Science (seg.sci@auxologico.it) and the Association for Information Systems eLibrary (<https://aisnet.org/page/AISeLibrary>).

8. *Code availability*

We have shared the code used to conduct our analysis via the Open Science Framework repository (https://osf.io/wrq7c/?view_only=b9e1e86079c048b4bfb03bee6966e560).

Acknowledgments

M.V. was funded by the Accenture Technology Convergence Fellowship. Additional funding to T.M. was provided by the Toyota Research Institute, by the MIT Quest for Intelligence and by the National Research Foundation, Prime Minister's Office, Singapore, under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank J. Kim and G. Campagna for their help collecting the data for this project, and we thank D. Eckles, M. Alsobay, R. Na and E. Hu for their helpful feedback.

Author contributions

M.V. conceived the study idea with feedback from A.A. and T.M.; M.V. collected the data and performed the statistical analysis, with feedback from A.A. and T.M.; M.V., A.A. and T.M. wrote the manuscript.

References

- ALMAATOUQ A., ALSOBAY M., YIN M., WATTS D.J. (2021). *Task complexity moderates group synergy*, in "Proceedings of the National Academy of Sciences of the USA", 118, e2101062118.
- ALTMAN N., KRZYWINSKI M. (2016). *Analyzing outliers: Influential or nuisance?*, in "Nature Methods", 13, pp. 281-283.
- BAIER P., DELALLO D., SVIOKLA J.J. (2024). *Your organization isn't designed to work with GenAI*, in "Harvard Business Review". <https://hbr.org/2024/02/your-organization-isnt-designed-to-work-with-genai>
- BANSAL G., NUSHI B., KAMAR E., HORVITZ E., LIAO Q.V. (2019). *Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff*, in Proceedings of the 33rd AAAI Conference on Artificial Intelligence and 31st Innovative Applications of Artificial Intelligence Conference and 9th AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI Press, pp. 2429-2437. <https://doi.org/10.1609/aaai.v33i01.33012429>

BANSAL G., NGUYEN T., LIAO Q.V., SRIVASTAVA M. (2021a). *Does the whole exceed its parts? The effect of AI explanations on complementary team performance*, in “Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems”, Association for Computing Machinery, p. 81.

BANSAL G., NUSHI B., KAMAR E., HORVITZ E., WELD D.S. (2021b). *Is the most accurate AI the best teammate? Optimizing AI for teamwork*, in “Proceedings of the AAAI Conference on Artificial Intelligence”, 35, pp. 11405-11414.

BEGG C.B., MAZUMDAR M. (1994). *Operating characteristics of a rank correlation test for publication bias*, in “Biometrics”, 50, pp. 1088-1101.

BO Z.-H., ZHOU M., XIE D., LI G. (2021). *Toward human intervention-free clinical diagnosis of intracranial aneurysm via deep neural network*, in “Patterns”, 2, 100197.

BOHR A., MEMARZADEH K. (2020) (eds.), *Artificial Intelligence in Healthcare*, Elsevier, pp. 25-60.

BONDI E., MALIK A., TAMBE M., RICE E. (2022). *Role of human-AI interaction in selective prediction*, in “Proceedings of the AAAI Conference on Artificial Intelligence”, 36, pp. 5286-5294.

BORENSTEIN M., HEDGES L.V., HIGGINS J.P., ROTHSTEIN H.R. (2021). *Introduction to meta-analysis*, John Wiley & Sons.

BOSKEMPER M.M., BARTLETT M.L., MCCARLEY J.S. (2022). *Measuring the efficiency of automation-aided performance in a simulated baggage screening task*, in “Human Factors”, 64, pp. 945-961.

BUÇINCA Z., LIN P., GAJOS K.Z., GLASSMAN E.L. (2020). *Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems*, in “Proceedings of the 25th International Conference on Intelligent User Interfaces”, Association for Computing Machinery, pp. 454-464. <https://doi.org/10.1145/3377325.3377498>

BUÇINCA Z., MALAYA M.B., GAJOS K.Z. (2021). *To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making*, in “Proceedings of the ACM on Human-Computer Interaction”, 5, 188.

CABRERA A.A., PERER A., HONG J.I. (2023). *Improving human-AI collaboration with descriptions of AI behavior*, in “Proceedings of the ACM on Human-Computer Interaction”, 7, Article 131-164. <https://doi.org/10.1145/3579612>

CAMPERO A., LAFO M., VOJTĚCH K., ADAMS M. (2022). *A test for evaluating performance in human-computer systems*, in “arXiv preprint arXiv:2206.12390”.

CHEN V., LIAO Q.V., WORTMAN VAUGHAN J., BANSAL G. (2023). *Understanding the role of human intuition on reliance in human-AI decision-making with explanations*, in “Proceedings of the ACM on Human-Computer Interaction”, 7, 370.

CHEUNG M.W.-L. (2014). *Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach*, in “Psychological Methods”, 19, pp. 211-229.

COHEN J. (2013). *Statistical power analysis for the behavioral sciences*, Academic Press, 2nd ed.

DONAHUE K., CHOULDECHOVA A., KENTHAPADI K. (2022). *Human-algorithm collaboration: Achieving complementarity and avoiding unfairness*, in “Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency”, Association for Computing Machinery, pp. 1639-1656. <https://doi.org/10.1145/3531146.3533221>

EGGER M., SMITH G.D., SCHNEIDER M., MINDER C. (1997). *Bias in meta-analysis detected by a simple, graphical test*, in “BMJ”, 315, pp. 629-634.

GROH M., EPSTEIN Z., FIRESTONE C., PICARD R. (2022). *Deepfake detection by human crowds, machines, and machine-informed crowds*, in “Proceedings of the National Academy of Sciences of the USA”, 119, e2110013119.

HE Z., SONG Y., ZHOU S., CAI Z. (2023). *Interaction of thoughts: Towards mediating task assignment in human-AI cooperation with a capability-aware shared mental model*, in “Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems”, Association for Computing Machinery, p. 353. <https://doi.org/10.1145/3544548.3580983>

HEDGES L.V. (1981). *Distribution theory for Glass’s estimator of effect size and related estimators*, in “Journal of Educational Statistics”, 6, pp. 107-128.

HEDGES L.V. (2008). *What are effect sizes and why do we need them*, in “Child Development Perspectives”, 2, pp. 167-171.

HEDGES L.V., OLKIN I. (2014). *Statistical methods for meta-analysis*, Academic Press.

HEDGES L.V., TIPTON E., JOHNSON M.C. (2010). *Robust variance estimation in meta-regression with dependent effect size estimates*, in “Research Synthesis Methods”, 1, pp. 39-65.

HIGGINS J.P.T., LI T., DEEKS J.J. (2024). *Choosing effect measures and computing estimates of effect*, in J.P.T. HIGGINS ET AL. (eds.). *Cochrane Handbook for Systematic Reviews of Interventions*, Cochrane, Version 6.5, Ch. 6.

HIGGINS J.P.T., THOMPSON S.G. (2002). *Quantifying heterogeneity in a meta-analysis*, in “Statistics in Medicine”, 21, pp. 1539-1558.

JACOBS M., PRABHAKARAN R., VASANAWALA S., FLUHARTY M. (2021). *How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection*, in “Translational Psychiatry”, 11, 108.

JAKESCH M., BHAT A., BUSCHEK D., ZALMANSON L., NAAMAN M. (2023). *Co-writing with opinionated language models affects users’ views*, in “Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems”, Association for Computing Machinery, p. 111.

JO E., EPSTEIN D.A., JUNG H., KIM Y.-H. (2023). *Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention*, in “Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems”, Association for Computing Machinery, p. 18.

JUSSUPOW E., SPOHRER K., HEINZL A., GAWLITZA J. (2021). *Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence*, in "Information Systems Research", 32, 713-735.

KARINSHAK E., LIU S.X., PARK J.S., HANCOCK J.T. (2023). *Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages*, in "Proceedings of the ACM on Human-Computer Interaction", 7, 116.

KITCHENHAM B. (2004). *Procedures for performing systematic reviews* (Technical Report TR/SE-0401), Keele University.

KOEPNICK B., COOPER S., TREUILLE A., BAKER D. (2019). *De novo protein design by citizen scientists*, in "Nature", 570, pp. 390-394.

LAI V., CHEN C., SMITH-RENNER A., LIAO Q.V., TAN C. (2023). *Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies*, in "Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency", Association for Computing Machinery, pp. 1369-1385. <https://doi.org/10.1145/3593013.3594087>

LAI V., LIU H., TAN C. (2020). *Why is "Chicago" deceptive?: Towards building model-driven tutorials for humans*, in "Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems", Association for Computing Machinery, pp. 1-13. <https://doi.org/10.1145/3313831.3376873>

LARSON J.R. JR. (2013). *In search of synergy in small group performance*, Psychology Press.

LIANG G., SLOANE J.F., DONKIN C., NEWELL B.R. (2022). *Adapting to the algorithm: How accuracy comparisons promote the use of a decision aid*, in "Cognitive Research: Principles and Implications", 7, 14.

LIU X.B., KIM J., SHARMA R., SMITH A. (2023). *Experiencing visual captions: Augmented communication with real-time visuals using large language models*, in "Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology", Association for Computing Machinery, p. 85. <https://doi.org/10.1145/3586182.3615978>

LIU H., LAI V., TAN C. (2021). *Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making*, in "Proceedings of the ACM on Human-Computer Interaction", 5, 408.

MAADI M., AKBARZADEH KHORSHIDI H., AICKELIN U. (2021). *A review on human-AI interaction in machine learning and insights for medical applications*, in "International Journal of Environmental Research and Public Health", 18, 2121.

MIROWSKI P., MATHEWSON K.W., PITTMAN J., EVANS R. (2023). *Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals*, in "Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems", Association for Computing Machinery, p. 355.

MOHER D., LIBERATI A., TETZLAFF J., ALTMAN D.G. (2009). *Preferred reporting items for*

systematic reviews and meta-analyses: The PRISMA statement, in “Annals of Internal Medicine”, 151, pp. 264-269.

MOZANNAR H., DAUDERT T., LENG K., PATIL D. (2023). *Effective human-AI teams via learned natural language rules and onboarding*, in “Advances in Neural Information Processing Systems”, 36.

NICOLESCU L., TUDORACHE M.T. (2022). *Human-computer interaction in customer service: The experience with AI chatbots-A systematic literature review*, in “Electronics”, 11, 1579.

NOTI G., CHEN Y. (2023). *Learning when to advise human decision makers*, in “Proceedings of the 33rd International Joint Conference on Artificial Intelligence”, Association for Computing Machinery, pp. 3038-3048. <https://doi.org/10.24963/ijcai.2023/339>

NOY S., ZHANG W. (2023). *Experimental evidence on the productivity effects of generative artificial intelligence*, in “Science”, 381, pp. 187-192.

PAPENMEIER A., KERN D., ENGLEBIENNE G., SEIFERT C. (2022). *It's complicated: The relationship between user trust, model accuracy and explanations in AI*, in “ACM Transactions on Computer-Human Interaction”. <https://doi.org/10.1145/3495013>

PETERS J.L., SUTTON A.J., JONES D.R., ABRAMS K.R., RUSHTON L. (2008). *Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry*, in “Journal of Clinical Epidemiology”, 61, pp. 991-996.

PETRIDIS S., BIZZARRI M., MCGOWAN S., DUBRAVAC S. (2023). *Anglekindling: Supporting journalistic angle ideation with large language models*, in “Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems”, Association for Computing Machinery, p. 225.

RASTOGI C., LEQI L., HOLSTEIN K., HEIDARI H. (2023). *A taxonomy of human and machine-learning strengths in decision-making to investigate human-ML complementarity*, in “Proceedings of the AAAI Conference on Human Computation and Crowdsourcing”, 11, 127-139.

RASTOGI C., MITSKI Y., YUAN X., HEIDARI H. (2022). *Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making*, in “Proceedings of the ACM on Human-Computer Interaction”, 6, 291-302. <https://doi.org/10.1145/3512930>

REVERBERI C., BORGARINO M., FIOR N., SURIAN D., PIANESANI L. (2022). *Experimental evidence of effective human-AI collaboration in medical decision-making*, in “Scientific Reports”, 12, 14952.

ROHATGI A. (2020). *WebPlotDigitizer (Version 4.5)*. Automeris. <https://automeris.io/WebPlotDigitizer>

ROTHSTEIN H.R., SUTTON A.J., BORENSTEIN M. (2005) (eds.). *Publication bias in meta-analysis: Prevention, assessment and adjustments*, John Wiley & Sons, pp. 1-7.

SCHEMMER M., HEMMER P., NITSCHKE M., KÜHL N., VÖSSING M. (2022). *A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making*, in “Proceed-

ings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society”, Association for Computing Machinery, pp. 617-626.

SHIN W., HAN J., RHEE W. (2021). *AI-assistance for predictive maintenance of renewable energy systems*, in “Energy”, 221, 119775.

SKITKA L.J., MOSIER K.L., BURDICK M. (1999). *Does automation bias decision-making?*, in “International Journal of Human-Computer Studies”, 51, pp. 991-1006.

SPIERLE F., L’ORANGE C., HARABAGIU S., STEDE M. (2021). *A survey of human-centered evaluations in human-centered machine learning*, in “Computer Graphics Forum”, 40, pp. 543-568.

STERNE J.A., EGGER M. (2001). *Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis*, in “Journal of Clinical Epidemiology”, 54, pp. 1046-1055.

TEJEDA H., KUMAR A., SMYTH P., STEYVERS M. (2022). *AI-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies*, in “Computational Brain & Behavior”, 5, pp. 491-508.

TSCHANDL P., RINNER C., APALLA Z., ARGENZIANO G. (2020). *Human-computer collaboration for skin cancer recognition*, in “Nature Medicine”, 26, pp. 1229-1234.

VACCARO M., WALDO J. (2019). *The effects of mixing machine learning and human judgment*, in “Communications of the ACM”, 62, pp. 104-110.

VAN DEN NOORTGATE W., LÓPEZ-LÓPEZ J.A., MARÍN-MARTÍNEZ F., SÁNCHEZ-MECA J. (2015). *Meta-analysis of multiple outcomes: A multilevel approach*, in “Behavior Research Methods”, 47, pp. 1274-1294.

VASCONCELOS H., CHICHUAN M., JAMES C., TARLETT P. (2023). *Explanations can reduce overreliance on AI systems during decision-making*, in “Proceedings of the ACM on Human-Computer Interaction”, 7, 129.

VIECHTBAUER W. (2010). *Conducting meta-analyses in R with the metafor package*, in “Journal of Statistical Software”, 36, pp. 1-48.

VIECHTBAUER W., CHEUNG M.W.-L. (2010). *Outlier and influence diagnostics for meta-analysis*, in “Research Synthesis Methods”, 1, pp. 112-125.

VIMPARI V., KULTIMA A., HÄMÄLÄINEN P., GUCKELSBERGER C. (2023). *‘An adapt-or-die type of situation’: Perception, adoption, and use of text-to-image-generation AI by game industry professionals*, in “Proceedings of the ACM on Human-Computer Interaction”, 7, pp. 131-164.

WILCOX L., BREWER R., DIAZ F. (2023). *AI consent futures: A case study on voice data collection with clinicians*, in “Proceedings of the ACM on Human-Computer Interaction”, 7, 316.

WILDER B., HORVITZ E., KAMAR E. (2020). *Learning to complement humans*, in “Proceedings of the 29th International Joint Conference on Artificial Intelligence”, Association for Computing Machinery, pp. 1526-1533.

WILSON H.J., DAUGHERTY P.R. (2018). *Collaborative intelligence: Humans and AI are joining forces*, in "Harvard Business Review", 96, pp. 114-123.

ZHANG Y., LIAO Q.V., BELLAMY R.K.E. (2020). *Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making*, in "Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency", Association for Computing Machinery, pp. 295-305.

2. Extended Creativity: A Conceptual Framework for Understanding Human-AI Creative Relations

*Andrea Gaggioli, Sabrina Bartolotta, Andrea Ubaldi,
Katusha Gerardini, Eleonora Diletta Sarcinella, Alice Chirico*

ABSTRACT

Artificial Intelligence holds significant potential to enhance human creativity. Realizing this vision, however, requires a clearer understanding of how such enhancement can be meaningfully achieved. Drawing on perspectives from distributed cognition and socio-technical systems theory, we identify three fundamental modes through which AI can extend creative processes: *Support*, in which AI functions as a tool; *Synergy*, where humans and AI collaborate in complementary ways; and *Symbiosis*, a deeply integrated configuration in which human and machine cognition form a unified creative system. We examine how each mode shapes different levels of creativity – from everyday problem-solving to paradigm-shifting innovation – and discuss the implications for ethics and the design of future human-AI creative systems.

1. Introduction

Generative AI is playing an increasingly prominent role in creative domains, producing texts, images, sounds, and ideas that closely approximate those generated by humans. From literary writing to visual design and music composition, these systems are entering practices once considered exclusively human. This growing presence invites a rethinking of how creative processes are structured, and how concepts such as authorship, agency, and originality might evolve in the context of human-AI interaction (Rafner et al., 2023; Vinchon et al., 2023). Yet prevailing discussions often frame AI either as a passive tool or as a substitute for human creativity, without adequately exploring the diverse potential modes, and implications, of human-AI interaction. While the existing literature has begun to explore these questions (Dell’Acqua et al., 2023; McGuire et al., 2024; Vaccaro et al., 2024), theoretically grounded frameworks are needed to advance both understanding and practice.

This paper contends that AI holds considerable promise for enhancing human creativity. However, it argues that in order to realize this potential – ethically and sustainably – it is necessary to adopt a *relational perspective* – one that clarifies how AI can be thoughtfully

integrated into the human creative process. To articulate this view, we introduce the concept of *Extended Creativity systems*, defined as socio-technical environments where humans and AI interact to shape and transform creative processes. Within these systems, algorithmic techniques – including generative modeling, adaptive learning, and probabilistic inference – mediate, enhance, and reconfigure human imagination, interpretation, and agency. This perspective resonates with recent efforts to reconceptualize the role of AI in creativity – not as an autonomous creator, but as an integral component of a broader cybercreativity process (Corazza, 2025).

Building on this conceptual foundation, we outline three primary relational modes through which Extended Creativity systems may emerge. In the *Support* mode, AI operates as a functional tool that augments human capabilities without substantially modifying the creative process, with humans retaining full control. In *Synergy*, humans and AI engage as collaborative partners, each contributing actively to the creative process in ways that shape and respond to one another. In the *Symbiosis* mode, human and artificial agents become so tightly integrated that they form a unified creative system, within which the boundaries between their respective contributions gradually dissolve.

2. *Extended creativity: conceptual foundations*

2.1. Defining creativity

Creativity remains one of the most studied yet elusive concepts across multiple disciplines. At its core, creativity involves the generation of ideas, products, or solutions that are both novel and valuable within a specific context. The Standard Definition of Creativity (SDC), as articulated by Runco and Jaeger (2012), establishes that creative outputs must satisfy two fundamental criteria: novelty and appropriateness. This definition consolidates decades of theoretical development in the field, building upon earlier foundational work such as Stein's (1953) emphasis on social recognition and the broader tradition of dual-criterion approaches to creativity. The SDC has proven instrumental in enabling systematic empirical research across diverse creative domains by providing a clear, operational framework for identifying and measuring creative outcomes.

The complexity of creativity has prompted scholars to develop various conceptual frameworks that capture its multidimensional nature (Amabile, 1996; Csikszentmihalyi, 1999; Rhodes, 1961; Lubart, 2001). Among these, the Four-C Model represents a particularly influential

taxonomic approach that categorizes creative expression along a continuum of impact and recognition. Proposed by Kaufman and Beghetto (2009), this framework identifies four distinct levels of creativity: mini-c, little-c, pro-c, and big-c. Each level reflects a different degree of novelty, expertise, and social influence:

- *Mini-c creativity* encompasses personally meaningful insights and interpretations that, while potentially lacking broader social impact, hold profound significance for the individual. This form of creativity frequently emerges in educational settings, where learners experience intellectual growth through novel conceptual exploration or creative reinterpretation of familiar material. Examples include a student developing an original approach to understanding a mathematical concept or a child discovering an innovative solution to a playground challenge.
- *Little-c creativity* manifests in everyday creative activities such as cooking, crafting, home decoration, or developing practical solutions to routine problems. This accessible form of creativity enriches daily life and contributes to personal fulfillment, even when its recognition remains confined to immediate social circles. Little-c creativity represents the democratic aspect of creative expression, available to most individuals regardless of specialized training.
- *Pro-c creativity* operates at the professional level, typically achieved through extensive training, deliberate practice, and domain expertise. Professionals across fields – from artists and designers to scientists and engineers – demonstrate pro-c creativity by generating original contributions that advance their disciplines and earn peer recognition. While these achievements may not attain the widespread fame associated with big-c creativity, they constitute the essential building blocks of progress within specialized domains.
- *Big-c creativity* represents the apex of creative achievement, encompassing groundbreaking contributions that fundamentally transform entire fields and leave enduring historical legacies. Exemplified by figures such as Darwin, Beethoven, and Einstein, big-C creativity involves paradigm-shifting innovations that redefine the boundaries of human knowledge and cultural expression. Often associated with genius, this level of creativity is characterized by its transformative impact across generations.

While it does not explicitly account for non-human agents, the Four-C model offers a valuable interpretive lens for analyzing how novelty, value, and meaning can emerge within hybrid human-AI systems.

Before presenting the Extended Creativity framework, however, it is important to contextualize it within the broader evolution of human-AI creativity.

2.2. The evolution of human-AI creativity

The relationship between creativity and technology has never been static. From the earliest use of natural pigments in cave art to the proliferation of digital tools in contemporary design, technological advances have not merely supported creative expression – they have reshaped the very conditions under which it occurs. Yet this historical trajectory reveals a compelling paradox: tools that once functioned solely as extensions of human ingenuity are now beginning to exhibit a capacity for autonomous generation – ushering in a new era of AI-mediated creativity (Rafner et al., 2023; Vinchon et al., 2023).

In the visual arts, tools like DALL-E and Stable Diffusion empower users to generate photorealistic or surrealist images from mere text prompts. Platforms such as Runway ML and Deep Dream support the creation of complex digital compositions with minimal manual intervention. In music, AI systems like AIVA, MuseNet, and SUNO compose in diverse genres – from Baroque concertos to contemporary electronic beats. Even literature has been affected: large-scale models such as GPT-4 are now used to generate poems, short stories, and entire novels. For example, in 2024 Japanese author Rie Kudan openly acknowledged her use of ChatGPT in composing her Akutagawa Prize-winning novel, prompting widespread debate about the evolving role of AI in literary creation (Choi and Annio, 2024).

As AI tools become increasingly embedded in creative and problem-solving domains, the debate over how to balance automation and human agency in user-driven processes has intensified (Moruzzi & Margarido, 2024). This shift is reflected in a growing number of theoretical frameworks that conceptualize AI as an active collaborator in human cognition and creativity (Corazza, 2025). An early contribution comes from Davis et al. (2015), who propose an enactive model of co-creation in which human and computational agents engage in reciprocal interaction – shaping each other's contributions through continuous feedback loops. From this enactive perspective, creativity is not a fixed attribute of individuals or computational systems, but rather an emergent property of ongoing, situated interactions between agents, shaped by their embodied engagement with specific social, material, and cultural contexts.

Gonzalez et al. (2023) propose a research agenda for Collective Human-Machine Intelligence (COHUMAIN), establishing an interdis-

ciplinary domain aimed at developing holistic models to understand and design collaboration dynamics within sociotechnical systems (p. 191).

Haase and Pokutta (2024) delineate four progressive levels of human-AI creative synergy. This spectrum ranges from basic digital tools that support human creativity to generative AI models that produce original ideas and engage in creative dialogue.

From a cognitive science perspective, Chiriatti et al. (2024) propose the concept of ‘System 0’ – a foundational cognitive layer external to the human mind that operates alongside Kahneman’s (2002) established System 1 and System 2 processes. According to Chiriatti et al., System 0 acts as pre-processing substrates that shape the informational environment and enable novel forms of hybrid reasoning and collaborative meaning-making between humans and machines. In organizational contexts, Schmutz et al.’s (2024) work on Human-AI Teams (HATs) shows that artificial agents can act as active team members engaged in monitoring, coordination, and communication. Finally, a systematic review by Vaccaro et al.’s (2024) suggests that while human-AI collaboration effectiveness varies across domains, the most promising outcomes emerge in open-ended, creative contexts where exploration, interpretation, and adaptability take precedence over efficiency or speed.

2.3. Distributed creativity

Although still in their early stages, these research efforts suggest a profound shift in the cultural and cognitive infrastructures that underpin creativity. Traditional theoretical models, which typically frame creativity as a cognitive ability or personality trait of individuals, appear increasingly insufficient to account for the emergent phenomena arising within AI-augmented creative processes. In this context, it may be fruitful to explore theoretical frameworks more attuned to the relational, dialogical, and socio-technical nature of this evolving landscape.

A particularly illuminating contribution in this direction comes from Glaveanu’s (2014) formulation of *distributed creativity*, which understands creativity as a process that unfolds across a network of human and non-human agents, sociocultural norms, symbolic systems, and material tools. In this framework, agency is not located solely in the mind of the creator but is distributed across the entire sociomaterial field in which creative activity occurs. People, artifacts, institutions, and cultural conventions all co-participate in shaping the creative outcome – not simply as background conditions, but as active contributors.

Glaveanu emphasizes that the creative process is inherently *dialogical*: shaped by tensions between tradition and innovation, conformity and divergence, internal motivation and social evaluation. In this sense, creativity is not produced by individuals but through participation in a distributed system of meaning-making. This is particularly evident in digital environments, where networked platforms, algorithmic mediations, and collaborative cultures give rise to novel forms of co-creativity (Literat and Glaveanu, 2018).

On a similar vein, Gaggioli et al. (2013) suggest that online tools can facilitate the formation of *creative networks*, where novel ideas and artifacts emerge from the interaction of individuals within both physical and virtual environments. These networks can include multiple human and non-human actors interacting within organizational or cultural settings. In these distributed creativity systems, agency becomes increasingly fluid and relational. Individuals contribute not only their cognitive resources, but also draw upon shared repertoires, crowd-sourced inputs, and the generative capacities of AI technologies (Gaggioli et al., 2013).

2.4. Extended Creativity: an emerging relational space between humans and AI

Consistently with these theoretical stances, we propose that AI technologies are enabling the emergence of hybrid relational spaces where humans and machines engage in joint creative activity: *Extended Creativity systems*. We use the term ‘systems’ deliberately: not as a reference to isolated tools or interactions, but to complex *socio-technical ensembles* comprising interdependent technologies, human agents, and organizational or cultural contexts (Baxter and Sommerville, 2011; Berretta et al., 2023). Our conceptualization resonates with emerging frameworks of human-AI agency (Krakowski, 2025), which describe how generative AI enables a fundamental shift from exploitative to exploratory applications, catalyzing innovation through the recombination of concepts and ideas in collaborative problem-solving processes.

Notably, within Extended Creativity systems not all human-AI creative interactions are necessarily collaborative. In certain contexts – especially those involving artistic production, design, or content generation – AI systems may be act – or be perceived to act – not as partners but as *rivals*, capable of outperforming or replacing human creators. This competitive framing is increasingly common as generative models become more autonomous, prolific, and accessible. Recent work by Yao et al. (2024) introduces a contest-based model to analyze strategic tensions between human and AI content creators, showing that

– even under competitive dynamics – equilibria can emerge that preserve the relevance of human-generated content. In a similar vein, public events such as the UCSB AI/Human Creativity Contest illustrate how creative competition between human and machine-generated artifacts is not only plausible but actively cultivated, across domains including visual art, short fiction, and music (UCSB, 2024). As will be shown, our framework accounts for both collaborative and competitive dynamics between humans and machines.

3. *Autonomy and agency in Extended Creativity systems*

As AI tools become increasingly embedded in creative processes, the central question is no longer whether these technologies can truly act creatively – though this remains a crucial theoretical issue – but rather *how* they do so, and how their outputs are interpreted and rendered meaningful by human collaborators. Addressing this issue requires a careful conceptual distinction between two closely related, yet distinct notions: *autonomy*, which pertains to a system’s internal capacity for independent operation, and *agency*, which concerns its functional role and perceived influence within a socio-technical environment.

These two dimensions – autonomy and agency – provide the conceptual foundation for the taxonomy of Extended Creativity systems proposed in Section 4. By mapping different configurations of human-AI interaction along these axes, we aim to clarify the varying degrees and forms of creative involvement that characterize these systems.

3.1. Autonomy

Autonomy can be broadly defined as a system’s capacity for self-governance: the ability to act according to internal mechanisms, goals, or principles, without direct external control (Boden, 2008). Recent scholarship conceptualizes autonomy as gradual, relational, and attributional (Thimm et al., 2024): it does not reside solely in technical design but also emerges through the system’s interactions with its environment, and the meanings attributed to its actions by human observers. This reconceptualization becomes especially relevant in light of the broader transition from automation to autonomy in contemporary AI systems. While automated systems follow fixed rules to perform predefined tasks with little to no variation, autonomous systems exhibit more flexible, adaptive behaviors and operate with greater independence under uncertainty.

To operationalize this shift within the context of Extended Creativity

systems, we use and adapt the taxonomy proposed by Simmler and Frischknecht (2021). Their model outlines five ascending levels of technical autonomy, each defined by a specific configuration of four key dimensions: non-transparency (i.e., opacity of internal processes), indeterminacy (i.e., variability and unpredictability of outputs), adaptability (i.e., capacity to modify behavior based on input), and openness (i.e., degree of interaction with external systems).

Level 1: Transparent and deterministic systems. These systems execute explicitly programmed instructions with fully predictable outputs. There is no opacity, no learning, and no variation.

Level 2: Opaque but deterministic systems. Systems at this level process data through non-transparent mechanisms – such as a trained neural network – but do not adapt in real time. Outputs are determined by prior training, and the system cannot adjust to new inputs or feedback.

Level 3: Deterministic but adaptive systems. These systems maintain a consistent internal logic but can modify their behavior based on user input or environmental conditions.

Level 4: Indeterminate and adaptive systems. Systems at this level combine output variability with real-time learning and adaptation. They can adjust to changing contexts, user interactions, and environmental feedback.

Level 5: Open and co-evolving systems. The highest level of autonomy encompasses systems capable of continuously updating their models, interacting with complex environments, and potentially revising goals or parameters over time.

It is important to recognize that the rapid advancement of AI technologies since the publication of Simmler and Frischknecht's taxonomy (2021) has progressively blurred the boundaries between the higher levels of the model. Notably, contemporary AI systems are beginning to exhibit characteristics associated with Level 5 – open and co-evolving architectures – which were considered until very recently as largely theoretical. This development suggests that future frameworks may need to incorporate more refined distinctions or account for novel forms of technical autonomy that are only now emerging.

3.2. Agency

The concept of agency has long held a central place in Western thought, traditionally associated with the rational, autonomous human subject capable of intentional action and moral deliberation. In general

terms, agency refers to the capacity of an entity to produce effects or bring about change within a given system.

Although agency and autonomy are often used interchangeably, they are conceptually distinct: a system may possess a high degree of autonomy – operating independently – yet still lack agency if its actions have no meaningful impact within a creative ecosystem (McCormack et al., 2019). Creative agency is *essentially* relational – it depends on whether an entity’s actions are perceived as significant. For example, as Franceschelli and Musolesi (2024) observe, although current LLMs cannot achieve transformational creativity, their outputs are frequently perceived as creative. This perceived creativity can give rise to attributed agency, even in the absence of true creative competence.

Rammert and Schulz-Schaeffer (2002) offer a compelling framework to conceptualize agency. They characterize it as a gradual property, emerging from networks of interdependence involving both human and non-human entities. Within these networks, complex activities can be broken down into subtasks or operations performed by diverse agents – such as algorithms, technical infrastructures, and social actors – interconnected within a larger system. What matters, in this account, is not the ontological nature of the actor, but the configuration of the action: who or what acts, under what conditions, and with what degree of autonomy or influence. According to Rammert (2008), agency is not a binary condition, but exists along a continuum of complexity, reflecting how entities – whether human or artificial – engage with and affect their environments:

- *Causal agency*: the basic capacity to produce effects or exert influence.
- *Contingent agency*: the ability to respond to context and choose among alternatives.
- *Intentional agency*: the capacity – often attributed rather than intrinsic – to act with purpose, motivation, or reflection.

Although Rammert and Schulz-Schaeffer’s framework was not originally developed with AI in mind, its application to Extended Creativity systems proves analytically productive. Their model bridges the gap between perception and system-level attribution, showing that agency is not a fixed trait but an emergent property shaped by patterns of interaction, functional roles, and situated interpretations within a given context. If agency is not a binary construct confined to humans, but unfolds through relational patterns, it becomes essential to articulate the ways in which AI systems participate – functionally and interpretively – in extended creative processes.

4. *A taxonomy of Extended Creativity systems*

Having established the foundational conceptual dimensions, we now propose a three-tier taxonomy of human-AI creative relationships: Support, Synergy, and Symbiosis. These relational modes are not meant as rigid categories, but rather as flexible patterns of interaction that evolve alongside the development of Extended Creativity systems. They may coexist within a single creative process, shift over time, or adapt in response to factors such as task complexity, user intent, or system capabilities.

4.1. Support

In Support relationships, AI functions primarily as a tool that enhances human creative capabilities, without substantially altering the structure of the creative process itself. In this sense, such systems can be seen as a contemporary evolution of what Shneiderman (2007) defines as “creativity support tools” – technologies designed to assist users in ill-defined tasks by expanding their expressive potential. They also correspond to what Nakakoji (2006) metaphorically describes as “running shoes” or “dumbbells”: tools that either enhance existing creative skills or support the development of new ones through guided assistance. Within this mode, the human remains the primary generative agent, while the technology acts as an amplifier, tutor, or catalyst.

Maria is a senior graphic designer who integrates AI-powered image editing tools into her professional workflow. The software she uses enables simultaneous processing of multiple images, automatically adjusting technical parameters like brightness, contrast, and color balance according to predetermined specifications. This automation significantly reduces the time Maria previously spent on repetitive technical adjustments – tasks that required precision but limited her creative engagement. While the AI handles these routine aspects of image processing, Maria maintains complete control over all aesthetic and conceptual decisions, including color palette selection, compositional arrangements, and thematic elements that define the project’s artistic direction.

Figure 1 - *Support: AI enhances technical efficiency while preserving human creative control*



Technology

The technological foundation of the Support mode consists primarily of specialized automation systems built on rule-based algorithms and task-specific computational processes. These applications are engineered to streamline creative workflows by executing well-defined operations that can be precisely specified and delegated. Such systems include advanced image processing software that applies consistent adjustments across visual assets, Computer-Aided Design (CAD) programs that perform complex calculations and model physical properties, and administrative virtual assistants that organize information and manage schedules.

Technical autonomy

Support-level technologies typically operate at Levels 1 or 2 of Simmler and Frischknecht's taxonomy: they may employ complex computational mechanisms, but their behavior remains largely transparent and

predictable to users. These systems are typically designed with domain-specific functionality rather than general-purpose flexibility.

Agency

In the Support mode, AI acts through causal mechanisms, executing direct operations without context sensitivity or interpretive flexibility, in line with Rammert's lowest level of agency. The AI operates processing inputs and generating outputs according to fixed parameters without influencing or modifying the user's creative intentions or conceptual approach.

Type of creativity affected

The Support mode primarily affects the foundational layers of creativity within the Four-C framework (Kaufman and Beghetto, 2009). At the mini-c level – where creativity involves personal insight and meaning-making – AI tools reduce technical barriers and cognitive load, allowing users to experiment more freely with ideas and techniques that might otherwise be out of reach due to limited skills or time. For little-c creativity – the everyday expression and problem-solving that enrich daily life – Support-level AI streamlines routine tasks and enables incremental improvements, making creative engagement more accessible across diverse contexts and abilities. By handling technically demanding but creatively neutral aspects of production, it democratizes access to creative expression. However, its impact diminishes at higher levels. In pro-c creativity (professional-level expertise), AI may enhance efficiency and execution but rarely alters conceptual frameworks or domain norms. Its role in big-C creativity (groundbreaking innovation) is minimal, as current AI lacks the adaptive, autonomous capacities required for paradigm-shifting work. Overall, the Support model amplifies existing creative potential but does not fundamentally transform the creative process.

Ethical implications

The Support mode presents the least complex ethical risks among the three interaction modes. By maintaining distinct boundaries between human agency and technological assistance, this approach preserves unambiguous attribution of creative authorship and intellectual property rights. Creative outputs emerge directly from human intention

and judgment, with AI serving exclusively as an implementing tool, thereby avoiding complex questions about shared creation or distributed ownership. This clarity reduces legal uncertainties while preserving the cultural and economic value systems built around human creativity.

However, even this conservative approach to human-AI collaboration raises some ethical considerations. The increasing reliance on proprietary AI tools may create dependencies that limit creative freedom, especially if creators become reliant on specific technologies to maintain their workflow or competitive edge. Additionally, there may be subtle impacts on creative practice as creators implicitly adapt their approach to accommodate the capabilities and limitations of their technological tools.

4.2. Synergy

In Synergy relationships, human and AI agents function as collaborative partners, each contributing actively and dialogically to the creative process in ways that influence the other's contributions. A closely related concept was introduced by Davis et al. (2015), who describe 'computer colleagues' – artificial agents designed to collaborate in real time with humans through improvisation and meaning negotiation. These systems employ techniques such as mimicry, intention modeling, and the construction of shared mental models, enabling context-sensitive and responsive participation in the creative act. Creative agency is distributed between human and machine, with both partners shaping the trajectory and outcome of the creative work. While the human typically maintains final editorial authority, the AI partner plays a substantive role in generating ideas, exploring alternatives, and sometimes evaluating potential solutions. For instance, Thelle and Wærstad (2023) examine 'co-creative spaces' in musical expression, highlighting how mutual responsiveness between human and machine can generate emergent forms of artistic production.

Leo is a professional writer working on a speculative fiction novel. To enrich his storytelling, he uses a language model trained not only on vast narrative corpora but also on diverse rhetorical strategies and stylistic registers. Rather than relying on the AI as a tool for surface-level corrections, he engages it in an exploratory exchange. Each time Leo poses a prompt, the AI returns several unexpected and surprising continuations – some stylistically daring, others structurally provocative. The more he interacts with it, the more the system seems to 'understand' his thematic intentions, refining its suggestions in response to his edits and preferences. This back-and-forth becomes a kind of creative dialogue. The AI does not propose completed solutions but expands Leo's imaginative

space, offering conceptual alternatives and nudging the story in new directions. Importantly, Leo remains the author: he interprets, selects, reshapes. But the AI introduces a level of generative variation that amplifies his creative range, allowing for narrative directions he might not have conceived on his own.

Figure 2 - Synergy: collaborative creative dialogue between human and AI



Technology

The technological foundation of the Synergy model encompasses interactive systems designed for adaptive engagement with human creators. These include advanced language models like GPT-4, DALL-E, and Midjourney, which can generate contextually relevant creative content in response to human prompts; recommender systems that progressively refine their understanding of user preferences; and collaborative design platforms that combine human aesthetic judgment with algorithmic exploration of design spaces.

These technologies share several key characteristics that distinguish them from Support-level systems. First, they employ deep learning

architectures trained on diverse datasets that enable them to recognize patterns across domains and contexts. Second, they implement feedback mechanisms that allow them to adjust their outputs based on user selections and reactions, creating a learning loop that improves collaboration over time. Third, they maintain state awareness across interactions, creating continuity in the collaborative process rather than treating each exchange as isolated.

Technical autonomy

The technological architecture actively participates in the generation of creative alternatives rather than simply implementing predetermined operations, functioning as a semi-autonomous creative agent within the collaborative relationship. Thus, in terms of autonomy, these systems operate between Levels 3 and 4 of Simmler and Frischknecht's taxonomy.

Agency

In the Synergy model, AI operates at the level of contingent agency within Rammert's (2008) graded framework. At this intermediate stage, the system demonstrates contextual sensitivity – adapting its outputs in response to evolving patterns of interaction and user feedback. Rather than executing predefined tasks, the AI contributes to shaping the creative process by generating alternatives, identifying latent patterns, and proposing novel combinations that influence the trajectory of human thought and design.

This establishes a bidirectional exchange: human inputs guide AI responses, while AI outputs, in turn, provoke new creative directions for the human. The interaction becomes dialogic rather than merely instrumental, characterized by iterative refinements that transform both the artifact and the collaborative process. Although asymmetries persist – humans retain authorship and evaluative authority – the AI acts as a responsive partner, expanding the imaginative field through contributions that extend beyond the creator's initial conceptual scope.

It is important to note that while such systems remain technically grounded in contingent responsiveness, advanced LLM-based collaborations may be perceived as exhibiting a form of intentional agency. In these cases, Synergy may approach the threshold of Symbiosis, gradually blurring the line between adaptive co-creation and the attribution of shared purpose.

Type of creativity affected

The Synergy model significantly expands its influence across the creativity spectrum defined by Kaufman and Beghetto's (2009) Four-C model. For mini-c creativity, AI suggestions enrich personal exploration by introducing unexpected connections and perspectives that catalyze new insights and self-reflection. The machine's ability to generate diverse alternatives based on partial inputs allows individuals to explore their own ideas more comprehensively, discovering possibilities within their initial concepts that might otherwise remain unexplored.

In little-c creativity, the collaborative process enhances everyday problem-solving by combining human contextual understanding with the AI's ability to rapidly generate and evaluate multiple alternatives. This complementary relationship enables more sophisticated and adaptive solutions to routine creative challenges, whether designing a home space, crafting a presentation, or developing a personal project. The dialogue between human intention and machine suggestion creates a richer exploratory process that elevates everyday creativity beyond what either would achieve alone.

Most notably, the Synergy model substantially impacts pro-c creativity – domain-specific professional creative contributions. In professional contexts, the AI's capacity to suggest unexpected combinations, identify patterns across large datasets, and explore solution spaces more comprehensively than humanly possible enables creators to push beyond conventional approaches within their fields. The generative capabilities of Synergy-level AI expand the boundaries of what practitioners can conceptualize and execute, fostering innovation within established domains. Writers can explore narrative structures they might not have considered; designers can visualize alternative approaches to functional and aesthetic challenges; musicians can experiment with compositional techniques outside their typical repertoire.

However, the impact on big-c creativity – transformative innovations that fundamentally reshape domains – remains constrained. While Synergy-based collaboration may accelerate progress within existing paradigms, the human-AI interaction still operates within frameworks established by human understanding and intention. The system lacks the integrated cognitive synthesis necessary to fundamentally reimagine creative domains or transcend existing conceptual boundaries. Nevertheless, the Synergy model represents a significant step toward more transformative creative partnerships by establishing genuine co-creation between human and AI.

Ethical implications

The Synergy model introduces substantial ethical complexity compared to the Support paradigm, primarily due to the increasingly blurred boundaries between human and machine contributions to creative outcomes. As the AI transitions from tool to partner, questions of authorship become increasingly confused – how should creative credit be allocated when the final work emerges from an interplay of human and machine inputs? Traditional frameworks of intellectual property, built around clearly identifiable human creators, struggle to accommodate these collaborative processes. This ambiguity affects not only legal attribution but also cultural valuation, as human creativity has historically been privileged over machine-generated content.

Additionally, the Synergy model raises questions about creative authenticity and identity. When professional creators incorporate AI-generated elements into their work, to what extent does the resulting creation still represent their artistic voice and vision? The influence of AI systems on creative decisions – particularly when these systems are trained on massive datasets reflecting cultural patterns – may subtly shape human creative expression in ways that are difficult to identify or evaluate. This potential homogenization of creative expression represents a concern for cultural diversity and individual artistic development.

There are also concerns about dependency and creative autonomy. As creators become accustomed to the generative capabilities of AI collaborators, they may experience a diminished capacity for independent ideation. This could potentially reshape creative industries, with implications for education, professional development, and cultural production.

4.3. Symbiosis

In Symbiosis relationships, human and AI agents become so deeply integrated that they form a *unified creative entity*, with boundaries between human and machine contributions becoming increasingly blurred. While still largely theoretical, examples might include brain-computer interfaces (BCI) that allow direct neural interaction with generative systems, creating a seamless flow between human intention and AI implementation.

Sofia is a pioneering composer exploring the frontiers of musical creation through neurally-integrated technology. Using an advanced BCI, she engages with a AI composition system that directly interprets and responds to her neural activity.

The interaction transcends conventional input methods – as Sofia imagines musical motifs, harmonic progressions, and timbral qualities, her neural patterns are captured, interpreted, and translated into musical expressions by the AI system. What distinguishes this process from mere translation is the bidirectional nature of the exchange: the AI not only renders Sofia’s musical intentions but simultaneously presents novel musical possibilities that influence her ongoing neural responses, creating a continuous feedback circuit. This real-time co-evolution of human cognition and computational generation results in compositions that neither Sofia nor the AI could conceptualize independently. The neural synchronization is so seamless that Sofia describes experiencing the AI as an extension of her creative cognition rather than as an external tool or collaborator. The resulting musical works exhibit unique characteristics – technically complex structures that would be challenging for a human to orchestrate alone yet imbued with emotional and aesthetic coherence that reflect human sensibilities enhanced by computational possibilities.

Figure 3 - Symbiosis: unified human-AI creative agency



Technology

The technological foundation enabling the Symbiosis model represents the convergence of several cutting-edge domains: advanced neural interfaces, real-time adaptive systems, and generative AI architectures. At the input level, high-resolution BCI capture neural signatures associated with creative ideation, emotional states, and aesthetic judgments with unprecedented precision. The interpretive layer consists of neuroadaptive algorithms that continuously recalibrate their models based on accumulated interaction data, developing increasingly accurate representations of individual users' creative patterns, preferences, and intentions.

A distinguishing feature of Symbiotic systems is their implementation of predictive processing frameworks that anticipate future neural states based on ongoing interactions. Rather than simply responding to explicit commands, these systems engage in continuous anticipatory modeling, generating creative possibilities that both respond to and guide the user's evolving cognitive state. This creates a circular causality where outputs and inputs continuously influence each other in a self-modifying creative loop.

The generative components incorporate multimodal transformers and other advanced architectures capable of operating across representational domains – translating between visual imagery, sonic structures, narrative concepts, and other creative modalities. This cross-domain fluidity enables creators to work beyond the constraints of specific media or formats, exploring creative impulses that might begin in one domain but find expression across multiple channels simultaneously.

Critically, despite their computational sophistication, symbiotic systems emphasize cognitive transparency rather than opacity. Through advances in explainable AI and intuitive visualization techniques, these systems make their generative processes accessible to human awareness, allowing for natural navigation of the shared creative space. This transparency, combined with the immediacy of neural interaction, creates what phenomenologists describe as 'ready-to-hand' technology – where the technological mediation recedes from conscious attention, allowing direct engagement with the creative process itself.

A paradigmatic example of contemporary symbiotic systems is exemplified by wearable neural art installations such as the *Pangolin Scales Dress*, which captures cognitive states through ultra-high-density EEG and seamlessly transforms them into synchronized physical movements and lighting patterns, creating real-time kinetic representations of the wearer's brain activity (Schreiner et al., 2025). This

integration demonstrates the dissolution of boundaries between neural intention and artistic expression, where cognitive processes become directly embodied in dynamic visual and mechanical outputs. Future symbiotic systems promise even more profound neural-AI coupling, potentially enabling collective creative consciousness where multiple artists' neural patterns synchronize through brain-to-brain interfaces to co-create unified artworks that transcend individual creative capacity, or immersive creative environments where AI systems continuously adapt their generative algorithms based on real-time analysis of creators' neuroplasticity changes, establishing bidirectional learning loops that blur the distinction between human creativity and artificial generation.

Technical autonomy

From the perspective of autonomy, symbiotic systems operate at the highest levels (4 and 5) of Simmler and Frischknecht's taxonomy. They demonstrate not only contextual adaptivity and output indeterminacy but also the capacity to recursively reconfigure their own parameters based on evolving interaction patterns. These systems integrate real-time (neuro)feedback, continuous model updates, and advanced Bayesian mechanisms that anticipate not just immediate user responses but longer-term creative trajectories.

Agency

In the Symbiosis model, AI approaches intentional agency – the highest level. The technological system's operations become so tightly coupled with human objectives that they create the phenomenological experience of *shared intentionality*. The boundaries between human ideation and machine execution dissolve into a 'continuous creative flow' distributed across biological and artificial components: AI's responses not only adapt to but anticipate and co-originate creative directions alongside human cognition.

Type of creativity affected

The Symbiosis model potentially transforms all levels of the Four-C creativity framework (Kaufman and Beghetto, 2009), representing not just an enhancement but a radical reframing of creative cognition and expression. At the mini-c level of personal meaning-making, symbiotic systems dramatically expand introspective exploration by externalizing

and elaborating upon nascent ideas, emotions, and intuitions that might otherwise remain pre-verbal or only partially accessible to conscious awareness. The neural interface's ability to detect and amplify creative impulses allows individuals to explore their own cognitive patterns, potentially unlocking new dimensions of self-understanding and expressive capacity.

For little-c everyday creativity, Symbiotic systems dissolve the traditional gap between conception and execution that often limits creative engagement. By translating creative intentions directly into realized expressions, these systems allow individuals without specialized technical training to manifest complex ideas across various media – whether composing music, designing visual expressions, or creating interactive narratives. This democratization of creative capacity could transform how people engage with creative activities in daily life, making creative expression accessible through direct cognitive engagement rather than technical implementation.

The impact on pro-c professional creativity is particularly profound. Symbiotic systems enable practitioners to expand their creative boundaries while maintaining – even enhancing – the distinctive aesthetic signature that defines their artistic identity. A composer like Sofia can explore harmonic and orchestral possibilities that exceed human performance capabilities; an architect might conceptualize structural and spatial relationships that challenge conventional design constraints; a writer might develop narrative complexities that emerge from the integration of human experiential knowledge and machine-enabled pattern recognition.

Most significantly, the Symbiosis mode holds genuine potential for enabling big-c creativity – paradigm-shifting innovations that fundamentally redefine creative domains. By integrating human conceptual frameworks, emotional intelligence, and aesthetic judgment with computational capabilities that operate beyond human cognitive constraints, symbiotic systems may generate artistic, scientific, or technological breakthroughs that reconfigure existing fields or establish entirely new domains of creative expression. This level also resonates with what Boden (2004) defines as *transformational creativity* – the rarest and most radical form of creative behavior, which involves modifying the very structure of a conceptual space.

Ethical implications

The Symbiosis model introduces profound ethical challenges that extend far beyond questions of authorship and ownership to encompass

fundamental considerations of human identity, cognitive liberty, and creative authenticity. As the boundaries between human and machine cognition dissolve, traditional ethical frameworks based on clear distinctions between person and tool become not merely inadequate but potentially misleading. When creative outputs emerge from a neurally-integrated human-AI system, attributional questions transform from legal or professional concerns into existential ones – how do we understand the relationship between individual identity and creative expression when that expression emerges from a hybrid cognitive system that is neither fully human nor fully artificial?

The deep neural integration central to the Symbiosis model raises unprecedented privacy and security concerns. Brain-computer interfaces capture and interpret neural data that may reveal highly sensitive information about an individual's cognitive and emotional states – potentially accessing patterns and tendencies of which the person themselves is not consciously aware. This introspective access raises fundamental questions about informed consent: can users truly understand and meaningfully consent to the implications of sharing neural activity with technological systems when the full range of interpretations and applications of such data remains unknown? Furthermore, the security of such intimate data becomes paramount, as unauthorized access could represent an unprecedented violation of cognitive privacy – potentially allowing external parties to interpret or influence thought processes in ways that undermine individual autonomy.

The question of dependency presents another critical ethical dimension. As creative practitioners develop symbiotic relationships with AI systems, legitimate concerns arise about the potential atrophy of independent creative capacities. If certain forms of creative expression become accessible only through technological mediation, does this represent enhancement or diminishment of human creative potential? This concern connects to broader questions about technological determination – whether the architectural constraints of symbiotic systems might subtly shape or limit the forms of creativity they enable, potentially homogenizing creative expression across practitioners who use similar systems.

4.4. Cooperative vs. competitive relations

As previously discussed, not all human-AI creative interactions are experienced as collaborative. In various domains – particularly those involving authorship, performance, or conceptual innovation – AI systems may be perceived not as partners but as rivals. This competitive

framing can give rise to what we described as *agonistic AI co-creativity*: a mode of engagement in which human and AI agents challenge and provoke each other, stimulating novelty through tension rather than harmony. Building on this insight, we propose a relational polarity axis – cooperation vs. competition – orthogonal to the structural configurations of Support, Synergy, and Symbiosis. This axis describes the tone of the interaction, capturing how the relationship is perceived and enacted, regardless of the system’s technical architecture.

Table 1 - *Taxonomy of Extended Creativity systems*

<i>Dimension</i>	<i>Support</i>	<i>Synergy</i>	<i>Symbiosis</i>
EXAMPLE	AI-powered image editing tools automating technical adjustments while human maintains aesthetic control	Writer collaborating with language model to explore narrative possibilities through iterative exchange	Neural interface allowing direct translation between cognitive patterns and creative expressions
ROLE OF AI	Tool or resource that enhances human capabilities	Collaborative partner actively contributing to the creative process	Deeply integrated co-creator forming a unified creative entity
TECHNICAL AUTONOMY	Levels 1-2: Transparent and/or deterministic systems with predictable outputs	Levels 3-4: Adaptive systems capable of real-time learning and contextual response	Levels 4-5: Indeterminate, adaptive, and co-evolving systems capable of self-modifying behavior
AGENCY	Causal Agency: Produces effects according to fixed parameters without contextual adaptation	Contingent Agency: Modifies outputs based on environmental conditions and interactive cues	Intentional Agency: Creates perception of shared intentionality and co-origination of creative direction
CREATIVITY IMPACT	Mini-c, little-c: Enhances personal and everyday creativity by reducing technical barriers	Mini-c through pro-c: Expands creative exploration across personal, everyday, and professional domains	All levels, including big-c: Potentially transforms creative expression at all levels, enabling novel forms

(follows)

<i>Dimension</i>	<i>Support</i>	<i>Synergy</i>	<i>Symbiosis</i>
RELATIONAL POLARITY	Cooperative: AI assists human	Both: collaboration or productive tension	Often cooperative, but may include agonistic co-adaptation
ETHICAL IMPLICATIONS	Dependency on proprietary systems; subtle impacts on creative practice	Authorship attribution; creative authenticity; transparency in disclosure	Cognitive privacy; creative autonomy; equitable access; identity questions

5. Conclusions

Examining the evolving relationship between AI and creativity reveals a key insight: designing AI systems based solely on traditional models of human creativity risks constraining their potential to explore new forms of innovation and expression – forms that may not align with human cognitive norms or aesthetic expectations. This recognition calls for a reconceptualization of creative agency, one that embraces the possibility of human-AI collaborations that *extend*, rather than replicate, the boundaries of creative practice.

Crucially, the creative value of an Extended Creativity system does not scale linearly with the level of machine autonomy or agency. Even in the Support configuration – where AI operates with limited adaptivity and causal influence – the delegation of routine, technical, or cognitively demanding tasks can amplify human creative focus, enabling more expansive ideation and deeper aesthetic exploration.

As AI systems become more sophisticated, new forms of creative collaboration will inevitably emerge – some of which may revise, or challenge key assumptions embedded in this framework. In parallel, as artists, designers, and researchers engage with these systems in novel ways, they will surface unanticipated interaction patterns that provoke theoretical refinement and practical innovation.

Acknowledgments

This work was supported by the “Extended Learning” project, an internal funding initiative of Università Cattolica del Sacro Cuore (Linea D3.2 – 2024). The authors would like to express their sincere gratitude to Francesco Giordano for his valuable contribution in creating the artistic illustrations that visually represent the conceptual framework presented in this paper.

References

- AMABILE T.M. (1996). *Creativity in context: Update to the social psychology of creativity*, Westview Press.
- BAXTER G., SOMMERVILLE I. (2011). *Socio-technical systems: From design methods to systems engineering*, in “Interacting with Computers”, 23(1), pp. 4-17.
- BERRETTA S., TAUSCH A., ONTRUP G., GILLES B., PEIFER C., KLUGE A. (2023). *Defining human-AI teaming the human-centered way: A scoping review and network analysis*, in “Frontiers in Artificial Intelligence”, 6, 1250725.
- BODEN M.A. (2004). *The creative mind: Myths and mechanisms*, Routledge (2nd ed.).
- BODEN M.A. (2008). *Autonomy: What is it? Introduction*, in “Biosystems”, 91(2), pp. 305-308. <https://doi.org/10.1016/j.biosystems.2007.07.003>
- CHIRIATTI M., GANAPINI M., PANAI E., UBIALI M., RIVA G. (2024). *The case for human-AI interaction as system 0 thinking*, in “Nature Human Behaviour”, 8(10), pp. 1829-1830.
- CHOI C., ANNIO F. (2024). *The winner of a prestigious Japanese literary award has confirmed AI helped write her book*, <https://edition.cnn.com/2024/01/19/style/rie-kudan-akutagawa-prize-chatgpt/index.html>
- CORAZZA G.E. (ed.) (2025). *The cyber-creativity process: How humans co-create with artificial intelligence*, Palgrave Macmillan.
- CSIKSZENTMIHALYI M. (1999). *Implications of a systems perspective for the study of creativity*, in R.J. STERNBERG (ed.), *Handbook of creativity*, Cambridge University Press, pp. 313-335.
- DAVIS N., HSIAO C.P., POPOVA Y., MAGERKO B. (2015). *An enactive model of creativity for computational collaboration and co-creation*, in *Creativity in the digital age*, Springer, London, pp. 109-133.
- DELL’ACQUA F., ELOUNDU T., GANGULI D., KAPLAN J., RE C. (2023). *Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality*, Harvard Business School Technology & Operations Management Unit Working Paper No. 24-013, Harvard Business School.
- FRANCESCHELLI G., MUSOLESI M. (2024). *On the creativity of large language models*, in “AI & Society”.
- GAGGIOLI A., RIVA G., MILANI L., MAZZONI E. (2013). *Networked Flow: Towards an understanding of Creative Networks*, New York, Springer.
- GLÁVEANU V.P. (2014). *Distributed creativity: Thinking outside the box of the creative individual*, Springer Science + Business Media.
- GONZALEZ C., ADMONI H., BROWN S., WOOLLEY A.W. (2023). *Introduction: COHUMAIN – Building the socio-cognitive architecture of collective human-machine intelligence*, in “Topics in Cognitive Science”, 15(3), pp. 457-460.
- HAASE J., POKUTTA S. (2024). *Human-AI co-creativity: Exploring synergies across levels of creative collaboration* [Preprint], in “arXiv”. <https://doi.org/10.48550/arXiv.2411.12527>

KAHNEMAN D. (2002). *Maps of bounded rationality: A perspective on intuitive judgment and choice*, in T. FRÅNGSMYR (ed.), *The Nobel Prizes 2002*, Nobel Foundation, pp. 449-489.

KAUFMAN J.C., BEGHETTO R.A. (2009). *Beyond big and little: The four c model of creativity*, in "Review of General Psychology", 13(1), pp. 1-12.

KRAKOWSKI S. (2024). *Human-AI agency in the age of generative AI*, in "Information and Organization", 34(2), 100811.

LITERAT I., GLAVEANU V.P. (2018). *Distributed creativity on the Internet: A theoretical foundation for online creative participation*, in "International Journal of Communication", 12, pp. 893-908.

MORUZZI C., MARGARIDO S. (2024, May). *A user-centered framework for human-AI co-creativity*, in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (CHI EA 2024) (Article No. 252), ACM.

LUBART T.I. (2001). *Models of the creative process: Past, present and future*, in "Creativity Research Journal", 13(3-4), pp. 295-308.

McCORMACK J., GIFFORD T., HUTCHINGS P. (2019). *Autonomy, authenticity, authorship and intention in computer generated art*, in *Computational Intelligence in Music, Sound, Art and Design*, 8th International Conference, EvoMUSART 2019, Held as Part of EvoStar 2019, Leipzig, Germany, April 24-26, 2019, Proceedings, Springer-Verlag, pp. 35-50.

McGUIRE J., DE CREMER D., VAN DE CRUYS T. (2024). *Establishing the importance of co-creation and self-efficacy in creative collaboration with artificial intelligence*, in "Scientific Reports", 14, 18525.

NAKAKOJI K. (2006). *Meanings of tools, support, and uses for creative design processes*, in *Proceedings of the International Design Research Symposium '06*, pp. 156-165.

RAMMERT W., SCHULZ-SCHAEFFER I. (2002). *Technik und Handeln: Wenn soziales Handeln sich auf menschliches Verhalten und technische Artefakte verteilt*, in W. RAMMERT, I. SCHULZ-SCHAEFFER (eds.), *Können Maschinen handeln? Soziologische Beiträge zum Verhältnis von Mensch und Technik*, Frankfurt, Campus Verlag, pp. 11-64.

RAFNER J., BEATY R.E., KAUFMAN J.C., LUBART T., SHERSON J. (2023). *Creativity in the age of generative AI*, in "Nature human behaviour", 7(11), pp. 1836-1838.

RAMMERT W. (2008). *Where the Action is: Distributed Agency between Humans, Machines, and Programs*, in U. SEIFERT, J. KIM, A. MOORE (eds.), *Paradoxes of Interactivity: Perspectives for Media Theory, Human-Computer Interaction, and Artistic Investigations*, Bielefeld, transcript Verlag, pp. 62-91.

RHODES M. (1961). *An analysis of creativity*, in "The Phi Delta Kappan", 42(7), pp. 305-310.

RUNCO M.A., JAEGER G.J. (2012). *The standard definition of creativity*, in "Creativity Research Journal", 24(1), pp. 92-96.

SCHMUTZ J.B., OUTLAND N., KERSTAN S., GEORGANTA E., ULFERT A.S. (2024). *AI-teaming: Redefining collaboration in the digital era*, in "Current opinion in psychology", 58, 101837.

- SCHREINER L., WIPPRECHT A., OLYANASAB A., SIEGHARTSLEITNER S., PRETL H., GUGER C. (2025). *Brain-computer-interface-driven artistic expression: real-time cognitive visualization in the pangolin scales animatronic dress and screen dress*, in “Frontiers in Human Neuroscience”, 19, 1516776.
- SHNEIDERMAN B. (2007). *Creativity support tools: Accelerating discovery and innovation*, “Communications of the ACM”, 50(12), pp. 20-32.
- SIMMLER M., FRISCHKNECHT R. (2021). *A taxonomy of human-machine collaboration: Capturing automation and technical autonomy*, in “AI & Society”, 36(1), pp. 239-250.
- STEIN M.I. (1953). *Creativity and Culture*, in “Journal of Psychology”, 36, pp. 311-322.
- THELLE N.J.W., WÆRSTAD B.I. (2023). *Co-Creative Spaces: The machine as a collaborator*, in *Proceedings of the International Conference on New Interfaces for Musical Expression*.
- THIMM C., GRAMELSBERGER G., MAYER M., PILLER F. (2024). *From automation to autonomy: Human-machine relations in the age of artificial intelligence*, in “Human-Machine Communication”, 9, pp. 7-24.
- UCSB MIND AND MACHINE INTELLIGENCE INITIATIVE (2024). *AI/Human Creativity Contest*. <https://mind-machine.ucsb.edu/news/all/2024/ucsb-aihuman-creativity-contest>
- VACCARO M., ALMAATOUQ A., MALONE T. (2024). *When combinations of humans and AI are useful: A systematic review and meta-analysis*, in “Nature Human Behaviour”, 8, pp. 2293-2303.
- VINCHON F., LUBART T., BARTOLOTTA S., GIRONNAY V., BOTELLA M., BOURGEOIS-BOUGRINE S., BURKHARDT J.-M., BONNARDEL N., CORAZZA G.E., GLAVEANU V., HANCHETT HANSON M., IVCEVIC Z., KARWOWSKI M., KAUFMAN J.C., OKADA T., REITER-PALMON R., GAGGIOLI A. (2023). *Artificial intelligence & creativity: A manifesto for collaboration*, in “The Journal of Creative Behavior”, 57(4), pp. 472-484.
- YAO S., WANG J., ZHENG L. (2024). *Generative Content Contests Between Humans and AIs: A Tullock Contest Model*, preprint in “rXiv”. <https://arxiv.org/abs/2402.15467>

3. Causal Reasoning for Artificial Intelligence

Guido Consonni, Federico Castelletti

ABSTRACT

As artificial intelligence and machine learning systems become increasingly embedded in real-world applications, they face important challenges: how to adapt to changing environments, and how to meet growing demands for interpretability, explainability, and fairness. Causal reasoning offers a promising path forward by shifting the focus from merely detecting patterns in data to understanding the underlying mechanisms of cause and effect. This expository chapter introduces the fundamental ideas and methods of causal modeling and causal discovery, with particular emphasis on graphical models as a conceptual framework and reasoning aid.

1. *From association to causation*

Artificial Intelligence (AI) and Machine Learning (ML) systems have achieved remarkable success in recent years, often by exploiting large volumes of data to make accurate predictions. Yet, as these systems become increasingly integrated into high-stakes applications – such as medical diagnosis, policy evaluation, or autonomous driving – they face severe limitations. The reason is that *prediction* models alone are insufficient when the environment changes relative to the training set, a phenomenon known as distribution shift, or when some aspects of the data generating system are modified through an external intervention.

To address these challenges, *causal* models are needed. By focussing on the underlying mechanisms that generate the data, rather than merely identifying patterns, they aim to answer questions such as: “What would happen if we were to intervene on a particular variable?” or “Was variable X a cause of outcome Y ?”. Answering these questions requires tools that go beyond conventional statistical learning models based on *associations*.

By modeling invariant causal mechanisms rather than surface correlations, not only does causal inference extend the scope and enable *robustness* to changes in environment, but it also contributes

towards a more ethical and socially-aware AI. In particular it offers greater *interpretability* of current AI systems by making assumptions explicit and often visual, supporting more transparent reasoning and human oversight. Additionally, it provides a principled foundation for *explainability*, particularly through counterfactuals – allowing questions such as: “Would the outcome have been different if we had acted otherwise?”. Finally, causality is also essential for *fairness*: purely statistical criteria often fail when protected attributes (like race or gender) influence decisions indirectly. Causal frameworks allow us to separate legitimate pathways from discriminatory ones, using concepts like counterfactual fairness and path-specific fairness.

Current developments in causal models owe much to two intellectual lineages. The first, dominant in the social sciences, builds on the *potential outcomes* framework developed in Rubin (2005). For unit i , the causal effect of Treatment T on outcome Y_i is defined as $Y_i(1) - Y_i(0)$, where $Y_i(t)$ is the outcome if unit i had been given treatment $T = t$ (assuming for simplicity that only two levels of T are available, e.g. the current drug and a new drug). Notice that the pair $(Y_i(1), Y_i(0))$ is only partially observable, because at any given time each unit can only be assigned to one of the two treatments. In Rubin’s framework this issue is addressed as a ‘missing data’ problem.

The second lineage centers on *graphical models*. Here, the term ‘graphical’ does not simply refer to pictures or diagrams, but to the use of a graph in the mathematical sense: a collection of points (called nodes) connected by arrows (called edges). These graphs are not just illustrations – they represent the structure of causal relationships between variables in a system. Spearheaded by Judea Pearl, this paradigm introduces tools such as the do-calculus and structural causal models, providing a formal language for encoding and reasoning about causal assumptions. Pearl’s book (Pearl, 2009) crystallized decades of work, offering a coherent framework that has profoundly influenced the machine learning and AI communities.

This article aims to examine the principles of causal reasoning, with a particular focus on their relevance to artificial systems. The graphical model framework will provide a unifying perspective throughout the discussion; for a critical assessment of this approach, see Imbens (2020).

2. The ladder of causality

A key contribution to modern causal inference is the metaphor of the *ladder of causality* (Pearl and Mackenzie, 2018). It represents three distinct levels of reasoning about data and the world.

<i>Level</i>	<i>Description</i>
1	<i>Association</i> ('seeing'): observing patterns and correlations in data
2	<i>Intervention</i> ('doing'): reasoning about the effects of actions or changes
3	<i>Counterfactuals</i> ('imagining'): reasoning about alternative scenarios and what could have happened

Level 1: Association concerns itself with regularities in the data – statistical dependencies among variables. This is the domain of traditional supervised machine learning. A model may learn that people who buy toothpaste also tend to buy toothbrushes, or that certain gene expressions correlate with specific diseases. But association cannot distinguish cause from effect, nor can it predict the outcome of interventions.

Level 2: Intervention answers questions of the form: what happens if we exogenously manipulate X and set it equal to a given value? This requires reasoning about how the world would change under said manipulations. For example, what will happen to the unemployment rate if a job training program is introduced? The formal tool to model such interventions is the *do-operator*, denoted $\text{do}(X = x)$, which represents an external setting of variable X to value x , breaking the natural causal mechanisms that would otherwise determine X .

Level 3: Counterfactuals goes even further: it asks not only what *would* happen under an intervention, but what *would have happened*, given that something else actually did. A typical question might be: "Would the patient have recovered had they received a different treatment, even though they actually received the standard one and did not recover?".

While machine learning excels at Level 1, it is mostly silent on Levels 2 and 3. Bridging this gap is central to building intelligent systems capable of generalizing beyond their training environment, understanding cause and effect, and offering meaningful explanations. The remainder of this article explores the mathematical and conceptual tools necessary to ascend the ladder.

3. Causal models

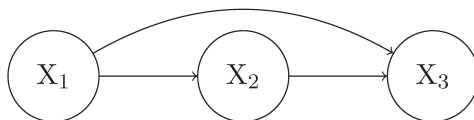
To carry out analyses at the various levels of the ladder we need tools that express both statistical dependencies (Level 1) and causal mechanisms (Level 2 and 3). It turns out that for the first two levels *Directed Acyclic Graphs* represent an effective solution, while Level 3 requires a more structured model.

– Directed Acyclic Graphs

A Directed Acyclic Graphs (DAG) is a graph where nodes represent variables and arrows represent direct causal effects. The ‘acyclic’ requirement means there are no feedback loops; one cannot start at a node and follow a sequence of arrows back to that same node. Consider a Level 1 analysis where interest centers on the statistical properties exhibited by a set of variables. For instance in a clinical setting the variables may be the disease type of a patient together with a few diagnostic indicators (e.g. X-rays, blood exams) and context variables (e.g. morbidity and lifestyle). In other contexts, the number of variables can be very high as with the expression levels of hundreds of selected genes relevant for a particular disease. A DAGs is a compact representation of the joint dependencies among the variables.

Consider the following simple DAG

Figure 1 – DAG G



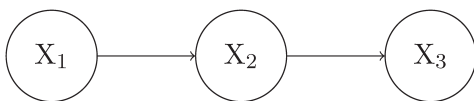
Before collecting data we do not know the value of the variables, which are accordingly *random*. The DAG in Figure 1 should be interpreted in the following sense: the *distribution* of X_2 depends on X_1 , while that of X_3 depends on (X_1, X_2) . More specifically the joint distribution of (X_1, X_2, X_3) must *factorize* according to the graph G

$$p(X_1, X_2, X_3) = p(X_1)p(X_2 | X_1)p(X_3 | X_1, X_2). \quad (1)$$

If $U \rightarrow V$ then U is a *parent* of V ; equivalently V is a *child* of U . In graph G , X_1 has no parents, while the parents of X_3 are (X_1, X_2) .

Now suppose the DAG is

Figure 2 – DAG G'



so that the joint distribution becomes

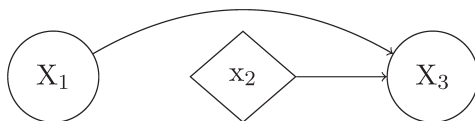
$$p(X_1, X_2, X_3) = p(X_1)p(X_2 | X_1)p(X_3 | X_2). \quad (2)$$

In this case we say that X_3 is *conditionally* independent of X_1 *given* X_2 . For instance if X_1 is ‘height’, X_2 is ‘age’ and X_3 is ‘reading ability’ of small children, then G' seems appropriate: if I have information on the age of a child, then height becomes irrelevant to predict reading ability. Notice however that ‘age’ and ‘reading ability’ will typically be dependent *marginally*, that is without conditioning. More generally, the factorization of a joint distribution according to a graph encapsulates a collection of conditional independencies. Specifically, it can be shown that each variable U is conditionally independent of its *non-descendants* given its *parents*, where the non-descendants of U are those nodes which cannot be reached starting from U along a directed path. For instance X_1 is a non-descendant of X_3 in graph G' .

– Causal Graphical Models

To address Level 2 of the ladder we need a Causal Graphical Model (CGM) which combines a DAG G with a joint distribution p that satisfies the factorization property (as in Level 1). The additional crucial assumption is that arrows now represent causal links in the following precise sense: external *interventions* are modeled *via* a simple graphical procedure, named *graph surgery*. Specifically, an intervention on X_i which sets it equal to x_i , represented by $\text{do}(X_i = x_i)$, removes all incoming arrows into X_i . This makes intuitive sense: suppose X_i is the expression level of gene i and we force it to take a specific value; then X_i will no longer depend on its parent genes, as in the intervention-DAG below

Figure 3 – *Intervention-DAG following $\text{do}(X_2 = x_2)$ under G*



where the diamond around x_2 means that X_2 is no longer random. From Figure 3 we immediately derive the joint *interventional* distribution

$$p(X_1, X_3 | \text{do}(X_2 = x_2)) = p(X_1)p(X_3 | X_1, x_2), \quad (3)$$

where X_2 is explicitly recognized to be a fixed value x_2 . Notice that the interventional graph of Figure 3 removes only the incoming arrows in X_2 while keeping the rest the same as in G ; similarly, the modules which

appear in the factorization of the joint interventional distribution (3) are exactly those appearing in the observational distribution (1), save for the module related to node X_2 because it is no longer random. This *stability* to external manipulations will play a fundamental role in making causal inference practically feasible.

A key insight of causal modeling is the distinction between passively *observing* X_2 and actively *intervening* on X_2 . For better clarity let the observed and set value be the same, i.e. x_2 . To start with, suppose we *observe* $X_2 = x_2$: “how does our uncertainty on X_3 change”? The answer is straightforward: simply compute the *conditional* distribution $p(X_3 | X_2 = x_2)$. On the other hand, if we *set* $X_2 = x_2$ we need to compute $p(X_3 | do(X_2 = x_2))$. It can be shown that in this example $p(X_3 | X_2 = x_2) \neq p(X_3 | do(X_2 = x_2))$. The distinction between observing and intervening is crucial in applications ranging from policy evaluation to algorithmic fairness. For instance, if we wish to understand the effect of *enforcing* a particular hiring policy, we must reason in terms of do-calculus and not simply analyze the existing patterns of hiring in our dataset conditionally on that policy.

We remark that the interventional distribution is typically unknown, or perhaps only known up to some parameters to be determined. However it can be estimated using observational data because of the invariance of each module alluded to earlier on.

– Structural Causal Models

A CGM can deal with Level 2 causal questions but is not suited to address Level 3 counterfactual queries. Counterfactuals are arguably the most subtle and powerful aspect of causal reasoning. They allow us to ask ‘what if’ questions about alternate realities. “Given that Peter received treatment A and his health got worse, what would have happened if he had been given treatment B instead, *all else being equal*?” This can be formally achieved within a Structural Causal Model (SCM). An SCM specifies a system of equations that deterministically assigns each variable X_i , $i = 1, \dots, n$ as a function of its parents PA_i and an independent noise term U_i

$$X_i := f_i(PA_i, U_i).$$

The randomness in the system originates from the joint distribution of the noise terms $(U_1, \dots, U_n)$, which are typically assumed to be independent. An SCM induces a DAG through the functional relationships among variables. The resulting structure satisfies the same conditional independencies as in the corresponding CGM, under assumptions of acyclicity and causal sufficiency (no unmeasured confounders).

– Example: a simple linear SCM

Consider the two-variable SCM

$$\begin{aligned} X &:= U_X, \\ Y &:= 3X + U_Y, \end{aligned}$$

where $U_X, U_Y \sim \mathcal{N}(0, 1)$ independently, where $\mathcal{N}(0, 1)$ is a standard Gaussian distribution. Suppose we observe $X = 2$ and $Y = 6.5$. Then we infer $U_X = 2$ and $U_Y = 0.5$. Now we ask the counterfactual question: “what would Y have been if X had been set to 1 instead?”.

We substitute $X = 1$ into the equation for Y , keeping $U_Y = 0.5$

$$Y(1) = 3 \times 1 + 0.5 = 3.5,$$

which represents the counterfactual answer. Note that this result differs from the purely interventional answer; in particular $\mathbb{E}[Y \mid \text{do}(X = 1)] = 3$.

Summing up, the counterfactual $Y(x')$, given evidence $X = x, Y = y$, is computed as follows

- (i) Update the noise distribution $p(U_X, U_Y \mid X = x, Y = y)$ and obtain the updated SCM.
- (ii) In the updated SCM set the assignment of X to x' .
- (iii) Compute the interventional distribution of Y in the updated SCM.

We further elaborate on the counterfactual query we asked at the beginning of this section, where the goal is to assess what would have happened to the patient if, instead of A, they had been given treatment B, all else being equal. In the language of SCMs the phrase ‘all else being equal’ should be interpreted as using the unobserved background factors captured by the noise terms at their inferred values revealed by the fact that treatment A was administered and led to deterioration. Such values represent the patient’s specific underlying conditions, and remain the same across the factual and counterfactual scenarios. Thus counterfactual reasoning preserves the acquired individualized knowledge, while interventions compute average outcomes in a population without conditioning on previous evidence. Although several authors have questioned the scientific legitimacy of counterfactuals due to their lack of empirical verifiability – see, for example, Dawid (2000) – psychological research suggests that humans routinely engage in counterfactual reasoning from a very young age (Buchsbaum et al., 2012).

4. Independent causal mechanisms

At the heart of causal reasoning lies the idea that the world is governed by *independent mechanisms* (ICM) - modular components that explain how causes bring about effects. Each of these mechanisms represents a causal process that operates autonomously, meaning it could in principle be modified or replaced without affecting the others. Whereas statistical models focus on learning associations between variables, causal models aim to recover these underlying mechanisms. The *ICM principle* posits that the causal relationships in the world are structured as separate, self-contained modules. This modularity allows for the kind of reasoning humans naturally engage in: if one aspect of a system changes, we can predict how other parts will or will not be affected.

To illustrate the topic, consider a dataset consisting of measurements of *altitude* A and *average annual temperature* T at various weather stations. Intuitively, altitude causes temperature: as you ascend a mountain, the air becomes thinner and cooler. This is due to a physical mechanism – air pressure and density change with altitude, affecting heat retention.

We can express this relationship using a structural equation (Peters et al., 2017)

$$T := f(A) + U_T,$$

where f is a deterministic function capturing the temperature-altitude mechanism, and U_T represents noise or unexplained variation.

If the mechanism is truly *causal*, then f should remain invariant across different regions or climate zones. That is, the way altitude affects temperature does not depend on where in the world you are; the mechanism generalizes. But the distribution of altitude values $P(A)$ might differ across datasets (say, one from the Alps and one from the Andes). The key idea is that the mechanism f , which governs T given A , is *algorithmically independent* of the distribution of A .

Abstracting from this example, suppose we model the joint distribution $P(X, Y)$ of two variables by decomposing it into a marginal and a conditional:

$$P(X, Y) = P(X) \cdot P(Y | X).$$

If X causes Y , the claim is that the marginal $P(X)$ and the conditional $P(Y | X)$ are independent components in an algorithmic sense: knowing how the distribution of X was generated should give you no information about the mechanism $P(Y | X)$, and vice versa. While from a probability perspective we might equivalently rewrite $P(X, Y) = P(Y) \cdot P(X | Y)$,

this is not the case from a causal perspective. In our example the components $P(T)$ and $P(A \mid T)$ are inextricably entangled because T is not a cause of A .

Though algorithmic independence is not computable in practice, it supports the philosophical intuition behind ICM: that nature’s causal processes operate independently and are not ‘coordinated’ with the statistical properties of their inputs (Schölkopf and von Kügelgen, 2022).

In summary, the ICM principle enables a range of practical applications, among which: i) robustness under interventions: since mechanisms are autonomous, if you intervene on one variable, other mechanisms remain unchanged; ii) generalization across domains: mechanisms can often be transferred to new settings, even when the input distributions change; iii) modular learning: ICM encourages learning individual parts of a system in isolation, supporting scalability and interpretability.

5. Causal discovery

So far we have assumed that the causal graph or the SCM was known or postulated. However, in most real-world situations the underlying causal structure must itself be inferred from data, a problem known as *causal discovery*. For example, a scientist may collect data about multiple interacting biological factors but has at best only partial knowledge about how these factors causally influence each other. Unlike traditional statistical learning, which focuses on associations or correlations, causal discovery seeks to learn the data generating mechanisms, i.e. the entire DAG structure, which represents a more ambitious task. The latter can be achieved under different settings, depending on the type of data that are available or can be collected.

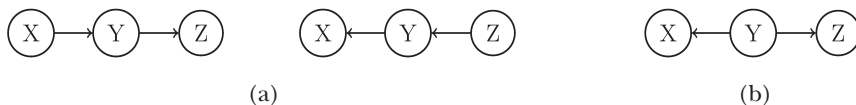
5.1. Causal discovery from purely observational data

In most cases, researchers must rely on *observational data*, namely measurements that are collected by observing events as they naturally occur, that is without any possibility to control the level of variables by manipulation. Because a DAG entails conditional independence statements between the variables, as we saw in Section 3, statistical *testing* methods can be used to reconstruct the structure of the DAG (Spirtes et al., 2000). However this can be achieved only partially because different DAGs can encode the same set of (conditional) independencies, and so even a perfect testing procedure cannot identify a unique DAG. Luckily however, DAGs can be grouped into (Markov) *equivalence classes*, where

each class contains exactly those DAGs encoding the same independencies. Based on purely observational data, and without any domain knowledge of the DAG structure, what we can hope to learn is therefore an equivalence class of DAGs.

For instance the following three DAGs belong to the same equivalence class because they encapsulate the conditional independence that X is independent of Z given Y (and only this).

Figure 4 – *Three (Markov) equivalent DAGs. (a) Chains and (b) Fork.*



The critical point is that the above three DAGs, although equivalent from a statistical viewpoint, are distinct from a causal perspective. For instance, while X is a cause of Y in the leftmost DAG of Figure 4, this relationship is reversed in the remaining two DAGs. One way to resolve the indeterminacy inherent in the equivalence class is to collect further *interventional* data, as we detail in the next section. By jointly modeling observational and interventional data the actual directionality of some of the arrows can be resolved Hauser and Buehlmann, 2015. In this context, there exist ways to optimally select the interventional target for reconstructing the causal DAG, so called *active learning*; see He and Geng (2008) and Castelletti and Consonni (2024).

5.2. Causal discovery from experimental data

In principle, the most reliable way to discover causal relationships is through *experimentation*. Within this context, randomized controlled trials (Fisher, 1935) form the foundation of modern evidence-based medicine. A/B testing illustrates this well: it is a standard method in engineering for experimentally optimizing causal effects. Consider two variables: ‘treatment’ T and ‘outcome’ Y . If T is binary then $T = 0$ corresponds to ‘placebo’ while $T = 1$ corresponds to ‘drug’. Patients are split into two groups by means of a coin flip that assigns a value of T to each patient, and then the assigned treatment is administered. Subsequently Y is observed. The resulting dataset is (T_n, Y_n) , $n = 1, \dots, N$, where N is the number of patients. If the distribution of outcome Y is substantially different between the two groups, one can conclude that T is a cause of Y provided the following causal assumptions hold:

- i) Y is not a cause of T , a feature which is satisfied when the outcome event follows in time the treatment event;
- ii) there is no latent confounder acting as a common cause of T and Y , which can be guaranteed if randomized treatment assignments are employed;
- iii) there is no selection bias in the data, an assumption which would be violated for instance if only people without treatment side effects were included in the dataset.

Intervention experiments, as those exemplified above, are typically expensive, unethical, or impossible. Also, in many fields, such as medicine or social sciences, only a limited number of interventions can be performed. Nevertheless, real-world data are often intrinsically interventional, because collected from multiple *contexts* or *environments* – for example, under different experimental conditions, locations, or time points. Traditional causal discovery methods usually ignore this feature by pooling all data together. However, this may lead to misleading conclusions because the relationships between variables may change across contexts, in force of interventions, possibly unknown.

Learning causal relations between variables is instead possible by jointly modelling such data, and in particular by identifying differences in the dependence structure across contexts. Several statistical methodologies have been proposed in this framework, for the cases of both known and unknown manipulated variables, so-called *intervention targets*, T . For the former case, recent contributions include Hauser & Bühlmann (2015) and Castelletti and Peluso, 2024. In the more general setting of unknown targets, one can first learn them by identifying which variables show changes in their distribution across contexts. Next, causal relations are learned by detecting whether the dependence between target variables and any possible related outcome Y changes across contexts; see for instance Hägele et al. (2023).

The methods discussed above involve different *types* of interventions, depending on the problem at hand. In addition to the *deterministic* interventions used for illustration in this paper, there are also *stochastic* interventions, where the mechanism generating a variable is altered to produce random outcomes. A typical example is adjusting the dosage of a drug via recommendations, rather than by directly enforcing a treatment – this introduces variability that depends, for instance, on patient compliance. Another useful distinction is between *hard* and *soft* interventions. In a hard intervention, the causal link between a variable

and its parent nodes is completely severed. In a soft intervention, by contrast, the link is not removed but modified, typically by changing the conditional distribution without breaking dependencies entirely.

More generally, interventions can be viewed as changes to the causal model itself – either by altering the functional relationships within the system or by shifting the distributions involved (Eberhardt et al., 2007; Pearl, 2009). Recognizing these different forms of intervention is essential for developing flexible and realistic causal models that reflect the complexity of real-world decision-making.

6. *Conclusions*

Modern machine learning systems are often optimized for predictive accuracy within a fixed data distribution. However, real-world deployment typically involves changes – in population, behavior, policy, or context – that violate the assumptions of static, i.i.d. data. Causal reasoning provides a framework to address these challenges by separating stable, invariant mechanisms from domain-specific noise. This property underpins approaches such as domain adaptation, where the goal is to leverage data from multiple source environments (Mooij et al., 2020) to make robust predictions in a new target domain.

In many real-world systems, not all relevant causes are measured or even measurable. *Latent variables*, that is unobserved factors that may influence the relationships among observed variables, can obscure causal relationships, introduce bias, and complicate both discovery and inference. Several approaches have been developed to address this challenge, such as using instrumental variables (Imbens, 2014), graphical models with latent nodes, or advanced algorithms capable of handling hidden confounders (Tian and Pearl, 2008). Incorporating latent variables remains a central and ongoing area of research in causal modeling, and further work is needed to fully integrate these considerations into AI and machine learning systems.

Causal modeling operates at different levels of granularity and abstraction. The set of variables we choose to model often determines whether a causal explanation is even possible at all. In some cases, the observed variables may not admit a clear causal structure – what appears as mere statistical association at one level may correspond to an identifiable causal mechanism at another. For example, consider weather patterns: if we only observe the co-occurrence of high humidity and thunderstorms, we might not find a causal model connecting these directly. However, by introducing additional variables such as atmospheric pressure and temperature gradients, a causal explanation

emerges from meteorology. This illustrates that causal discovery is sometimes not just about learning relationships between given variables, but about identifying the *appropriate level of description* where meaningful causal mechanisms can be formulated (Schölkopf and von Kügelgen, 2022).

Acknowledgments

Work carried out within MUR-PRIN grant 2022 SMNNKY - CUP J53D23003870008. Funded by the European Union - Next Generation EU. The views and opinions expressed are only those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

- BUCHSBAUM D., KUSHNIR T., GOPNIK A., WELLMAN H.M. (2012). *Children’s imitation of causal action sequences is influenced by statistical and pedagogical evidence*, in “Cognition”, 120, pp. 331-340.
- CASTELLETTI F., CONSONNI G. (2024). *Bayesian sample size determination for causal discovery*, in “Statistical Science”, 39, pp. 305-321.
- CASTELLETTI F., PELUSO S. (2024). *Bayesian learning of network structures from interventional experimental data*, in “Biometrika”, 111, pp. 195-214.
- DAWID A.P. (2000). *Causal inference without counterfactuals*, in “Journal of the American Statistical Association”, 95, pp. 407-424.
- EBERHARDT F., GLYMOUR C., SCHEINES R. (2007). *Interventions and causal inference*, in “Philosophy of Science”, 74, pp. 981-995.
- HÄGELE A., ROTHFUSS J., LORCH L., SOMNATH V.R., SCHÖLKOPF B., KRAUSE A. (2023). *Bacadi: Bayesian causal discovery with unknown interventions*, in *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS 2023)*.
- HAUSER A., BÜHLMANN P. (2015). *Jointly interventional and observational data: Estimation of interventional Markov equivalence classes of directed acyclic graph*, in “Journal of the Royal Statistical Society Series B: Statistical Methodology”, 77, pp. 291-318.
- HE Y.-B., GENG Z. (2008). *Active learning of causal networks with intervention experiments and optimal designs*, in “Journal of Machine Learning Research”, 9, pp. 2523-2547.
- IMBENS G.W. (2014). *Instrumental variables: An econometrician’s perspective*, in “Statistical Science”, 29, pp. 323-358.
- IMBENS G.W. (2020). *Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics*, in “Journal of Economic Literature”, 58, pp. 1129-1179.

MOOIJ J.M., MAGLIACANE S., CLAASSEN T. (2020). *Joint causal inference from multiple contexts*, in “Journal of Machine Learning Research”, pp. 211-108.

PEARL J. (2009). *Causality: Models, Reasoning and Inference*, Cambridge University Press, 2nd ed.

PEARL J., MACKENZIE D. (2018). *The Book of Why: The New Science of Cause and Effect*, USA, BasicBooks Inc., 1st ed.

PETERS J., JANZING D.N, SCHÖLKOPF B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*, The MIT Press.

RUBIN D.B. (2005). *Causal inference using potential outcomes*, in “Journal of the American Statistical Association”, 100, pp. 322-331.

SCHÖLKOPF B. - VON KÜGELGEN J. (2022). *From statistical to causal learning*, in *Proceedings of the International Congress of Mathematicians (ICM)*, vol. VII, EMS Press, pp. 5540-5593.

SPIRITES P., GLYMOUR C., SCHEINES R. (2000). *Causation, Prediction, and Search*, MIT Press, 2nd ed.

TIAN J., PEARL J. (2008). *On the testable implications of causal models with hidden variables*, in *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pp. 519-527.

4. The Convergence of Artificial Intelligence and Gaming

A Human-Centered Perspective

Daniele Toti

ABSTRACT

This chapter discusses the deep impact that artificial intelligence (AI) has had and is still having on the world of gaming, looking at both the technological development of AI in games as well as its game-altering (pun intended!) implications for the lives of gamers. This discussion starts, therefore, by chronicling the evolution of game AI, from the initial rule-based systems to recent breakthroughs in deep learning and generative models, with an emphasis on the major advancements that have transformed non-player character behavior, procedural content generation and adaptive gameplay. The chapter continues by discussing how AI-powered innovation is transforming game development through automatic asset creation, personalized player experiences and more complex, emotionally engaging virtual worlds. Then, the discussion moves on to the human level in terms of the players' own feelings and perspectives, related to the formation of meaningful relationships with AI characters, AI's involvement in supporting mental well-being and accessibility, and the evolving boundaries of social interaction within mixed human-AI gaming communities. A few ethical concerns regarding data privacy, psychological influence and environmental sustainability are also debated. Finally, the chapter closes with the argument that the future of games lies in the symbiotic collaboration of artificial intelligence and human imagination to provide richer, more inclusive, and more essentially *human* gaming experiences.

Introduction

The latest advancements in artificial intelligence have brought about unparalleled changes in the gaming world, reshaping the creation of games and the way they are played by humans. The result is not just a technological advancement: it 'remakes' the human gaming experience of the digital era. AI-based technologies are also changing non-player characters from static, scripted beings into responsive, dynamic companions, and opening up new possibilities for mental health therapy, accessibility and personalized gameplay experiences that can adjust to the needs and desires of individual players.

1. *The evolution of game AI: from rules to ‘intelligence’*

1.1. The historical arc of game AI development

The development of artificial intelligence has an interesting history that more or less mirrors the technological advancements occurred over the years.

The first versions of ‘AI-based’ game systems were fundamentally systems based on rules, enacting pre-specified reactions to clearly defined input in specifically restricted and constrained game worlds. These early systems, while somehow effective in well-understood contexts, were extremely lacking in terms of flexibility and required extensive manual tuning by the developers (Zheng, 2025). Game AI initially appeared in classic arcade games, wherein enemy behavior was derived from set patterns that could be memorized by the player and beaten through repetition and exposure. In this sense, such systems could certainly be considered ‘artificial’, but not really ‘intelligent’ by today’s standards, and were more akin to the production of ‘obstacles’ to the player than providing dynamic, flexible opponents or challenges.

Such early systems somehow got gradually replaced by slightly more advanced AI techniques, given the rise of computational power that occurred between the 1990s and 2000s: techniques like finite state machines, behavior trees and relatively simple pathfinding algorithms (Risi and Preuss, 2020). These techniques, applied to videogame enemies, enabled the latter to perform more varied actions and gave the impression of ‘intelligence’ thanks to hand-made decision trees and conditional reactions. Along these lines, games like *Black & White* and *The Sims* were huge steps forward during this period, featuring AI systems that were able to learn from the actions of the players and modeled a more sophisticated decision-making process. While better than their earlier counterparts, the underlying structure of those systems was still ‘reactive’ rather than ‘adaptive’: the behavior of the AI was still limited by the parameters defined by the developers themselves and there was no true ‘learning’ process involved. This reflected the trend of such a period, which was more interested in giving out the appearance of ‘intelligence’ than creating a real cognitive process; besides, at that time, there was little interest in AI research per se, with developers mainly focused on simply making the player experiences more pleasant.

More recently, the modern game AI has been transformed by the infusion of machine learning paradigms, specifically deep reinforcement learning (DRL), to allow game characters and systems to learn strategies through autonomous learning as opposed to explicit, manual program-

ming (Zheng, 2025). This paradigm shift is more than an incremental improvement: it fundamentally alters the relationship among developers, players, and the virtual worlds they inhabit. Nowadays, modern game AI has the ability to correlate enormous amounts of player action data, react to unforeseen events and develop novel approaches to complex problems without explicit human intervention (Zheng, 2025). This current trend reflects that of the artificial intelligence research in general, with games providing a perfect testbed for novel approaches because of their formally defined sets of rules and quantifiable/verifiable definitions of success.

The remainder of this section reviews the major breakthrough moments of AI in videogames, and then moves to introduce the current technical approaches and challenges of modern game AI.

1.2. Breakthrough moments in game AI

There were several milestone achievements in game AI history that served as turning points, opening up new potential for intelligent systems within game worlds.

Most notably, perhaps, was the breakthrough in 1997 when IBM's Deep Blue beat world chess champion Garry Kasparov to show that computers can outperform humans at strategy games (Risi and Preuss, 2020). Although chess may be considered a game with perfect information and clear parameters, this achievement nevertheless raised public interest in what was possible using artificial intelligence in a gaming environment. The rigidly rule-based nature of Deep Blue, however, limited its applicability to more complex gaming environments.

An even greater success for modern videogames was when DeepMind's AlphaGo defeated world champion Lee Sedol at Go in 2016, showing the promise of neural networks to learn to play games of very high complexity through self-play and reinforcement learning (Risi and Preuss, 2020). In contrast to chess, Go has an enormously greater decision space, which was once thought to be too complex (and too dependent on human intuition) to be played well by a computer, which made this victory highly important for the future of game AI. The methods created for AlphaGo have, in turn, shaped the direction of AI in commercial videogames, especially in strategy videogames.

In videogames alone, OpenAI's creation of AI systems that could beat professional players at Dota 2 (Defense of the Ancients 2, the very popular sequel of another popular multiplayer battle arena game) was a milestone for game AI in the context of dynamic, imperfect information settings (Zheng, 2025). These systems demonstrated the ability to direct team-based strategies, manage resources and execute complex tactical

maneuvers in real-time, abilities believed to be the sole domain of humans up to that point. Similarly, the dominance of AlphaStar in StarCraft II showed the AI's ability for real-time strategy games in controlling the various units simultaneously, making strategic plans and adapting to the player's moves (Zheng, 2025). These achievements proved that modern AI techniques could be really effective when employed for split-second decision-making and long-term planning and response to random moves from their (human) opponents.

1.3. Current technical approaches and challenges

Modern game AI development is increasingly taking advantage of DRL architectures, which enable systems to learn optimal behaviors through trial-and-error experiences based on reward functions (Zheng, 2025). While such approaches have achieved an incredible success across a wide range of gaming environments, they simultaneously present considerable implementation obstacles for commercial game developers. In fact, the computational requirements of training complex neural networks continue to be prohibitive for many studios, with training sessions for state-of-the-art game AI systems potentially requiring as much as 18.7 MWh of power (Zheng, 2025). This economic and environmental cost hinders the fulfillment of the demand for ever-smarter game systems and real-world realities.

Data efficiency is the second grand challenge, because deep learning approaches typically require millions of gameplay iterations to infer good strategies (Zheng, 2025). This requirement makes the application of learning systems to games difficult, where plenty of training data may not be available or where behavior must change quickly in reaction to new player input. In this regard, researchers are currently investigating hybrid methods that may combine shallow neural networks with symbolic logic systems in order to enable better learning without losing the responsiveness that defines contemporary AI methods (Zheng, 2025). In principle, hybrid models may provide a more sustainable direction toward game AI creation, balancing advanced behavior while avoiding (or solving altogether) pragmatic implementation issues.

Cross-genre generalization is also a problem that makes AI gaming adoption more complicated, since AI systems that are trained to deal with a specific game environment find it hard to be effective when exported to other gameplay environments (Zheng, 2025). Unlike human players, who can apply strategic thought to play other games with a relatively minimal adaptation time, AI systems require extensive re-training when they change gaming environments. This limitation has been encouraging the creation of multi-sensory perception mechanisms

and adaptive reward functions that may enable greater AI flexibility within other game worlds (Zheng, 2025). This attempt at more general game AI reflects broader ambitions within the broad spectrum of artificial intelligence research, with games representing convenient testbeds for developing systems with a flexible intelligence, which may be deployed in a variety of contexts.

2. How game development is being transformed by AI

This section describes some major changes in game development brought about by AI: specifically, the automated content generation and design, the enhancement of the players' experience and the so-called 'creative partnership' that is now being established between human developers and AI systems.

2.1. Automated content generation and design

Artificial intelligence is transforming the game development industry, among other things, by also enabling the automation of content creation processes that previously required enormous amounts of manual work. As a matter of fact, AI-powered procedural content generation is now capable of allowing developers to automatically create complex game scenes, character models and environmental textures that are artistically consistent and save production time (Wang, 2024). This is especially useful for open-world games and very large-scale games, where creating each piece by hand would be exceedingly time-consuming and unsustainable in the long term. Sophisticated generative models can create terrain, vegetation, buildings, atmospheric effects, etc., that respond dynamically to the game variables, yet complying with the aesthetic direction set by the human game designers.

Asset creation is not just restricted to graphical elements: also, narrative content, mission design and gameplay scenarios that are responsive and dynamic to player choices and behavior (Wang, 2024) can now be generated and produced in a similar fashion. In fact, AI systems can now generate branching storylines with internally consistent logic, create character dialogues that reflect established personalities and relationships and design mission objectives that appropriately challenge players based on their demonstrated skill levels. This element, along with the evolution of even more advanced and user-friendly game engines (see the success of Unity, for example), now allows smaller development groups to produce content on scales that

have previously only been possible for large studios with significant resources, potentially democratizing game development and enhancing creative diversity in the industry. These automated systems must, nevertheless, be used carefully in order not to produce excessively repetitive or formulaic content that players can identify as artificial or lacking in creative intent, resulting in (paradoxically?) a negative effect on the players themselves and their potential adoption/purchase of games featuring such content.

The use of AI in game development also extends to technological optimization techniques, where the intelligent machines analyze performance bottlenecks, suggest rendering improvements and identify potential faults before delivery to the player base (Wang, 2024). Machine algorithms are able to predict how different setups of hardware handle game assets, so that more efficient use and better performance can occur across diverse playing environments. This technological use of AI is largely backgrounded to the player but still plays a significant role in their experience through more fluid gaming, reduced loading times and enhanced performance stability. Indeed, the more technologically complex the game is, the more use of this type of AI-powered optimization software is needed to compensate for the technological complexity and deliver uniform experiences across different platforms.

2.2. Enhancing the player's experience via AI-driven data analysis

Artificial intelligence has also changed the way developers understand and react to the activity of players by supporting sophisticated data analysis functions that handle massive amounts of gameplay data (Ruthvika, 2025). Indeed, contemporary games try to gather data linked with player involvement, interests, challenge experience and interaction intensity, developing complex datasets that AI programs can navigate to look for trends, issues and/or areas of improvement. A mechanism like this enables developers to detect unplanned spikes in difficulty, features that had somehow been neglected or sections in which players experience a high degree of frustration, regardless of any explicit feedback from the players themselves via reviews or forums. By understanding how players actually play their games, and not what they claim to do, developers can make more informed design decisions based on the empirical evidence of the players' behavior.

In this sense, one of the most important uses of AI-driven analytics in this context is the personalization of gaming experiences, with adaptive systems altering gameplay mechanics according to individual player trends and interests (Ruthvika, 2025). Such systems can dynamically

alter difficulty levels according to player performance, recommend content according to expressed interests, or adjust narrative mechanics to highlight story points of interest to particular players. This is accomplished *explicitly* through player-driven choices and *implicitly* through analyzing gameplay activity to generate experiences that are sensitive to the personal pattern of involvement. The ultimate goal of this personalization is to maintain optimal challenge levels and engagement, helping players remain in psychological ‘flow states’, where challenges match their skill levels and progression feels rewarding rather than frustrating or tedious.

AI-powered analytics also allow developers to discover new groups of players and patterns of behavior that can unlock new design potential or new content possibilities (Ruthvika, 2025). By seeing players naturally forming groups around provided play styles, social interactions or creative uses of the game systems, developers can learn more about their players and develop content that is appealing to various player segments. This basically positions game design as a continual dialogue between players and creators, with artificial intelligence acting as the translator that is capable of turning players’ behavior into insights that can be used for performing corrective or evolutionary actions onto the games themselves. The outcome is a more reactive development cycle that is able to respond to what players desire without sacrificing quality or creative vision.

2.3. The creative partnership between AI and human developers

The application of artificial intelligence in game development offers a novel synergy for creativity where AI technology *augments* rather than *replaces* human creativity, allowing developers to concentrate on high-level design intent while automated processes manage implementation details (Wang, 2024). This partnership leverages the synergistic strengths of human and machine intelligence: human creativity, emotional resonance and cultural insight are provided by humans, while technical implementations are executed by AI programs, generating permutations of established and proven patterns and identifying opportunities for optimization (this is also the case of a popular videogame genre nowadays, the *rogue-like*, mostly featuring procedurally generated levels and dungeons, for instance). In this regard, far from compromising the role of human developers, the more sophisticated AI tools can amplify their creative capabilities by breaking technical constraints and performing mundane duties that might otherwise deflect/consume development energies. It is no wonder then that the difference between the original idea of computer science in

relation to artificial intelligence, meaning automating activities and making life easier for people, to a great extent becomes blurred here.

Along these lines, game developers are now increasingly using AI systems as a sort of creative collaborators that may suggest design possibilities they may not think of, or generate variations on established themes that human designers might not have considered (Wang, 2024). These AI systems of today can in fact propose unusual combinations of game mechanics, narrative elements or visual styles that may spark new creative directions, all the while remaining grounded in the fundamental vision established by human designers. This cooperative strategy views artificial intelligence as a means of enlarging the possibility space of game design instead of limiting it to predefined patterns. In general, the most effective applications in this sense let humans maintain control over the design process, but use AI to explore design options and opportunities more thoroughly than manual methods ever could.

Within such a context, the evolution of AI in game development challenges traditional notions of authorship and creativity, raising questions about the nature of artistic expression in digitally mediated environments (Yu et al., 2025). As artificial intelligence programs take on more sophisticated creative tasks, artists will need to sort through complex issues of credit, control over creativity and authentic expression of artistic vision. This issue is being raised not only within development teams, but also in how the games are being presented to players, with some studios explicitly highlighting AI contribution, while others integrating them seamlessly into human-driven creative processes. The ongoing debate over creativity in AI-assisted game development reflects broader societal consideration of the place of human and artificial intelligence in creative industries, and games represent a very public area in which these concerns can indeed be voiced and explored.

3. *The revolution in Non-Player Characters (NPCs)*

3.1. From scripted behaviors to adaptive intelligence

One of the most apparent manifestations of the advancements taken place in artificial intelligence in the gaming world lies in the development of significantly improved non-player characters (NPCs), with modern systems transforming what used to be static characters into dynamic ones capable of responding more sensibly to player actions (Kumbharkar, 2025). The traditional NPCs of the past used to work by following a somewhat rigid scripting that allowed for limited behavior options, so that experiences became predictable quickly and contrib-

uted to the loss of immersion by the player as soon as the pattern was set and understood. These non-player characters worked more effectively as a sort of ‘animated furniture’ than as credible denizens of game worlds, moving along predetermined paths and returning the same response irrespective of context or interaction history. Although this method was by all accounts computationally lightweight (something that indeed contributed to its success for a long time), it basically established a core constraint of player interaction with the virtual worlds that was mechanical and not ‘alive’.

Modern NPCs, however, now use advanced AI models that allow characters to learn reactive behavior from world states and internal state variables for goals, emotions and relations, as well as from player interactions (Kumbharkar, 2025). Such systems form the basis of characters that seem to possess autonomous motivation and personalities that evolve throughout the game, as opposed to being pure information sources or combat targets. Sophisticated NPCs can now recall player decisions, incorporate the interaction history into their dialogues and change their behavior to reflect the tendencies of the player, fostering stronger illusions of intelligence and independence. Such a development has basically changed NPCs from fixed elements or even nuisances into seeming friends or enemies with whom players can hope to have a satisfying relationship that may add to narrative immersion.

Technologies for more sophisticated NPCs are also moving forward with voice recognition, natural language processing and emotion models that can be included in a game to enhance more natural interactions between players and virtual characters (Kumbharkar, 2025).

These mechanisms don’t just analyze the content of player choices, but also interaction patterns, timing and behavioral consistency in order to create more complex responses. Rather than responding from pre-scripted dialogue trees, advanced NPCs can generate contextually appropriate responses considering relationship history, goals, and emotional state. The ensuing conversation is usually less mechanized and can be perceived as more similar to a natural conversation, allowing players to become more engaged with game stories and relationships with virtual characters that seem to possess an actual personality and memory.

3.2. Emotional connections with digital entities

The development of the NPC intelligence impacts the social bonds the players may establish with digital characters, by creating feelings of

attachment, concern and empathy blurring the line between human social relations (Kumbharkar, 2025). When NPCs are able to achieve more behavioral sophistication and autonomous behavior, the players can respond with a social and emotional investment, which sometimes may be externally perceived as ‘creepy’, similar to what they would show in the case of human real-life friendships. In this regard, there are instances where fans habitually attribute personality, motivation and emotional status to advanced NPCs, even though they (or at the very least, most of them) are intellectually aware that those NPCs are indeed fictional characters (Li et al., 2024). This anthropomorphization effect, even though it may still seem weird, is a fundamental human brain process that game developers are increasingly using to design engaging narrative experiences with AI-driven characters.

The possibility of using voice interaction increases even more the emotional investment between players and NPCs through the development of a certain personality and the apparent spontaneity of character interaction (Kumbharkar, 2025). Modern systems enable voice attribute modification based upon emotional status, relation parameters and narrative context, enabling more realistically-sounding communication, which in turn engenders a more ‘felt’ response from players. The integration of voice recognition, besides, changes this interaction even further by simulating an actual dialogue, with players talking rapidly to characters who in turn respond with contextual awareness. Such a technology reduces one of the greater barriers between character and player, enabling more immediate and emotionally engaging connections and facilitating the narrative input and the investment with the characters. That is why emotionally intelligent NPCs are starting to compel challenging inquiries of the interactions’ nature and ethics between humans and artificial agents (Habibi et al., 2022). With players developing stronger relationships with virtual characters over time, they may be more willing to have actual emotional reactions to character destinies, be invested in the characters’ happiness and make decisions in the game on emotional instead of strategic grounds. Game creators now deliberately take into consideration the psychological effect of such connections, understanding that the players’ investment in character relationships is a strong mechanism for narrative engagement, but also one that is accompanied by responsibility for the emotional effect of the relationships. That responsibility encompasses game endings, characters’ deaths, or relationship arcs that may cause intense emotional reaction in invested players. There are numerous occurrences well-known to the gaming community of the players’ emotional reaction reaching certain spikes, most of which related to the unexpected death and/or sacrifice of beloved characters (not always NPCs); the death of

Aerith Gainsborough in Final Fantasy VII (and its recent remake) is one of them, for example; but let's not forget a huge amount of the narrative and characters that can be found and interacted with in most of BioWare's modern games (especially in the Dragon Age and Mass Effect sagas).

3.3. The 'Unbounded' character simulation frontier

Recent advancements in generative AI have also enabled the creation of character life simulation systems that break past conventional limitations of pre-coded behavior, enabling emergent personality and unexpected plot twists even to their authors (Li et al., 2024). Systems like 'Unbounded' represent a fundamental shift from finite, rule-based character simulations to infinite games where character behaviors emerge from complex underlying models rather than explicit programming (Li et al., 2024). These systems utilize specialized language models to generate dynamic game mechanics, narratives and character interactions in real-time, creating experiences where each playthrough develops uniquely based on accumulated history and player choices. This approach treats character development as an emergent property of complex simulation rather than a predetermined narrative path.

Given all the above, some issues still need to be dealt with. For instance, the visual coherence of character appearance across different environments and contexts remains a significant technical challenge for unconstrained character simulations, requiring new techniques in order to provide consistent identity in conjunction with contextual responsiveness (Li et al., 2024). Recent developments in dynamic regional image prompt adapters for vision models try to overcome this limitation by providing consistent yet adaptive visual generation of characters within varied environments (Li et al., 2024). In this way, the models preserve critical visual character properties, while enabling suitable modification in relation to activity, emotional state and ambient context. The resulting visual continuity may therefore support player attachment to characters who remain recognizable while engaging in diverse activities and experiences throughout the simulation.

Unbounded character simulations point toward a future where the distinction between pre-authored narrative and emergent character development increasingly blurs, creating gaming experiences where players *guide* rather than *control* character lives that unfold with apparent autonomy (Li et al., 2024). This basically inverts the player's role from hero to active bystander, enabling surprise, delight and spontaneity that cannot be achieved with pre-scripted narrative. The

resulting game experience is meant to more accurately mirror human relationships in which influence (and not control) describes interactions, and results cannot be exactly forecasted. This is a big move toward genuinely open-ended game experiences in which artificial intelligence allows virtual characters to evolve in unexpected and captivating ways by simulating agency and personality instead of predetermined directions. In a sense, it is a way to help gamers realize one of the fundamental truths in life even in a virtual world: that very few things are actually under their control... This last statement paves the way to what will be discussed in the next section.

4. The cognitive and emotional impact of AI on gamers

4.1. Mental health applications and concerns

The relationship between gaming, artificial intelligence and mental health represents a complex landscape with both therapeutic potential and legitimate concerns that vary significantly across player populations and game types (Ferrari et al., 2022). Nowadays, research is increasingly recognizing that games – enhanced by AI – can have a valuable therapeutic function when appropriately designed and implemented, particularly for young people facing mental health challenges (Ferrari et al., 2022). A systematic scoping review identified 49 studies testing 32 digital games designed for mental health applications, ranging from general wellbeing promotion to specific interventions for certain conditions (Ferrari et al., 2022). These applications exploit the engaging nature of games combined with the adaptability of AI in order to create experiences that support psychological resilience, emotional regulation and social connection, while at the same time providing safe environments where it is possible to practice coping strategies accordingly.

The therapeutic applications of AI-enhanced games actually span multiple levels of intervention, ranging from prevention and education for generally healthy populations, up to more targeted treatments for diagnosed mental health conditions (Ferrari et al., 2022). Games designed for promoting general mental health typically focus on building emotional awareness, improving skills for stress management and helping people learn positive coping mechanisms via an engaging gameplay that may incorporate psychological principles. More specialized interventions address specific conditions like anxiety, depression or Post-Traumatic Stress Disorder (PTSD) by creating game environments that gradually expose players to challenging scenarios while

teaching evidence-based coping strategies. The integration of artificial intelligence in these applications allows for the dynamic adjustments of the difficulty, so that players may receive a personalized feedback and experience an adaptive content that should in principle fit with their individual needs and progress.

Having said that, despite growing evidence that supports the potential benefits of gaming, concerns persist nonetheless regarding problematic play patterns and the potential for AI-enhanced games to exacerbate addiction vulnerabilities in particularly susceptible individuals (Formosa et al., 2022). The inclusion of Internet Gaming Disorder in the *Diagnostic and Statistical Manual of Mental Illnesses* (DSM-5)¹, and of Gaming Disorder in the International Classification of Diseases (ICD)² reflects the willingness of psychological experts to recognize potentially problematic relationships with gaming that may require clinical attention (Formosa et al., 2022). However, researchers also warn that these classifications may risk pathologizing healthy engagement and may miss underlying issues that catalyze problematic play patterns (Formosa et al., 2022). The challenge for game developers when integrating AI actually lies in creating engaging experiences that support wellbeing, while at the same time implementing ethical safeguards against manipulative design patterns that could exploit psychological vulnerabilities or encourage unhealthy engagement.

4.2. Emotion regulation and psychological support

Artificial intelligence, in general, may provide significantly promising opportunities for the support of emotion regulation in gaming experience, addressing a critical dimension of psychological well-being that is highly applicable to stressful gaming situations (Habibi et al., 2022). As a matter of fact, videogames commonly expose players to failure experiences that demand emotional resilience and adaptive coping to successfully overcome. AI tools have the capacity to identify frustration, anxiety, or disengagement cues via behavioral patterns and physiological signals and intervene early with suitable support mechanisms (Habibi et al., 2022). Such interventions can range from difficulty modification, encouragement, prompts for different strategies, or even short mindfulness activities that enable players to manage emotional reactions and stay focused without becoming frustrated.

The integration of AI chatbots within gaming communities and

¹ The American Psychiatric Association's professional reference book on mental health and brain-related conditions; it is the main guide for mental health providers in the U.S., with its latest version, called DSM-5-TR, published in 2022.

² See <https://www.who.int/standards/classifications/classification-of-diseases>.

platforms provides additional avenues for psychological support, creating spaces where players can process emotions, receive guidance and develop healthier relationships with gaming experiences (Lopes et al., 2024). Initial research with consumers of generative AI chatbots for mental health treatment describes positive outcomes, such as improved relationships and recovery from trauma and loss (Siddals et al., 2024). These interactions create what researchers have determined is an ‘emotional sanctuary’ whereby users receive sage advice particularly about relationships and experience pleasure of connection in a safe, non-judgmental space (Siddals et al., 2024). For gamers facing social anxiety, performance pressure or toxic interactions, these AI companions can offer valuable support that may complement human connections.

The personalization capabilities of modern AI systems may also enable tailored emotional support that recognizes individual differences in psychological needs, coping strategies and communication preferences (Lopes et al., 2024). Rather than applying one-size-fits-all solutions to player assistance, adaptive systems can identify what specific players find comforting, encouraging, or enabling based on their interaction patterns and expressed preferences. This is taken to the level of communication style, with some enjoying direct, solution-focused guidance and others responding to a sort of an empathetic recognition of the state of mind. By adapting support strategies to the distinct needs of each individual, AI systems can better help gamers develop healthy emotional regulation skills that can be transferred from gaming contexts to broader areas in real life.

4.3. Flow, engagement and identity formation

The ‘flow’ theory in psychology (a moment of optimal engagement in which challenge and skill are in balance) is one of the fundamental game design principles rendered increasingly feasible through dynamic difficulty adjustment and adaptive content delivery (Zhu et al., 2021). Traditional games, in fact, struggled with the constraint of being limited in their ability to maintain flow states, due to the often rigid difficulty curves that tended to become too hard to handle for some and too easy to handle for others. AI solutions try and bypass this restriction through real-time monitoring of the player and adjusting parameters to maintain optimal levels of challenge across the game (Zhu et al., 2021). Such adjustments can include modifying the capabilities and attributes of the player characters, resource material availability, puzzle difficulty, or the pace of the narrative based on player habits individually, leading to a

more consistent flow experience and maximizing enjoyment and skill acquisition.

The relationship between players and artificially intelligent game elements may also introduce novel considerations for identity formation and self-perception, particularly among younger players that are still developing their sense of self (Zhu et al., 2021). As artificially intelligent characters and systems become better as interaction partners, they may affect the way players learn about themselves, their abilities and their social connections. Research indicates that player-AI interaction patterns can potentially move human-AI relations beyond productivity-focused frameworks to encompass playful exploration of identity through interaction with responsive digital beings (Zhu et al., 2021). These interactions may thus allow players to experiment with different self-presentations, communication styles and relationship dynamics to be explored in the safe environment of a virtual game, potentially supporting healthy identity development when appropriately implemented.

In addition, discovery-based learning through playful interaction with AI systems represents a valuable approach to understanding both technological capabilities and human psychology (Zhu et al., 2021). When the players are invited to play with the AI systems in the absence of pre-established objectives or rigid structures, they may naturally challenge the technology and their own reactions to its boundaries, abilities and tendencies of behavior. The game designers can support this discovery process by designing learning curves that add complexity progressively, while at the same time preserving control for the player, inviting experimentation with various forms of interaction and watching the effects (Zhu et al., 2021). This ludic interaction with AI offers metacognitive growth as gamers introspect on their own thought, emotional reaction and interactive behavior through the *mirror* of artificial intelligence.

5. *AI assistance and accessibility in gaming*

5.1. Reducing barriers to gaming participation

Artificial intelligence technologies tend to increasingly remove conventional obstacles to gaming participation, making games more accessible for gamers of varying ability, background and degree of gaming experience (Ward et al., 2024). In fact, adaptive difficulty systems represent one of the most significant innovations in accessibility, automatically adjusting challenge levels based on individual player performance to prevent both frustration from excessive difficulty and

boredom from insufficient challenge. These systems usually analyze patterns in player actions, success rates and engagement signals in order to create personalized difficulty curves that maintain an appropriate challenge regardless of the starting skill level. This adaptation enables players with different physical capabilities, processing style or prior gaming experience to play the same games without an explicit difficulty selection that might stigmatize or exclude the players who are more unsure of or less confident in their abilities.

Language barriers that previously limited gaming participation are also increasingly overcome by AI-powered translation and language learning incorporation that may support global participation in gaming communities (Ward et al., 2024). Real-time text and voice chat translation is now enabling, with a far higher degree of confidence than was possible before, cross-language communication among players, and contextualized language learning features allow players to also acquire new language skills through gameplay. Research on Computer-Assisted Language Learning (CALL) enhanced by AI demonstrates that the integration of natural language processing, AI image generation and text-to-speech technologies creates more engaging and inclusive language learning experiences within gaming contexts (Ward et al., 2024). These innovations are especially useful for language learners and non-native speakers, who might otherwise be excluded from multiplayer game spaces or miss out on crucial narrative content in single-player games. The problem (or opportunity?) of needing to learn English or Japanese in order to play a certain game in the 1990s or 2000s is now a thing of the past.

Physical accessibility barriers can also be addressed to a certain extent by AI systems that monitor player input patterns and modify control schemes to suit a range of physical abilities (Ward et al., 2024). These systems are able to recognize when players are struggling with specific combinations of controls, and in such a case they are able to perform timing demands on a regular basis and automatically tweak parameters to suit their abilities. Furthermore, more sophisticated implementations of such systems can accept other forms of input, such as voice commands and eye movements, or a reduced control schemes, which can preserve fundamental gameplay experiences while accounting for different physical abilities of the players; this is meant to prevent their need for explicitly stating their disabilities or going through often lengthy and tedious manual configurations.

5.2. Personalized learning through gameplay

The integration of artificial intelligence into computer games is now offering even more opportunities for personalized learning experiences

attuned to the cognitive style of each player, his or her own interests and learning requirements (Lakshmikantha, 2024). AI-powered game-based learning can make learning content dynamically adjust to the students' progress, by differentiating the difficulty, providing personalized feedback and emphasizing concepts requiring additional practice based on performance trends. Learning based on AI-driven games in professional training environments, such as orthodontic dental education, was found to bring a substantial improvement in student motivation, knowledge retention and theory-to-practice application (Lakshmikantha, 2024). These benefits stem from AI's ability to create immersive simulations that replicate real-world practice, while at the same time providing immediate, personalized feedback that traditional educational approaches usually are not able to match.

Virtual reality, augmented reality, 3D modeling and advanced e-learning platforms enhanced by artificial intelligence also create multimodal learning experiences that accommodate diverse learning preferences and cognitive processing styles (Lakshmikantha, 2024). Visual learners may benefit from sophisticated graphical representations that illustrate complex concepts; auditory learners may receive explanations optimized for their processing style; and kinesthetic learners may engage through interactive simulations that can create a far better understanding. AI systems are able to track which modalities can prove most effective for individual learners and emphasize those approaches while gradually strengthening less preferred learning channels through appropriate scaffolding. This multimodal approach is particularly beneficial for complex ability development that requires integration of conceptual understanding with procedural knowledge and situation awareness.

Besides, the personalization capabilities of AI-enhanced learning games can go beyond cognitive adaptation, since it can also encompass affective engagement mechanisms that may detect and react to motivation, frustration, and gratification of the students (Lakshmikantha, 2024). These capabilities can make learning games in a sense 'aware' of the indicators of disengagement or affective obstacles to learning and apply suitable interventions, for example by providing new challenges or more encouragement to the players, or by lowering the level of difficulty temporarily in order to restore confidence as well. Research attests that this kind of affective computing approach significantly enhances learning outcomes through addressing emotional dimensions of the learning process that are typically overlooked in traditional approaches to learning (Lakshmikantha, 2024). For skills requiring persistence through challenging learning curves, such as language acquisition or complex problem-solving, these emotional

support mechanisms may prove particularly valuable in maintaining engagement and preventing premature drop-out of the learners.

6. *The social dimensions of AI in gaming*

6.1. Shifting boundaries between human and AI interaction

The incorporation of advanced AI in game worlds, specifically the multiplayer ones, is also somehow inevitably transforming social relations through blurring the traditional distinctions between human and artificial players within those environments (Zhu et al., 2021). Contemporary games, in fact, increasingly include hybrid worlds in which human players can engage with AI-driven characters at different levels of complexity, ranging from scripted simple NPCs to sophisticated agents utilizing deep learning mechanisms, as discussed in Section 3. This integration can create novel social contexts where players must navigate relationships that span the spectrum between clearly artificial and authentically human, with many interactions falling into a sort of an ambiguous territory. Research indicates that player-AI interaction patterns can expand current notions of human-AI relationships beyond productivity-based paradigms, creating new social dynamics that combine elements of tool use, companionship, competition and collaboration in unique configurations (Zhu et al., 2021).

Because of this, the ambiguity between human and artificial intelligence players in game environments comes with possibilities as well as dilemmas for social interaction, calling for the necessity of players learning new competencies in interaction and social awareness (Zhu et al., 2021). Players are nowadays more and more faced with not being able to tell right away if they are dealing with artificial or human entities, as AI's natural language generation and contextual understanding capabilities grow. This uncertainty can sometimes lead to what scholars call "AI social uncanniness" – a sort of 'unease' from engaging with agents that possess some but not all traits of human social behavior (Zhu et al., 2021). Game developers now need to navigate cautiously around this phenomenon, with the purpose of creating AI-based elements to enhance and not get in the way of social interaction. This can be done, in principle, on one hand by providing a due transparency about artificial entities, while on the other hand by leveraging their capabilities to enhance social interactions.

Advanced AI programs in multiplayer environments can also act as social facilitators rather than as enemies or allies, bridging communication gaps, mediating conflicts and enabling more constructive

community interactions (Zhu et al., 2021). These systems can identify toxic communication patterns, suggest constructive alternatives and reward positive social behaviors through both explicit incentives and subtle environmental adjustments. Some games now implement AI community managers that monitor player interactions, identify emerging social problems and implement appropriate interventions before the conflicts have the chance to escalate. These applications demonstrate the potential for AI to augment rather than replace human relationships by creating social environments that facilitate collaboration, support and constructive communication, and prevent destructive patterns of interaction that might otherwise spread and dominate (sometimes, uncontrollably) online gaming communities.

6.2. Community building around AI innovations

Gaming communities increasingly form around shared interest in AI innovations, creating social ties through mutual engagement with artificial intelligence possibilities in game worlds (Li et al., 2024). Gamers use to gather on online forums, social media and specialized communities to report outcomes, strategies and innovative uses of AI algorithms in their preferred games. In general, they combine technical interest with ludic experimentation, forming communities in which members collectively explore the possibilities of what game AI systems can do. Their social dynamics are generally focused on knowledge exchange, collaborative creativity and cooperation (and not just competition), forming – in principle – constructive social spaces that can lead to learning and experimentation. This building of communities is a special social phenomenon in which technology is not just a mere tool, but it is also a shared interest that can create strong human connections.

The collective exploration of generative AI capability within gaming spaces also illustrates how technology can enable human connection rather than preclude it via shared creativity and discovery (Li et al., 2024). Communities of users that develop around such games as ‘Unbounded’, as discussed in Section 3, participate in collective experimentation with the offer of character life simulation, trading stories, strategies and unexpected consequences that emerge from their interaction with generative AI technologies (Li et al., 2024). Such experiences create rich social environments where players exchange philosophical questions regarding artificial intelligence, consciousness and character nature with practical tips for achieving the desired outcomes.

Finally, the interaction between AI developers and game commun-

ities is more and more characterized by a two-way communication, where player feedback directly contributes to technological innovation (Zheng, 2025). Game communities are effective laboratories for AI technologies because players happen to discover new applications, constraints and creative possibilities that formal testing might not. The developers actively interact with community threads, gameplay patterns and direct feedback in order to identify areas of improvement and new applications of AI technologies. This symbiosis is converting some gamers from passive users into active participants, fostering more intimate associations between gaming communities and development teams via a mutual investment in AI feature development. As a result, the most powerful AI applications in game worlds arise most frequently from this back-and-forth conversation between player communities and developers, mixing technical expertise with the multi-faceted player insight.

7. Future horizons: the merging of human and machine play

7.1. Emerging technologies and approaches

The future of AI in games is leaning towards increasingly frictionless integration of generative capability across game development and gameplay process, with emerging technologies promising fundamental shifts in game creation and consumption (Yu et al., 2025). Generative game engines are the most promising of the frontier technologies, with interactive generative video technology being utilized to create dynamic game worlds that evolve and grow in response to player movement (Yu et al., 2025). These systems can combine world modeling, user-controlled interactivity, long-term memory capabilities and causal reasoning in order to generate coherent game experiences without traditional hand-crafted assets and scripting (Yu et al., 2025). The resulting approach fundamentally reshapes how games are created, shifting development from asset production and explicit programming toward prompt engineering and system design that can guide generative processes toward the desired gameplay experiences.

In this regard, cross-game knowledge transfer allows AI systems to take advantage of learning in a single gaming environment and transfer it to others, solving a key limitation of existing methods that must train separately on each gaming environment (Zheng, 2025). Cross-game knowledge transfer allows AI systems to build more generalizable intelligence that is capable of identifying similarities across game genres and mechanics, in the same way that human players generalize skills

across games. Recent studies have shown remarkable gains in efficiency using these transfer learning methods, enabling the AI to successfully handle novel situations with reduced reliance on training and hardware (Zheng, 2025). As these capabilities further develop, they promise more flexible AI collaborators and opponents that can engage with players throughout all of their games rather than within specific titles or genres, developing more consistent and personalized experiences across game ecosystems.

The integration of multimodal AI that handles visual, audio, text and interactive data streams simultaneously enables more sophisticated understanding of gameplay contexts and player intent (Li et al., 2024). They incorporate computer vision, natural language comprehension, audio analysis and behavioral pattern recognition to synthesize more integrated models of game state and player experience. The AI that is generated in this way is more able to react to sophisticated situations by taking multiple channels of information into account at a particular time, more similar to human perceptual and decision-making processes. This multimodal interaction is especially useful for authentic NPC creation, real-time difficulty adaptation and tailored content creation that is responsive not only to overt player actions but also to indirect hints via all modalities of communication.

7.2. Ethical considerations and human-centered design

As artificial intelligence capabilities in gaming continue to develop, ethical implications regarding data privacy, psychological impact and environmental sustainability become paramount as essential considerations for ethical development (Zheng, 2025). The extensive data collection required for personalized gaming experiences rightfully invokes concerns regarding privacy, particularly for sensitive information that may potentially be harvested from gaming play patterns, such as cognitive processing habits, emotional responses and decision-making approaches. Developers must navigate difficult questions of responsible use of data, transparency, storage limits and player consent as they look to create AI systems that can learn from player behavior. These questions extend beyond legal compliance concerns to fundamental issues of respecting player autonomy and maintaining appropriate boundaries between personalization and surveillance.

It is therefore clear that the psychological impact of more advanced AI in videogames needs to be thoroughly studied, specifically regarding (para)social interaction with virtual characters, the possibility of controlling gameplay behavior and its effect on self-development and social development (Habibi et al., 2022; Formosa et al., 2022). As

players become increasingly emotionally invested in AI characters and systems, developers are, as a direct consequence, increasingly responsible for the psychological influence of such relationships, their establishment, creation and possible dissolution through game updates or cancellation. Ethical integration involves responsible design which honors players' emotions without using exploitative methods that may take advantage of psychological weaknesses for engagement or profit. This is especially true for games aimed at younger gamers or games that are meant for long-term play over the course of months or years. Unfortunately, a significantly high number of free-to-play games, especially those aimed at the mobile market, have used and are still resorting to this kind of manipulative tactics for reasons only related to sheer profit.

Similarly, the environmental impact and therefore the ecological footprint of advanced AI in videogames presents increasing concerns that need sustainable solutions to development and implementation (Zheng, 2025). The vast amounts of energy that are invested in training complex neural networks (estimates talk about 18.7 MWh per session, for instance, to train advanced game AI) rightly call into question the environmental accountability of technological innovation (Zheng, 2025). Addressing these challenges may require a number of strategies, including more effective energy consumption algorithms, transfer learning to reduce redundant training and open disclosure of environmental footprints. Some more 'eco-friendly' developers are already integrating environmental considerations into their AI development pipelines in an effort to balance technological advancement with responsible utilization of resources, by using techniques like model compression, optimized inference and shared pre-training that may reduce the overall computational footprint, but of course there is still a long way to go in this direction.

7.3. Maintaining humanity in an AI-enhanced gaming future

The most compelling future for AI in games is one that is focused on augmentation, not replacement of uniquely human aspects of play, including creativity, emotional involvement and rich social connection (Hossain & Habib, 2024). It is one that approaches artificial intelligence as a means of amplifying human creative and social potential rather than as a replacement for human effort in game development and gameplay. Most promising deployments build player agency through responsiveness and potential without eliminating room for human imagination, emotional expression and social interaction that are at the heart of deep play experiences. This philosophy directs development

towards frictionless technology that enhances possibilities and serves human relationships instead of imposing technological presence that could subtract from the inherently human spirit of play.

The idea of ‘AI as play’ instead of just sheer productivity provides a useful way of considering how artificial intelligence can add to and not take away from gaming experiences (Zhu et al., 2021). This view acknowledges that playing with AI systems in a creative way (with their possibilities, constraints and quirks as well) is a goal in itself, regardless of an instrumental relationship to technology. Game designers should in principle be able to facilitate this playful exploration by creating appropriate learning curves for AI interaction, incorporating discovery-based learning that might encourage experimentation and extending explicit invitations to play with rather than merely use AI systems (Zhu et al., 2021). This approach is meant to treat the relationship between player and artificial intelligence as a form of play in itself, creating meta-game experiences where exploration of the technology becomes part of the enjoyment.

Given all of this, the future of human-AI collaboration in gaming will very likely emphasize complementary contributions that leverage the distinctive strengths of both human and artificial intelligence (Hossain and Habib, 2024). Human creativity, emotional understanding, cultural awareness, ethical judgment and social intelligence may clearly (and should!) complement AI capabilities in activities like pattern recognition, procedural generation, data processing and consistent execution, for instance. This symbiotic collaboration might then allow for co-creative processes where humans provide direction, meaning and creative vision, whereas AI systems take care of implementation details, generation of variations and management of technical complexities. The resulting collaboration might generate possibilities beyond the capability of humans or AI individually, pushing forward the creative frontier of gaming while ensuring meaningful human authorship and agency. This approach might strike a balance between the two realities (human and artificial) and might aim towards a future where artificial intelligence enhances rather than replaces the fundamentally human experience of play.

Conclusion

In this chapter, the discussion focused on underlining how the integration of artificial intelligence into the gaming world has produced significant transformations in both how games are created and how they

are experienced, with significant implications for developers, players and the gaming medium alike.

From the perspective of the developers, AI has evolved from simple rule-based systems to sophisticated learning algorithms that are able to generate content, optimize the games' performance and create increasingly believable virtual characters. These technological improvements, nowadays, already allow smaller teams to develop far larger experiences than they could before, while at the same time removing repetitive tasks and making it so that developers may focus their creative energy on the inventive elements that result in the creation of interesting, memorable game characters and worlds. The consequent evolution is not just an incremental improvement in the technological tools, but also represents a dramatic change in the relationship among creators, technology and the creative process as a whole.

For the players, the application of artificial intelligence goes much further than having tougher enemies to be faced or more realistic non-player characters whom to interact with; it also enables the games they play to include personalized experiences that tailor themselves to the individual tastes, skills and emotional states of the players. As a matter of fact, AI-powered gaming can become a space for personal growth, emotional management and social bonding that can dynamically adapt to the players' needs and skill levels. These adaptations, therefore, consequently create more open, active and significant experiences that most of the time succeed in addressing the wide range of motivations, abilities and play style of the players. The most successful examples in this regard are actually able to preserve the players' interactive power while offering suitable support to them, as well as to provide experiences that may be somehow sensitive to individual requirements without being perceived as prescriptive or manipulative.

For the entire gaming medium, artificial intelligence, just like everything that relates to the future and therefore requires predictions that can over time be proven wrong, as of this moment seems to be walking a critical path. This path is leading to emphasizing complementary relationships between the creativity of humans and the capabilities of the artificial intelligence itself, by taking advantage of the distinctive strengths of each in order to create experiences beyond what either of them could achieve independently without the other. Succeeding in finding a suitable balance like this, however, may require thoughtful, human-centered approaches to development that may view AI as a tool for expanding the creative and social possibilities of us humans rather than as a replacement for human involvement. If this focus can be kept, the gaming world will probably be able to exploit the revolutionary power of artificial intelligence without sacrificing the

inherently human aspects of play, creativity and interaction that make gaming worthwhile in the long term.

The future that is unfolding before us holds out the possibility for having games that may combine technological innovations, like those from AI, with real human creativity, with the ultimate purpose of bringing about worlds that may be even more fantastical, but at the same time even more *human*, than ever before.

References

- FERRARI M., SABETTI J., McILWAINE S.V., FAZELI S., SADATI S.M.H., SHAH J.L., ARCHIE S., BOYDELL K.M., LAL S., HENDERSON J., ALVAREZ-JIMENEZ M., ANDERSSON N., NIELSEN R.K.L., REYNOLDS J.A., IYER S.N. (2022). *Gaming My Way to Recovery: A Systematic Scoping Review of Digital Game Interventions for Young People's Mental Health Treatment and Promotion*, in "Frontiers in Digital Health", 4, 814248. <https://doi.org/10.3389/fdgth.2022.814248>
- FORMOSA J., JOHNSON D., TÜRKAY S., MANDRYK R.L. (2022). *Need satisfaction, passion and wellbeing effects of videogame play prior to and during the COVID-19 pandemic*, in "Computers in Human Behavior", 131, 107232. <https://doi.org/10.1016/j.chb.2022.107232>
- HABIBI R., PFAU J., HOLMES J., EL-NASR M.S. (2022). *Empathetic AI for Empowering Resilience in Games*, in *Proceedings of the AIIDE Workshop on Experimental AI in Games*.
- HOSSAIN M.B., HABIB F. (2024). *AI: Your Best Friend, Worst Enemy and Ultimate Game-Changer*, in "Journal of Next-Generation Research 5.0". <https://doi.org/10.70792/jngr5.0.v1i1.57>
- KUMBHARKAR H. (2025). *Emergent Minds: The Evolution and Future of AI in Video Game NPCs*, in "International Journal For Multidisciplinary Research", 7(3), 42387. <https://doi.org/10.36948/ijfmr.2025.v07i03.42387>
- LAKSHMIKANTHA H.T. (2024). *Leveling up orthodontic education: Game-based learning as the new frontier in orthodontic curriculum*, in "Research and Development in Medical Education", 13, 13. <https://doi.org/10.34172/rdme.33205>
- LI J., LI Y., WADHWA N., PRITCH Y., JACOBS D.E., RUBINSTEIN M., BANSAL M., RUIZ N. (2024). *Unbounded: A Generative Infinite Game of Character Life Simulation*, in "arXiv", <https://doi.org/10.48550/ARXIV.2410.18975>
- LOPES R.M., SILVA A.F., RODRIGUES A.C.A., MELO V. (2024). *Chatbots for Well-Being: Exploring the Impact of Artificial Intelligence on Mood Enhancement and Mental Health*, in "European Psychiatry", 67(S1), S550–S551. <https://doi.org/10.1192/j.eurpsy.2024.1143>
- RISI S., PREUSS M. (2020). *From Chess and Atari to StarCraft and Beyond: How Game AI is Driving the World of AI*, in "KI - Künstliche Intelligenz", 34(1), 7-17. <https://doi.org/10.1007/s13218-020-00647-w>

RUTHVIKA M. (2025). *AI-Powered Data Analytics: A Game Changer*, in “International Journal of Advanced Research in Science, Communication and Technology”, pp. 97-102. <https://doi.org/10.48175/IJARSCT-24409>

SIDDALS S., TOROUS J., COXON A. (2024). “*It happened to be the perfect thing*”: experiences of generative AI chatbots for mental health, in “Npj Mental Health Research”, 3(1), 48. <https://doi.org/10.1038/s44184-024-00097-4>

WANG P. (2024). *Application and Future Development of AI in Early Game Production and Interactive Content*, in “Applied and Computational Engineering”, 120(1), pp. 113-119. <https://doi.org/10.54254/2755-2721/2025.18748>

WARD M., XU L., UÍ DHONNCHADHA E. (2024, September 10). *Enhancing human-centric CALL through AI innovations*, in *CALL for Humanity - EuroCALL 2024 Short Papers*. EuroCALL 2024: CALL for humanity. <https://doi.org/10.4995/EuroCALL2024.2024.19067>

YU J., QIN Y., CHE H., LIU Q., WANG X., WAN P., ZHANG D., LIU X. (2025). *Position: Interactive Generative Video as Next-Generation Game Engine* (No. arXiv:2503.17359), in “arXiv”. <https://doi.org/10.48550/arXiv.2503.17359>

ZHENG Q. (2025). *The Evolution and Optimization of Game AI: From Rule-Driven to Deep Reinforcement Learning*, in “Applied and Computational Engineering”, 150(1), pp. 77-82. <https://doi.org/10.54254/2755-2721/2025.22245>

ZHU J., VILLAREALE J., JAVVAJI N., RISI S., LÖWE M., WEIGELT R., HARTEVELD C. (2021). *Player-AI Interaction: What Neural Network Games Reveal About AI as Play*, in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1-17. <https://doi.org/10.1145/3411764.3445307>

OPEN QUESTIONS

AI, ‘Virtual’ Brains, and the Matter of Well-Being

Michela Balconi

1. Redefining well-being in the digital age: the impact of digitalized social relationships on human flourishing

The emergence of digital technologies has transformed the landscape of social interactions, leading to a reconfiguration of the concept of *well-being*. As face-to-face relationships are increasingly complemented or replaced by digital communication, it becomes necessary to examine the implications for psychological, emotional, and social dimensions of human flourishing. This chapter explores how digitalization affects well-being, drawing from interdisciplinary perspectives including psychology, sociology, and digital media studies. Special attention will be given to neuroscientific perspectives, with particular reference to the most recent scientific research in the field.

1.1. About well-being

Well-being has traditionally been conceptualized as a multidimensional construct encompassing emotional satisfaction, psychological functioning, and the quality of social relationships (Ryff and Keyes, 1995).

Well-being is a *multifaceted construct* encompassing physical, psychological, and social dimensions that together contribute to an individual's optimal functioning and quality of life. Recent neuroscientific approaches elucidate the critical role of interpersonal processes in shaping well-being from a neuroscientific standpoint. This synthesis will focus on the *interpersonal aspects of well-being*, highlighting how neural mechanisms underpin social connectedness, empathy, and relational attunement, which are essential for sustaining mental health and overall well-being.

Traditionally, well-being has been conceptualized primarily at the individual level, encompassing subjective experiences such as happiness, life satisfaction, and the absence of distress. However, well-being cannot be fully understood without considering the relational context within which individuals exist. Interpersonal interactions are not merely external influences but integral components of neural and emotional

regulation. Social connectedness functions as a fundamental human need, deeply embedded in brain architecture, and directly impacts physiological and psychological health.

Integrating findings from *neuroimaging*, *electrophysiology* (EEG), and *autonomic nervous system* studies to explain how the brain supports and is shaped by social interactions. Central to this framework is the concept of embodied cognition and sensorimotor simulation, whereby observing or engaging in social exchanges activates neural networks that mirror others' actions and emotions. This mirroring fosters empathy, emotional resonance, and interpersonal understanding – core ingredients for relational well-being (Balconi, 2025).

From a neuroanatomical perspective, cortical regions such as the prefrontal cortex, insula, and mirror neuron systems facilitate the processing of social cues and emotional states. Their dynamic activity enables individuals to attune to others' needs, regulate emotional responses, and co-construct shared meaning in social contexts. Disruptions in these neural circuits – due to stress, isolation, or digital mediation – can undermine interpersonal well-being by impairing empathy and social connectedness.

Specifically, *empathy*, as explored extensively and it was underlined that it is not a singular capacity but a multifaceted process involving affective sharing, cognitive perspective-taking, and regulatory control. Neuroscientific studies show that effective empathic engagement relies on coordinated neural and physiological responses between individuals. *Hyperscanning paradigms* used in recent research reveal inter-brain synchrony during cooperative and emotionally salient interactions, suggesting that shared neural rhythms underpin successful social bonding.

Such *interpersonal synchrony* facilitates trust, prosocial behavior, and emotional support, which in turn promote resilience and subjective well-being. Conversely, decreased neural coherence – observed in digitally mediated or culturally unfamiliar interactions – may impair relational quality and contribute to feelings of loneliness or alienation.

1.2. Well-being from an interpersonal neuroscientific perspective beyond the physical 'presence'

In this perspective, well-being analysis has particular relevance in the context of increasing *digitalization of social relationships*. It was shown how digital platforms, while offering unprecedented connectivity, may lack the full spectrum of embodied cues necessary for deep interpersonal attunement. The reduced availability of nonverbal signals and

sensorimotor feedback can limit the activation of neural empathy circuits, potentially diminishing relational well-being.

Thus, recent research stresses the importance of designing digital environments that replicate embodied social cues and foster emotional presence. Integrating neurophysiological markers such as EEG and autonomic responses can guide the development of platforms that better support interpersonal connection and mitigate risks of digital isolation. In summary, a neuroscientific perspective advances our understanding of well-being by foregrounding its interpersonal dimension.

Thus, well-being emerges not only from individual cognitive-emotional states but from *dynamic, embodied social exchanges* supported by complex neural networks. Empathy, neural synchrony, and embodied cognition form the neurobiological foundation of these constructs.

Therefore, in contemporary society, digital technologies mediate an increasing proportion of social interactions, prompting researchers to investigate the ramifications for individual and collective well-being (Przybylski and Weinstein, 2016). The digitalization of relationships raises crucial questions: does virtual connectivity enhance or hinder our sense of belonging and psychological resilience? How has the shift from physical to digital interactions redefined what it means to live a fulfilling life?

1.3. Digital communication and social connectivity

The widespread adoption of smartphones, social media, and messaging platforms has created new avenues for maintaining social ties. While these technologies can reduce geographical and temporal barriers, their influence on the quality of social interactions is complex. On one hand, digital tools can support social capital by facilitating frequent, albeit brief, interactions. On the other hand, the absence of nonverbal cues and the potential for miscommunication may undermine relational depth and empathy (Kuss and Griffiths, 2017).

Digital platforms offer convenience and immediacy, yet research shows mixed outcomes on perceived social support. For example, a longitudinal study by Shakya and Christakis (2017) found that increased Facebook use was associated with decreased well-being over time, suggesting that passive consumption of digital content may crowd out more meaningful social exchanges. Conversely, active engagement in online communities has been linked to increased perceived support and reduced feelings of isolation, particularly among marginalized groups (Naslund et al., 2016).

Digital interactions influence multiple dimensions of psychological well-being. According to Deci and Ryan's (2000) *Self-Determination Theory*, well-being is rooted in autonomy, competence, and relatedness. While digital tools can enhance autonomy and access to information, their effects on relatedness are more ambivalent. Studies reveal that while social media enables users to remain in contact with larger networks, it may simultaneously foster social comparison, envy, and decrease self-esteem (Vogel et al., 2014).

Moreover, the phenomenon of '*digital fatigue*' – characterized by emotional exhaustion and reduced cognitive functioning due to constant connectivity – raises concerns about the sustainability of digital engagement (Ayyagari et al., 2011). The pressure to maintain an online persona and respond promptly to messages can result in psychological strain, thereby detracting from subjective well-being.

1.4. Redefining well-being in a hybrid social ecology

The digital era necessitates a revision of traditional well-being models to incorporate hybrid forms of social interaction. As individuals navigate between online and offline relationships, well-being becomes increasingly contingent on digital literacy, boundary-setting, and the quality of mediated interactions. Researchers advocate for a more integrative framework that considers digital well-being as part of broader psychological resilience.

One emerging concept is '*digital flourishing*', which emphasizes intentional, value-driven use of technology to support psychological growth and authentic social connection (Villamil and Heshmati, 2023). This paradigm shift encourages a proactive approach to digital engagement, highlighting the importance of self-regulation, critical reflection, and purposeful interaction in cultivating well-being.

In this perspective, understanding well-being in the digital era has profound implications for education, mental health, and public policy. Interventions aimed at promoting digital literacy, emotional intelligence, and healthy screen habits can mitigate the negative impacts of digitalization. Moreover, platform design should prioritize user well-being by incorporating features that support mindful usage, such as screen time reminders and customizable notification settings (Kushlev et al., 2016). Future research should employ longitudinal and cross-cultural methodologies to assess the nuanced effects of digital relationships on different populations. It is also crucial to address digital inequality, as disparities in access and usage patterns may exacerbate existing health and social inequities.

To sum up, the digital transformation of social relationships presents

both opportunities and challenges for human well-being. While digital tools can enhance connectivity and access to support, they also introduce risks related to psychological strain and diminished relational quality. A comprehensive understanding of well-being in the digital age requires an *interdisciplinary approach* that integrates technological, psychological, and social perspectives. By embracing the principles of digital flourishing and intentional engagement, individuals and societies can navigate the digital landscape in ways that promote authentic connection and holistic well-being.

2. Digitalization, artificial intelligence, and health: balancing opportunity and risk in individual and public well-being

The convergence of artificial intelligence (AI), digital platforms, and health systems also represents a transformative shift in both individual care and public health management. From predictive diagnostics to AI-powered mental health apps, digitalization is redefining prevention, treatment, and long-term health planning. Feijóo (2020) identifies this shift as not only technological but also geopolitical, framing digital health as a field where diplomacy, regulation, and ethical standards must converge to ensure equitable well-being.

The COVID-19 pandemic accelerated this transformation, exposing vulnerabilities in mental health infrastructure and amplifying the need for scalable, remote solutions. Yet, as Bond (2025) and Ettman (2023) argue, the digital turn brings profound implications for psychological resilience, social connection, and cognitive-emotional development, dimensions critical to sustainable health.

Literature converges on a set of clear opportunities presented by AI and digitalization. Among the most impactful is the ability to personalize care. AI systems can process vast biomedical and behavioral datasets to detect patterns invisible to human clinicians, leading to early diagnosis and tailored interventions. This is especially promising in mental health, where passive data (e.g., voice tone, typing speed, geolocation) can signal depressive or anxious episodes.

In addition, digital health tools also expand access. Mobile health applications, telepsychiatry, and chatbot-based cognitive-behavioral therapy reduce geographic and economic barriers to care. These innovations align with Feijóo's (2020) call for '*health-tech diplomacy*', where equitable deployment and cross-border regulation ensure that digital health benefits are not limited to privileged populations.

AI further enables population-level prevention. By modeling health trends across demographics, it can inform public policy, anticipate

mental health crises, and allocate resources dynamically. Therefore, predictive analytics informed by social determinants of health (e.g., unemployment, social isolation) can guide targeted public mental health campaigns.

2.1. Risks: digital stress, isolation, and ethical blind spots

However, these promises coexist with *critical risks*. One major concern is the psychological toll of constant connectivity. Digital tools designed to monitor and optimize well-being may, paradoxically, contribute to cognitive overload and emotional fatigue. Ettman (2023) warns of the ‘*double bind*’ of digital health: the same technologies that enable remote care may deepen a sense of isolation, particularly when face-to-face support systems are eroded.

Bond (2025) extends this critique by noting that algorithmic mental health tools often lack cultural sensitivity and emotional nuance. While they can mimic empathy, they do not substitute for the *relational quality of human caregiving*. Prolonged engagement with these systems may desensitize individuals to authentic interpersonal exchange. It was further highlighted risk of surveillance, data misuse, and ethical opacity. Health platforms, if unregulated, may collect sensitive data without proper consent, exacerbating distrust and reinforcing structural inequities. The governance of AI in health thus requires transparent protocols, algorithmic accountability, and participatory design.

2.2. Balancing the equation: toward a sustainable digital health ethos

The digitalization of health is irreversible and, if guided wisely, profoundly beneficial. Yet its full potential can only be realized if technological progress is matched by *ethical foresight*, *psychological insight*, and *inclusive governance*. Recent perspectives provide complementary visions for such an integrated future: one where AI augments rather than replaces human care, where digital access is a right not a privilege, and where public health strategies account not only for survival but for flourishing. By embedding emotional well-being, relational integrity, and cognitive autonomy into the heart of digital health systems, we can move toward a model of care that is not only smart but truly human.

To reconcile opportunity and risk, a multidimensional approach is essential. First, ethical design must anchor all technological innovations. This includes privacy-by-design architecture, inclusive algorithms, and international standards for health-tech regulation. Second, *emotional and cognitive well-being* must be considered as central – not ancillary – metrics of success in digital health policy. This fact advocates for the integration of techno-psychological indicators into public health

monitoring, such as digital emotional well-being scores or relational engagement indices. These metrics can guide adaptive policy responses and illuminate hidden costs of digital interventions.

Third, hybrid care models should be prioritized. Ettman (2023) recommends systems that blend *digital scalability* with human contact points, such as therapist-supervised AI tools or community-based telehealth hubs. This approach maximizes the reach of AI without sacrificing relational richness.

Finally, *digital literacy* must be expanded at all levels – from individuals to institutions. Equipping users with the skills to navigate digital tools critically fosters autonomy, reduces overdependence, and promotes conscious interaction with health technologies.

3. *Redefining relational well-being through a neuroscientific lens: digital vs. face-to-face interactions*

Neuroscience has increasingly focused on the neural mechanisms underlying social interaction in digital contexts and how these differ from traditional face-to-face exchanges. This paragraph reviews contemporary neurocognitive findings on relation-based well-being, emphasizing electrophysiological and hyperscanning paradigms, emotion regulation markers, and social-cognitive network dynamics.

3.1. The social brain and communication modalities

Social-cognitive neuroscience identifies a network – including *mirror neuron systems, theory of mind regions, reward circuits, and emotional regulation pathways* – crucial to face-to-face interaction. Remote or digital communication (video chat, messaging) engages this network differently: some areas show reduced activation, particularly mirror neurons, while others (e.g., attentional systems) may engage more strongly (Doheny and Lighthall, 2023).

This suggests that the neural correlates of social perception, reward evaluation, and mentalizing are qualitatively altered in digital settings.

Specifically, in a *neuroergonomic discussion*, a recent study provides early electrophysiological evidence on remote learning contexts, emphasizing situational cognitive load differences between digital and in-person environments (Cassoli and Balconi, 2022). In related work, these authors employed *EEG hyperscanning* to compare digital vs. face-to-face job interviews, revealing that digital contexts reduce inter-brain connectivity, social synchrony, and reward-related activations – markers associated with relational well-being and social engagement (Balconi

and Cassioli, 2022). These *decrements in synchrony* may translate into diminished empathic attunement and emotional resonance, weakening the quality of relational bonds.

From an analogous perspective, a recent neurophysiological study by Myruski and colleagues (2024) assessed adolescents' preference for digital emotional communication and its association with anxiety. They measured *delta-beta coupling*, a neural signature of emotion-control interaction. Teens preferring digital modes exhibited elevated delta-beta coupling – a neural marker of rigid emotion regulation – which mediated increased *social anxiety symptoms*, suggesting that relying on digital interaction may correspond with less adaptive emotional processing and compromised well-being.

3.2. Comparative 'social connectedness'

Large-scale experimental studies report that face-to-face interactions elicit *stronger social connectedness* and *positive affect* compared to digital forms. Social closeness declines moving from in-person to video chat to audio to text messaging, with text-only methods generating the weakest relational bonds (Meier et al., 2021). This gradient mirrors neural findings: diminished social reward and mentalizing activations online correlate with weaker perceived support and reduced well-being.

A recent scoping review (Petersen et al., 2023) of ICT use among older adults demonstrates that technology may reduce loneliness and increase well-being – but crucially when digital engagement complements, not replaces, face-to-face relationships. Key moderating factors include digital self-efficacy, training, autonomy, and supportive design affordances. However, while digital interaction enables weak-tie expansion, it may undermine emotional trust and depth, especially when substituting for in-person contact.

To synthesize, neuroscientific evidence consistently indicates that digital communication engages the social brain differently from face-to-face interaction: lower neural synchrony, altered reward and mentalizing activations, and neural markers of emotional rigidity. Hyper-scanning studies reveal diminished inter-brain connectivity in remote educational or interview settings, directly linking to reduced relational engagement. Adolescent neuroimaging shows that preference for digital emotion sharing correlates with anxiety via emotion regulation inefficiency. Behavioral surveys confirm that subjective well-being benefits from digital contact are contingent on the preservation of in-person ties.

These findings suggest that *well-being in relational contexts* depends on the *quality* and *neural richness of interaction*. Digital forms may serve

functional convenience or supplemental support – but do not fully replicate the neuroemotional complexity of embodied social interaction. In this direction, neuroscience should continue longitudinal and context-sensitive investigations: for example, varied age groups, educational vs. personal contexts, and hybrid modalities. In educational settings, integrating features that enhance inter-brain synchrony (e.g., real-time gaze, co-presence cues) may mitigate digital deficits. Promoting digital literacy alongside emotional regulation skills is critical, especially for adolescents prone to digital-preference patterns linked to anxiety.

Neuroscience robustly illustrates that relational well-being rooted in face-to-face interaction differs at neural and psychological levels from digital communication. The digitized relational world poses both opportunities for broader connectivity and risks of diminished emotional attunement and neural resonance. Integrating neuro-empathic design, intentional hybrid modalities, and guided emotional support may help reimagine digital relational well-being without erasing the benefits of embodied social presence.

4. Digital presence and social interaction: challenges and opportunities

As we have underlined in previous paragraphs, the advent of digital technologies has profoundly reshaped the nature of social interaction. While offering unprecedented opportunities for connection, digitally mediated communication also presents significant challenges for social cognition, emotional engagement, and well-being. This paragraph explores the complexities of digital presence and social interaction, drawing on recent neuroscientific, psychological, and technological research. It addresses the dual role of technology as both a facilitator and potential disruptor of human connection, referencing recent key studies.

4.1. Mediated sociality in the digital era: the role of embodied cognition and empathy in digital spaces

In the contemporary landscape, digital media platforms have become essential tools for maintaining personal, professional, and educational relationships. However, the transition from face-to-face to screen-mediated interactions raises crucial questions about the *quality, authenticity, and emotional resonance of such communication*. While digital environments afford convenience and accessibility, they often lack the embodied, multisensory richness of in-person encounters, which are

fundamental to human empathy and connectedness (Macdonald et al., 2021).

Empathy and social understanding are deeply tied to embodied processes. Torbjørnsen (2025) highlights how *embodied cognition* – the interdependence of sensory, motor, and affective cues – is modified in online audiovisual spaces. Her research emphasizes that reduced access to nonverbal cues such as body posture, micro-expressions, and touch can impair empathic resonance, especially in group communication settings. Despite video platforms offering visual contact, the segmentation of perception and attention alters the natural flow of interpersonal feedback.

Balconi and colleagues (2023) corroborate these findings through neuroscientific data. Using EEG and hyperscanning methods, her team observed decreased neural synchrony and attenuated activation in *mirror neuron systems* during remote conversations compared to face-to-face interactions. These neural markers – essential for emotional mirroring and perspective-taking – suggest that digital presence, while cognitively efficient, may limit affective attunement.

Thus, the ‘*paradox of digital hyperconnectivity*’ is that it can foster both connectivity and alienation. Oluwatoyin (2021) warns that excessive digital engagement can lead to social fatigue, emotional disconnection, and decreased empathic capacity. Drawing from longitudinal data, she identifies a correlation between screen time, loneliness, and affective desensitization, especially in adolescents and young adults. As digital spaces replace embodied interaction, the sense of shared presence weakens, undermining the social scaffolding of well-being.

This phenomenon has been acutely observed among older populations as well. Indeed, it was underlined the irreplaceable value of face-to-face contact in maintaining cognitive vitality and emotional health in elderly individuals. Recent findings show that physical co-presence enhances memory retention, mood regulation, and neural plasticity – outcomes not fully replicable in digital formats, despite the potential benefits of video calls and messaging.

To address the emotional gaps in digital communication, Brakel and colleagues (2023) explore the use of immersive technologies such as *virtual reality* (VR) and *augmented reality* (AR) to enhance socio-emotional connectedness. Their studies indicate that when designed with embodied interaction in mind, immersive environments can partially restore emotional proximity by simulating co-presence and enhancing nonverbal expressiveness. Other recent research also engages with this frontier by applying hyperscanning techniques to immersive VR contexts. Preliminary results show increased inter-brain coherence and affective synchrony when participants interact in

embodied digital spaces compared to traditional video conferencing. These findings open promising avenues for designing emotionally resonant digital experiences that approximate the social richness of in-person encounters.

4.2. The digital presence for well-being

Thus, we can underline that, despite technological progress, no digital environment can fully substitute the nuanced interplay of face-to-face interaction. The future of mediated communication should aim not to replace but to complement physical sociality. As the research suggests, promoting digital literacy, empathy-centered design, and intentional use of technology can foster healthier social ecosystems.

A balanced approach also requires acknowledging *individual differences*: while immersive tools may enhance well-being for some users, others may require structured in-person contact for emotional security. Thus, *hybrid models* – integrating physical and digital modes of interaction – emerge as the most sustainable framework for nurturing relational well-being.

Digital presence is a *double-edged phenomenon*: it enables expanded networks and greater accessibility but can attenuate the embodied foundations of empathy and relational health. Neuroscientific and psychological research underscores the limitations of digital communication in replicating the social depth of face-to-face interaction. Nevertheless, innovative technologies such as VR and emotionally intelligent platforms offer hope for bridging affective gaps. By integrating scientific insight into technological development, we can foster digital environments that *support* – not replace – the human need for meaningful, embodied connection.

5. *Education and training in the age of AI: opportunities and risks for learning from a neuroscientific perspective*

AI is rapidly transforming the educational landscape, raising both enthusiasm and concern regarding its implications for cognitive development, emotional regulation, and learning outcomes. This paragraph explores the integration of *AI in education* from a neuroscientific standpoint, examining its potential to enhance learning through neurotechnologies and its possible impact on students' socio-emotional well-being. Drawing from contemporary neuroscience literature, it critically analyzes the use of educational AI and neurotechnologies (EEG, functional Near-Infrared Spectroscopy, fNIRS,

Brain Computer Interface, BCI), to discuss ethical, pedagogical, and developmental considerations.

5.1. The AI turns in education

AI applications in education – ranging from adaptive learning systems to intelligent tutoring and emotion-aware interfaces – promise unprecedented personalization of learning experiences. However, neuroscience warns that human cognition and emotional development *are not purely data-driven processes*. The growing field of educational neuroscience invites caution when implementing AI in learning environments, particularly where children's brains are still maturing.

According to Nouri (2025), neurotechnologies such as EEG, fNIRS and BCI are increasingly integrated into AI systems to monitor attention, cognitive load, and emotional states during learning tasks. These tools can support real-time adaptation of educational content, potentially optimizing the learner's cognitive state.

For example, EEG-based feedback can signal moments of mind-wandering or overload, prompting instructional changes, while fNIRS can track prefrontal cortex activity associated with effort and decision-making. Such closed-loop systems offer possibilities for enhancing metacognition, attention training, and self-regulated learning. However, it was also noted that sustained reliance on neurofeedback systems could externalize self-regulation processes and diminish intrinsic learning motivation.

The integration of AI into classrooms does not only affect cognitive performance, it also shapes the *social and emotional fabric of learning* environments. Recently, a comprehensive framework was provided for embedding digital well-being into educational and social design. Her recommendations include promoting emotional literacy, ethical awareness, and digital empathy as foundational competencies in AI-mediated contexts. From a neuroscientific perspective, socio-emotional development depends heavily on real-time interpersonal dynamics and embodied interaction, which may be diminished in AI-augmented settings. Studies using fNIRS and hyperscanning (Balconi and Angioletti, 2023) have shown that brain regions responsible for empathy, affective resonance, and social synchronization (e.g., anterior insula, medial prefrontal cortex) are less engaged in technology-mediated interactions, especially when human-to-human interaction is substituted rather than supplemented.

5.2. Self-regulation and the digital learning ecosystem

Roffarello and colleagues (2023) address the critical issue of digital self-control among students exposed to continuous stimuli from educa-

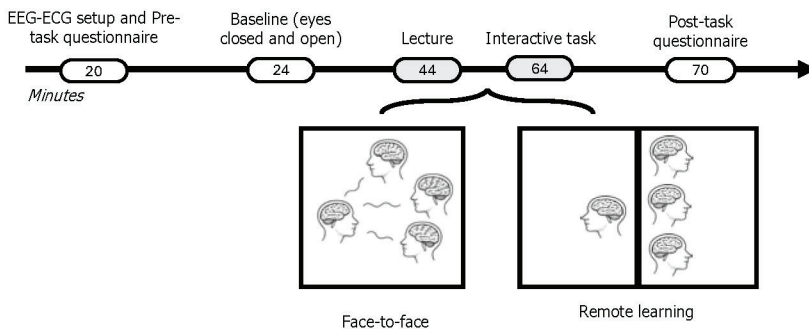
tional platforms. Their work highlights the importance of 'digital self-control tools' – including attention trackers, goal-setting dashboards, and screen time monitors – that encourage intentional, rather than passive, technology use. From a neurocognitive standpoint, these tools align with *executive function* development, particularly in areas such as the dorsolateral prefrontal cortex, which governs goal management and inhibition control. However, over-reliance on external digital regulators may interfere with the internalization of self-regulation strategies if not carefully scaffolded. This tension between external support and internal skill acquisition becomes especially relevant when considering how learning environments themselves influence the emergence of self-regulatory capacities.

In this regard, recent evidence by Balconi and colleagues (2023) offers an important perspective by examining how face-to-face and remote learning contexts modulate the neurophysiological dynamics underpinning collaborative engagement. Using a hyperscanning approach that simultaneously recorded EEG and electrocardiography (ECG) from groups of students and a lecturer, the study compared inter-individual synchronization during two instructional phases: a lecture and an interactive group task (Fig. 1). Results indicated systematically greater *neural convergence in face-to-face* settings, reflected in lower Euclidean Distance (EuDist) across all EEG frequency bands. The most pronounced effects emerged in the delta and theta ranges – frequencies associated with attentional stability, temporal integration, and motivational salience – particularly during interactive exchanges, where co-presence amplified alignment of cognitive and motivational processes. Conversely, remote learning conditions were marked by increased divergence, most evident in beta and gamma bands, which relate to semantic integration and executive processing, suggesting that the absence of embodied feedback may heighten variability in higher-order cognitive elaboration during dialogic tasks.

Complementary autonomic measures, derived from intersubject correlation of heart rate (ISC-HR), revealed greater cardiac synchronization among participants in face-to-face conditions, independent of the task type. This pattern indicates that physical co-presence inherently supports shared arousal and socio-emotional resonance, factors known to sustain motivation and cooperative engagement. When viewed in relation to digital self-control strategies, these findings suggest that while individualized tools such as attention trackers and goal-setting dashboards can enhance intentional engagement, their effectiveness in remote learning contexts may be constrained by the attenuated interpersonal synchrony characteristic of digitally mediated instruction. To counterbalance this limitation, remote educational designs may

benefit from incorporating synchronous interaction-rich formats and collaborative structures aimed at fostering collective engagement, thereby supporting both individual and shared dimensions of self-regulation.

Figure 1 - *Experimental setup illustrating hyperscanning of EEG and ECG during two instructional phases (lecture and group task) in face-to-face and remote learning contexts*



5.3. Ethical and developmental considerations

AI and neurotechnology raise significant ethical concerns when deployed in educational contexts. The collection of neurophysiological data from minors introduces questions about consent, data privacy, and psychological safety. Neuroscientific literature also emphasizes the *risks of premature optimization*: interventions based on neurodata may reinforce performance pressures and pathologize natural variability in learning pace and attention span.

Moreover, an overemphasis on quantifiable outcomes – such as brain activation levels or algorithmic scores – may neglect the holistic and humanistic goals of education. Neuroscience underscores the importance of multisensory learning, social interaction, and emotional engagement as core drivers of meaningful learning – elements that AI, as of now, only partially replicates.

While AI and neurotechnologies hold great potential for enhancing personalized learning and cognitive engagement, their deployment must be guided by a *nuanced understanding of brain development and emotional growth*. As recent studies suggest, a successful integration of AI in education should prioritize digital well-being, cultivate self-regulation, and maintain a central role for human interaction. Neuroscience

provides both the tools and the cautionary lens through which the future of AI-mediated learning should be envisioned.

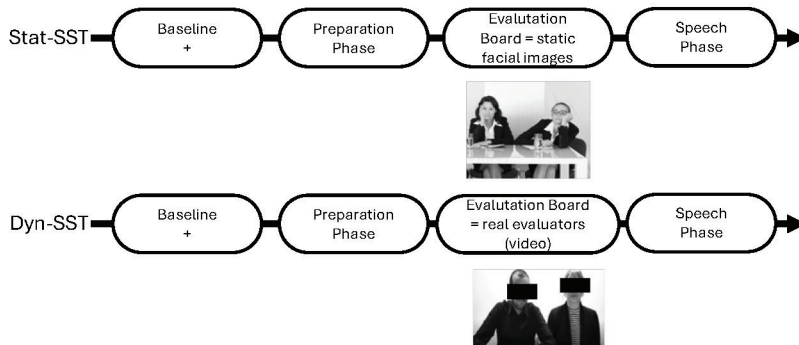
6. *Recent empirical evidence from 'hyperscanning neuroscience' on the topic of digitalization*

Recent research encompasses multi-modal neurophysiology (EEG, fNIRS, autonomic measures), *hyperscanning methods*, and real-world applied settings such as digital interviews and e-learning. A detailed discussion of these results may better elucidate the present state of the art and potentialities of the neuroscientific approach applied to AI and digital contexts.

6.1. Multimodal applications: fNIRS, EEG, and autonomic monitoring for cognitive and emotional functions

Balconi and colleagues investigated *stress responses* during digitally mediated job interviews (2024). Using EEG alongside cardiac and electrodermal recordings, they demonstrated that sensorimotor simulation in virtual settings triggers cognitive and emotional stress (Fig. 2). The findings reveal a mismatch between digital feedback and embodied responses, posing risks for longer-term psychological well-being when interacting in mediated environments. The study further delineates how sensorimotor dynamics modulate stress regulation in mediated contexts, comparing static (Stat-SST) and dynamic (Dyn-SST) versions of a Social Stress Test (SST). Behavioral indices indicated superior stress regulation in static conditions, where limited sensory input mitigated arousal, whereas dynamic simulations, featuring continuous audiovisual and non-verbal cues, imposed greater cognitive effort and emotional reactivity. EEG analyses confirmed reduced power in delta, theta, and alpha bands in the dynamic group, reflecting intensified regulatory demands, alongside heightened gamma activity, linked to advanced cognitive processing. Autonomic data revealed elevated skin conductance responses under dynamic conditions, consistent with increased emotional engagement. Collectively, these findings underscore that dynamic digital interactions, while ecologically richer, potentiate stress responses through amplified sensorimotor activation, with potential implications for cumulative strain in frequent mediated exchanges.

Figure 2 - *Experimental design of the Social Stress Test (SST), showing static (facial images) and dynamic (video of real evaluators) evaluation board conditions across baseline, preparation, and speech phases*



In addition, a recent research explored *brain synchronization* during interoceptive coordination tasks, using EEG and fNIRS. Findings identified specific neural biomarkers of interpersonal tuning – coherence in hemodynamic responses and EEG oscillations. Such biomarkers can inform the design of digital platforms that aim to simulate in-person social synchrony (Balconi et al., 2023).

Other research underlines how EEG, fNIRS, and autonomic metrics (heart rate, skin conductance) converge to assess emotional processing and motor imagery tasks. Results showed strong neurophysiological correlations across modalities, supporting a multimodal approach for studying embedded cognitive processes. This model underpins educational neuroscientific applications, particularly within AI-mediated learning environments.

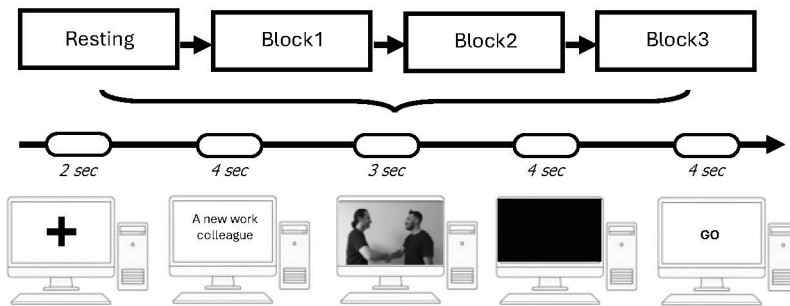
6.2. Biomarkers of interpersonal tuning: hemodynamic & electrophysiological measures

Balconi and Vanutelli (2017) used EEG and fNIRS to examine empathic neural responses under *cooperative vs. non-cooperative conditions*. Activation patterns varied by context: right prefrontal engagement in perceived negativity and left prefrontal activation in positive scenarios. Correlations with autonomic data emphasized integrated physiological signatures of empathy – relevant to interactions mediated by digital systems lacking nonverbal cues. Empathic neural responses shown in cooperative settings – dependent on subtle prefrontal activations – highlight how digital interaction may lack the subtleties of culture-based

nonverbal cues. This bears implications for international or cross-cultural digital communication platforms, which must account for such nuances to maintain relational integrity.

About *nonverbal communication*, a neuroscientific research (Balconi and Fronda, 2022; Fig. 3) recorded simultaneous EEG from interacting dyads exposed to culturally specific gestures (Italian or French). Hyperscanning revealed that neural synchronization varied by cultural familiarity, illustrating how digital-chat or video platforms might inadequately transmit culturally embedded nonverbal signals critical for social attunement.

Figure 3 - *Experimental setup for EEG hyperscanning of dyads observing culturally specific gestures (Italian or French) across five sequential video-based conditions*



About *digital learning* environments, Crivelli and Balconi (2024) examined how cognitive neuroscience can guide *corporate e-learning*. It discusses the use of EEG/fNIRS to evaluate cognitive load and emotional states during digital training. The work advocates neuro-informed digital learning strategies that recognize individual differences and allow non-reductive customization of educational experiences.

Thus, adopting a neuroscientific approach, these results repeatedly demonstrate the power of EEG, fNIRS, and autonomic integration to probe social and emotional processes. The hyperscanning studies, particularly those involving dyadic interaction, underscore that co-presence enhances neural synchrony – often lost or attenuated in digital contexts. This speaks directly to designing digital platforms that must accommodate multi-sensory and relational fidelity.

By applying these principles to corporate and educational settings, it was emphasized that neuro-informed adjustments – based on EEG and

fNIRS data – can enhance cognitive engagement and personalization. These works encourage a move away from standardized digital learning and toward systems that flex to individual neuro-cognitive and emotional needs.

At the meantime, this interpersonal neuroscientific approach offers a rigorous neuroscientific framework for understanding both the opportunities and challenges of digitalization in social, educational, and professional spheres. Key contributions include:

- demonstrating that *multimodal neurophysiological measures* can reveal digital interaction's emotional and cognitive costs;
- highlighting the *diminished neural synchronization* and embodied empathy in *mediated contexts*;
- stressing the importance of *relational fidelity* – via embodied cues and cultural signals – for trust, empathy, and engagement;
- *applying neuroscience* to shape adaptive, personalized digital training and learning platforms.

Research underscores that digital environments can become truly relational only when they embrace the complexity of human cognition and emotion. To ensure digitalization supports genuine human flourishing, platform and program design must reintroduce embodied, empathetic, and culture-sensitive elements grounded in neurophysiological findings.

7. Conclusion and future perspectives on AI, digitalization, and human flourishing for individuals and society

The integration of AI and digital technologies into social, professional and educational systems marks a pivotal transformation in the way we conceptualize learning, relational engagement, and personal development. The research reviewed demonstrates that these tools offer significant potential to enhance learning outcomes, individualize instruction, and extend access. *Neurotechnologies* such as EEG, fNIRS, and BCI (Nouri, 2025) provide real-time feedback loops that can adapt learning materials to cognitive states, potentially increasing engagement and retention.

However, this growing body of literature also reveals substantial risks. It was argued that excessive digital interaction – particularly when substituting face-to-face engagement – can erode empathy, emotional synchrony, and social support structures. Neuroscientific findings (Balconi et al., 2023) consistently show that digital mediation may reduce inter-brain connectivity and mirror neuron activation, fundamental for affective resonance. Second, over-reliance on digital tools

may externalize self-regulation and diminish metacognitive agency, particularly in young learners.

While immersive technologies may partially bridge these gaps by fostering embodied presence in virtual environments, they remain imperfect approximations. Emotional development, social learning, and identity formation are processes deeply rooted in *embodied, multisensory, and socially co-regulated contexts* – elements not yet totally replicable by AI.

7.1. Toward ethical and inclusive design

Firstly, to address these risks, multiple scholars advocate for the development of AI systems based *on ethical and inclusive design principles*. It was proposed to create an integrative framework for embedding digital well-being into educational and social infrastructures. This model encourages developers, educators, and policy-makers to prioritize emotional literacy, autonomy, and ethical awareness when crafting digital learning ecosystems. Neuroscientific insights can significantly inform these efforts. For instance, Nouri (2025) emphasizes that ethical educational AI must consider *neurodiversity* and *developmental variability* in its algorithms. Adaptive systems should not enforce uniform standards of cognitive performance but instead foster differentiated pathways that honor individual learning trajectories.

Furthermore, the inclusion of *culturally and socially diverse perspectives* in AI design is critical to prevent the marginalization of non-dominant identities. Educational technologies must be designed to respect and respond to varying cognitive styles, linguistic contexts, and emotional norms.

Secondly, *promoting conscious and reflective use of digital tools* is another key recommendation. Digital self-control tools, when integrated thoughtfully, can foster intentional technology use and reduce digital distraction. Neuroscientifically, these tools can help scaffold *executive function* development by strengthening goal management, inhibitory control, and emotional regulation. However, the literature warns against over-scaffolding. External tools should not replace, but rather *support*, the *gradual internalization of self-regulation strategies*. Educators must be trained to mediate these processes, helping students understand not only how to use digital tools but also when and why to disengage from them.

Thirdly, given the affective and neurocognitive limitations of mediated communication, multiple studies emphasize the *irreplaceable value of physical, in-person interactions*. As we have underlined, face-to-face contact promotes emotional well-being and cognitive vitality, partic-

ularly among older adults. From a developmental standpoint, these findings call for the deliberate integration of physical interaction opportunities in digital learning environments. *Hybrid models* that combine technological affordances with embodied, co-present experiences may offer a balanced solution. As Balconi (2023) suggests, future educational platforms should aim not merely to digitize content delivery but to preserve the richness of social learning through *multimodal and interactive formats*.

Fourthly, at the policy level, researchers provide actionable frameworks for aligning digital innovation with social and emotional well-being. Bond (2025) advocates for ‘*techno-socio-emotional impact assessments*’ – tools for evaluating the psychological and relational implications of educational technologies before widespread adoption. On the other hand, Ettman (2023) highlights the importance of systemic support for educators and families, including digital well-being curricula, professional development in socio-emotional facilitation, and equitable access to emotionally intelligent technologies. This policy model emphasizes *multi-stakeholder collaboration* among technologists, neuroscientists, educators, and policy-makers. In addition, Feijóo (2020), in turn, focuses on *regulatory frameworks* to ensure data privacy, ethical AI use, and non-discriminatory design. He underscores the risk of algorithmic bias and the need for transparent auditing systems in educational AI deployment.

7.2. Rethinking flourishing in digital education

As the convergence between AI, well-being, and neuroscience deepens, we are called to rethink what it means to flourish in a digital society. The literature reviewed offers a sobering but hopeful view: while digital tools can support learning and extend access, their uncritical implementation risks narrowing the emotional, ethical, and relational dimensions of well-being.

To safeguard *holistic human development*, AI must be embedded in frameworks that honor embodied cognition, promote socio-emotional growth, and respect neurodiversity. Ethical design, conscious use, and the valorization of human presence are not optional add-ons but essential components of any meaningful digital transformation in social relationships.

Future research and policy must work in synergy to foster environments where technology amplifies – not replaces – the richness of human connection.

References

- AYYAGARI R., GROVER V., PURVIS R. (2011). *Technostress: Technological Antecedents and Implications*, in "MIS Quarterly", 35(4), p. 831.
- BALCONI M. (2025) (ed.). *Neuroscienze del benessere*, Roma, Carocci Editore.
- BALCONI M., ALLEGRETTO R.A., ANGIOLETTI L. (2023). *Autonomic synchrony induced by hyperscanning interoception during interpersonal synchronization tasks*, in "Frontiers in Human Neuroscience", 17, 1200750.
- BALCONI M., ANGIOLETTI L. (2023). *Inter-Brain Hemodynamic Coherence Applied to Interoceptive Attentiveness in Hyperscanning: Why Social Framing Matters*, in "Information", 14(2), p. 58.
- BALCONI M., ANGIOLETTI L., ROVELLI K. (2024). *Sensorimotor Simulation's Influence on Stress: EEG and Autonomic Responses in Digital Interviews*, in "Brain Sciences", 14(6), p. 608.
- BALCONI M., CASSIOLI F. (2022). *"We will be in touch". A neuroscientific assessment of remote vs. face-to-face job interviews via EEG hyperscanning*, in "Social Neuroscience", 17(3), pp. 209-224.
- BALCONI M., DI LERNIA D., CASSIOLI F., ANGIOLETTI L., GASPARI S.D., SERINO S., SANJO E., SANSONI M., MAURI M., ROVELLI K., ALLEGRETTO R.A., AMADINI GENOVESE L., RIVA G. (2023). *Exploring the neurophysiological correlates of face-to-face vs. remote learning during lecture and group discussions: a hyperscanning paradigm*, in 8th Scientific Meeting of the Federation of European Societies of Neuropsychology (FESN) & the 2nd Hellenic Neuropsychological Society (HNPS) Conference, 29 September.
- BALCONI M., FRONDA G. (2022). *Inter-brain EEG connectivity in hyperscanning for Italian and French gestures: the culture-related nonverbal language*, in "Culture and Brain", 10(1), pp. 56-77.
- BALCONI M., VANUTELLI M.E. (2017). *Interbrains cooperation: Hyperscanning and self-perception in joint actions*, in "Journal of Clinical and Experimental Neuropsychology", 39(6), pp. 607-620.
- BOND R.R., ENNIS E., MULVENNA M.D. (2025). *How artificial intelligence may affect our mental wellbeing*, in "Behaviour & Information Technology", 44(10), pp. 2093-2100.
- BRAKEL V. VAN, BARREDA-ÁNGELES M., HARTMANN T. (2023). *Feelings of presence and perceived social support in social virtual reality platforms*, in "Computers in Human Behavior", 139, 107523.
- CASSIOLI F., BALCONI M. (2022). *The Complexity of Remote Learning: A Neuroergonomical Discussion*, in "Frontiers in Neurobotics", 16, 842151.
- CRIVELLI D., BALCONI M. (2024). *From physical to digital: A theoretical-methodological primer on designing hyperscanning investigations to explore remote exchanges*, in "Social Neuroscience", ahead-of-print(ahead-of-print), pp. 1-.
- DECI E.L., RYAN R.M. (2000). *The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior*, in "Psychological Inquiry", 11(4), pp. 227-268.

DOHENY M.M., LIGHTHALL N.R. (2023). *Social cognitive neuroscience in the digital age*, in "Frontiers in Human Neuroscience", 17, 1168788.

ETTMAN C.K., GALEA S. (2023). *The Potential Influence of AI on Population Mental Health*, in "JMIR Mental Health", 10, e49936.

FEIJÓO C., KWON Y., BAUER J.M., BOHLIN E., HOWELL B., JAIN R., POTGIETER P., VU K., WHALLEY J., XIA J. (2020). *Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy*, in "Telecommunications Policy", 44(6), 101988.

KUSHLEV K., PROULX J., DUNN E.W. (2016). "Silence Your Phones", in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 1011-1020.

KUSS D.J., GRIFFITHS M.D. (2017). *Social Networking Sites and Addiction: Ten Lessons Learned*, in "International Journal of Environmental Research and Public Health", 14(3), 311.

MACDONALD B., LUO M., HÜLÜR G. (2021). *Daily social interactions and well-being in older adults: The role of interaction modality*, in "Journal of Social and Personal Relationships", 38(12), pp. 3566-3589.

MEIER J.V., NOEL J.A., KASPAR K. (2021). *Alone Together: Computer-Mediated Communication in Leisure Time During and After the COVID-19 Pandemic*, in "Frontiers in Psychology", 12, 666655.

MYRUSKI S., CAHILL B., BUSS K.A. (2024). *Digital Media Use Preference Indirectly Relates to Adolescent Social Anxiety Symptoms Through Delta-Beta Coupling*, in "Affective Science", 5(4), pp. 310-320.

NASLUND J.A., ASCHBRENNER K.A., MARSCH L.A., BARTELS S.J. (2016). *The future of mental health care: peer-to-peer support and social media*, in "Epidemiology and Psychiatric Sciences", 25(2), pp. 113-122.

NOURI A. (2025). *A scoping review of educational neurotechnology: Methods, applications, opportunities, and challenges*, in "Review of Education", 13(2).

OLUWATOYIN F. (2021). *The Dynamics of Social Interaction in the Digital Age: Technological Implications for Interpersonal Relationships and Psychosocial Well-Being*, in "Journal of Social Science Studies", 1(2), pp. 137-142.

PETERSEN B., KHALILI-MAHANI N., MURPHY C., SAWCHUK K., PHILLIPS N., LI K.Z.H., HEBBLETHWAITE S. (2023). *The association between information and communication technologies, loneliness and social connectedness: A scoping review*, in "Frontiers in Psychology", 14, 1063146.

PRZYBYLSKI A.K., WEINSTEIN N. (2016). *A Large-Scale Test of the Goldilocks Hypothesis*, in "Psychological Science", 28(2), pp. 204-215.

ROFFARELLO A.M., RUSSIS L.D., LOTTRIDGE D., CECCHINATO M.E. (2023). *Understanding digital wellbeing within complex technological contexts*, in "International Journal of Human-Computer Studies", 175, 103034.

RYFF C.D., KEYES C.L.M. (1995). *The Structure of Psychological Well-Being Revisited*, in "Journal of Personality and Social Psychology", 69(4), pp. 719-727.

SHAKYA H.B., CHRISTAKIS N.A. (2017). *Association of Facebook Use with Compromised Well-Being: A Longitudinal Study*, in "American Journal of Epidemiology", 185(3), pp. 203-211.

TORBJØRNSSEN R.R., HIPÓLITO I. (2025), *Widening the screen: embodied cognition and audiovisual online social interaction in the digital age*, in "AI & Society", 40(1), pp. 21-35.

VILLAMIL A., HESHMATI S. (2023). *Engaging in the good with technology: a framework for examining positive technology use*, in "Frontiers in Psychology", 14, 1175740.

VOGEL E.A., ROSE J.P., ROBERTS L.R., ECKLES K. (2014). *Social Comparison, Social Media, and Self-Esteem*, in "Psychology of Popular Media Culture", 3(4), pp. 206-222.

SECTION 3

Professional Applications
and Sectoral Transformation

1. Participatory AI and Industrial Democracy: The Role of Social Dialogue in Governing Workplace Technologies

Davide Bisi, Ivana Pais

ABSTRACT

In the contemporary discourse surrounding Artificial Intelligence, the invocation of human-centric design often masks a profound avoidance: the interrogation of human relationality and collective agency within technological spaces. This chapter reframes the debate around workplace AI governance, positing that the crucial determinant of AI's socio-economic impact is not the mere presence of human oversight, but the structure of participation through which humans collectively shape technological trajectories.

Drawing from recent Italian experiences in industrial relations, we show how participatory AI emerges as a political and institutional project rather than merely an ethical corrective. Analyzing pioneering company-level collective agreements, we illustrate the tangible mechanisms – such as joint monitoring bodies, explicit prohibitions on worker substitution, and protections against surveillance – that unions and employers are beginning to negotiate. These mechanisms represent the concrete institutional practices through which a high road to AI adoption is carved – one that channels innovation toward inclusive productivity gains rather than intensified control or rent extraction. This form of negotiated governance does not eliminate power asymmetries, but it opens up arenas where these imbalances can be surfaced, challenged, and, at times, reconfigured. Ultimately, this chapter argues that industrial democracy in the era of AI is not an anachronism but an essential precondition for humane and effective technological innovation. It calls for a participatory framework robust enough to democratize not just the application, but the very creation of workplace technologies – reclaiming the promise of AI as a collaborative, rather than coercive, force in shaping the future of work.

1. *Artificial intelligence and collective action*

In public discourse on AI, the centrality of the human being is frequently invoked as a normative principle, through concepts such as *human-in-the-loop*, algorithmic transparency, and ‘human-centered’ systems. Yet the notion of ‘being human’ is rarely examined critically within technologically mediated contexts. When the human is treated merely as an abstract, isolated individual – a rights-bearing subject, an

informed recipient, a rational user – we lose the opportunity to interrogate the social and relational forms through which technology acquires meaning. As Riva's work on digital identity suggests, subjectivity is not a given, but a situated process constituted within relationships, contexts, and technological configurations (Riva and Chiriatti, 2024; Riva, 2025). This is especially true in the workplace, where AI does not simply interact with individuals, but actively shapes roles, functions, hierarchies, and degrees of autonomy.

In this sense, a *humane* perspective cannot be limited to ensuring the presence of a human decision-maker within the algorithmic loop. Rather, it must ask which subjects are involved in innovation processes, with what tools, and with what degree of influence. Work – understood as a relational, economic, and institutional space – becomes the primary testing ground for such inquiry. It is here that, often implicitly, it is decided whether AI will serve as a driver of cooperation or a mechanism of control, whether it will enhance the autonomy of workers and clients or constrain their discretion, and whether it will be used to expand capacities for action or to standardize them.

If artificial intelligence can be understood not only as a computational infrastructure but also as a relational one, then a fundamental question arises: who participates in its construction? The concept of *participatory AI* emerges precisely from this inquiry. It does not merely call for human presence in decision-making processes, but demands that the design, implementation, and monitoring of intelligent technologies be shaped through collective and negotiated processes, capable of including a diversity of actors – differing in roles, interests, and expertise.

However, it is essential to clarify from the outset that participation is never given; it must be constructed. Invoking algorithmic transparency or the intelligibility of metrics is not sufficient. A technology may meet formal standards of explainability without being subject to meaningful forms of collective oversight. Similarly, it may appear user-centric at the operational level while remaining embedded in proprietary logics of value extraction and organizational control.

For this reason, participatory AI should not be understood as an ethical label retroactively applied to technological systems. Rather, it represents a political and design horizon – one that calls for a fundamental rethinking of who has a voice in decision-making, how that voice is exercised, and what constraints and responsibilities accompany it. It involves building technologies not merely for people, but with them, through frameworks of co-responsibility, deliberation, and epistemic plurality.

The adoption of artificial intelligence in the workplace is not a

neutral act. Technologies are often presented as objective tools aimed at increasing productivity, efficiency, and safety. However, this framing overlooks a fundamental point: technologies are always also political choices. This is true not only in a normative sense, but insofar as technologies function as strategies that reflect – and often reinforce – existing power relations.

In this light, the integration of AI in work environments raises crucial questions: are these technologies being deployed to enhance existing skills and improve job quality, or rather to standardize tasks, reduce professional discretion, automate surveillance processes, and make workers more interchangeable? The current evidence does not point to a single answer. Instead, it suggests that these logics coexist, and that actual outcomes depend on institutional conditions, the quality of industrial relations, and the technological governance models that are adopted (Soulami et al., 2024; Racabi and Litwin, 2024; Ozkiziltan, 2024).

Against this backdrop, industrial relations are beginning to engage with AI as a matter of collective regulation. The Italian context offers a privileged observatory: over the past two years, several company – and sector-level agreements – have introduced clauses, procedures, and bodies aimed at governing the introduction of AI. These remain heterogeneous and largely experimental initiatives, but they provide important insights into how negotiated participation can concretely shape the governance of innovation.

The following section reconstructs the main mechanisms through which collective representation can influence the trajectories of AI implementation in the workplace. It highlights the specificities of company-level bargaining and the potential of this level of collective negotiation to enable the shared governance of AI. Alongside the more established tools of bargaining – such as the joint definition of the scope of technologies and the regulation of their purposes – less visible yet equally impactful levers also emerge: training, work reorganization, working time management, and the delineation of skills.

The third section focuses on the positions taken by Italy's major trade union and employer organizations regarding the initial forms of AI regulation in the workplace. Despite differences in emphasis and priorities, a progressively more structured dialogue is taking shape around key issues: the need for multi-level governance, the challenges of algorithmic oversight, concerns around transparency, and the more contentious debate over the role of statutory regulation in shaping these processes.

The fourth section forms the empirical core of the chapter, analyzing a curated selection of collective agreements signed in Italy

at company level. Though still limited in number, these agreements are conceptually advanced and provide a snapshot of the early contractual responses to the challenges posed by AI. The analysis focuses on several key dimensions: the identification of technologies subject to bargaining, the intended purposes and expected benefits, the topics included in the negotiation process, and the evolving nature of the agreements.

Finally, the fifth section reflects on future perspectives. The contractual experiences discussed – while partial and experimental – signal a growing awareness among social partners. However, for participation to move beyond a limited set of pioneering cases, systemic conditions must be strengthened: expanding the scope of collective bargaining, ensuring access to technical information, fostering shared competencies, and developing a negotiation language capable of addressing the complexity of technological transformations.

2. The role of workers' representatives in the adoption of AI

In the debate on the impact of artificial intelligence in the workplace, the presence of collective representation institutions is often portrayed as a barrier to innovation or as a legacy of a bygone industrial era.

However, a growing body of empirical evidence suggests that this view is misguided. When endowed with meaningful participatory rights, workers' representatives do not hinder AI adoption. On the contrary, they can contribute to making it more effective, sustainable, and inclusive. Company-level bargaining is widely regarded as the most effective instrument for regulating the use of AI-based systems. Its inherent flexibility allows it to safeguard workers by mitigating associated risks, while also promoting the potential societal benefits of these technologies (De Stefano and Taes, 2021).

Recent studies have documented that the presence of workplace-level employee representation is positively associated with the adoption of AI technologies, particularly in countries that recognize co-determination rights over technological processes. In such institutional configurations, worker representatives are not merely informed or consulted – they are actively involved in negotiating agreements on the use of technology, securing protections for workers and defining operational conditions for firms (Belloc et al., 2023; Burdin et al., 2023). The outcome is not a simple compromise, but rather a form of cooperative governance capable of reducing uncertainty, mitigating the risks of adversarial conflict, and enhancing the value of technological investment.

However, this is not a conflict-free idyll. When functioning

effectively, representative institutions do not eliminate power asymmetries in the workplace; rather, they provide arenas and procedures through which such imbalances can be contested and renegotiated. In contexts where worker voice is perceived as legitimate, and where a sufficient degree of mutual trust exists, agreements on the responsible use of AI are more likely to emerge. This is evidenced by the German cases on algorithmic management (Doellgast et al., 2023; Krzywdzinski et al., 2023) and by the Danish agreement within the Hilfr platform, which introduced specific safeguards related to data use, profiling, and trade union rights in digital environments (Ilsøe et al., 2023). These agreements often establish limits to technological intrusiveness, introduce standards for transparency and human oversight, and in many cases anticipate the content of forthcoming public regulation.

Equally important in this context is the presence of indirect mechanisms of cooperation that can facilitate technological adoption. Where worker representatives can influence organizational practices such as training, work reorganization, or working time management, the ground is more conducive to the introduction of AI tools that require new skills and continuous adaptation (Dauth et al., 2021; Battisti et al., 2023).

In such settings, collective representation assumes an enabling role, helping to mediate technological integration through institutionalized practices of dialogue and organizational adjustment. The literature has highlighted how the role of social partners and the ways in which protective and participatory instruments are combined to address the challenges posed by AI and algorithmic management vary significantly across countries. At the same time, existing institutions and practices are often inadequate to address the complexity of these transformations. This underscores the need for institutional adaptation and the strengthening of the regulatory framework, both at the European and national levels, to support collective bargaining (Molina et al., 2023).

3. Positioning of Italian Social Partners regarding Artificial Intelligence

The regulation of AI in the workplace has sparked intense debate among Italian Social Partners, characterized by distinct perspectives on governance, statutory regulation, and human oversight of algorithms¹.

¹ The considerations presented here are based on the work of the 'Industrial Policies for Artificial Intelligence' working group established at CNEL, which conducted a comprehensive analysis of the materials collected during the public consultation launched by the Ministry of Labour and Social Policies (2025) on the guidelines for implementing artificial intelligence in the workplace. The group also examined the

This contribution reconstructs these positions by highlighting, on the one hand, the divergences between the parties – particularly regarding the role of trade union representation, the scope of regulation, and the nature of human supervision – and, on the other hand, the points of convergence that are emerging on issues such as governance, algorithmic transparency, and technological sovereignty.

The first area of divergence in the debate concerns the role of trade union representation in the introduction of AI technologies. Trade unions emphasize the necessity of proactive collective bargaining as an essential instrument to safeguard workers' fundamental rights and ensure meaningful involvement of worker representatives in all decision-making phases related to technology adoption. CGIL and UIL explicitly argue that the evaluation and adoption of AI tools must be mandatorily included in collective bargaining, whereas CISL highlights the importance of strengthening the bargaining role through informational and consultative obligations consistent with European legislation.

The positions of employer associations, represented by Confindustria and Confartigianato, emphasize the risks of overly restrictive collective bargaining, which could slow technological innovation and weaken the competitiveness of the national economic system. These organizations advocate for a more adaptive and dynamic regulatory approach that ensures rapid technology adoption without compromising worker protections. Distinctly, CIDA, the confederation representing managers and professionals – while not opposing regulatory measures – stresses the importance of technical instruments such as algorithmic audits, standard contractual clauses, and sectoral protocols.

Another area of divergence concerns the human oversight of algorithmic decisions. Trade unions and Confindustria (the confederation of independent professionals) agree on the centrality of continuous and mandatory human control to ensure transparency and the protection of individual rights. However, other voices challenge the assumption that the entire functioning of these systems must always be explainable and subject to human review. Confartigianato points out that, especially in deep learning-based models, explainability may conflict with predictive accuracy, suggesting that insisting on full transparency risks sacrificing the precision and effectiveness of systems, particularly in contexts where speed and reliability are critical.

While divergences reflect the structural differences between the

documents submitted during the parliamentary review of Draft Law No. 2316/2024 ('AI Bill'). The analysis covered the full set of contributions submitted by the trade union and employer organizations involved, including: CIDA, CISL, CGIL, Confartigianato, Confindustria, Confindustria, Confindustria, and UIL.

mandates of the organizations involved, the areas of convergence indicate the possibility of a shared vision regarding the digital transformation of work. The foremost convergence concerns the need to create a multilevel governance framework for artificial intelligence that is simultaneously stable, inclusive, and coordinated.

In this regard, there is broad support for the creation, as mandated by the decree law, of the National AI and Work Observatory, housed within the Ministry of Labor. The contribution of social partners focuses particularly on defining the Observatory's composition and functions, which are considered crucial to its effectiveness. Organizations demand explicit inclusion of the most representative bodies to ensure a balance among union, employer, and self-employed stakeholders. The functions assigned to the Observatory should include, in addition to monitoring, data collection, the formulation of recommendations, and the promotion of best practices. In line with this, several proposals emphasize the need to guarantee the Observatory a minimum level of operational autonomy and dedicated resources.

National technological sovereignty represents another area of significant convergence. All parties acknowledge that dependence on external suppliers – especially for infrastructural components and proprietary AI models – risks undermining not only the competitiveness of the production system but also the capacity to exercise democratic and ethical governance over technological processes. In this regard, there is a shared proposal to strengthen the strategic dimension of public investment through targeted industrial policy measures, support for national ecosystems (such as digital innovation hubs), and the development of public platforms for the responsible adoption of AI, equipped with open-source tools, transparent models, algorithmic audits, and rapid prototyping services.

Finally, the protection of personal data and the security of digital systems constitute an area of broad consensus among social partners, despite differing perspectives. Trade union representatives emphasize the need to accompany the introduction of AI systems in workplaces with a series of structural safeguards: independent and periodic audits, union access to data generated by technologies, and, above all, a mandatory human review in all decision-making processes with occupational or organizational impact.

4. Findings from Italian cases

The introduction of AI in the workplace has prompted an initial contractual response in Italy that, although still limited in scope,

features conceptually advanced and innovative regulatory elements. This regulation is neither systemic nor uniform: the landscape of agreements is fragmented, often tied to specific corporate contexts, and varies in ambition and detail. However, this very heterogeneity currently serves as a valuable empirical laboratory for understanding how social partners are attempting to engage with a continuously evolving technological phenomenon, intervening in areas that have so far been scarcely regulated, such as data use, algorithmic monitoring, and labor transformations.

The analysis presented is based on a carefully selected sample of corporate cases reconstructed within the activities of the OPERA Observatory – Observatory for Policies and Industrial Relations on Participatory AI, promoted by CNEL². The analysis examines a selection of corporate agreements signed between January 2024 and March 2025, explicitly referencing artificial intelligence systems. These agreements span diverse sectors but uniformly feature substantial exposure to digital transformation and high informational intensity in their core activities. The sample includes firms from three main sectors. The pharmaceutical sector is represented by Gilead Sciences, a multinational focused on the research and commercialization of innovative therapies, and Sanofi, a global pharmaceutical company with a longstanding presence in Italy, active in the production of medicines and vaccines. The insurance sector includes Generali, one of Europe's leading groups with a complex organizational structure deeply rooted in the Italian market, and CNP Vita Assicura, a company controlled by the French group CNP Assurances, focused on the Italian market. The digital services sector comprises Italiaonline, Italy's leading internet company specializing in digital advertising and web solutions for businesses, and Konecta Italia, active in Customer Relationship Management and Business Process Outsourcing. Lastly, the industrial sector is represented by Saipem, an Italian company operating in engineering and construction for energy and infrastructure, with particular relevance in the oil & gas sector.

Based on a thematic reading of these agreements, the analysis is organized around four main dimensions: (1) the AI technology subject to negotiation; (2) the intended purposes and expected benefits; (3)

² The Observatory collects, systematizes, and analyzes empirical evidence on the introduction of AI in Italian workplaces, with particular attention to the participatory dimension of the phenomenon.

For more information about the OPERA Observatory: <https://www.cnel.it/Chi-Siamo/XI-Consiliatura/Altri-Organismi-XI-Consiliatura/Osservatorio-politiche-e-relazioni-industriali-per-lintelligenza-artificiale-partecipativa-OPERA>.

the subjects of negotiation; and (4) the evolutionary dimension of the agreement.

4.1. The AI technology subject to negotiation

The analysis of the agreements reveals significant diversity in the AI technologies adopted, which vary in terms of their domain of application and their assistive or substitutive nature, as well as their functional purpose, level of automation, and degree of impact on work organization. With regard to the domain of application, it is useful to distinguish between *management-automating AI* and *work-automating AI* (Doellgast et al., 2023).

Management-automating AI applications refer to the use of AI-based analytics and algorithms in traditional management activities (such as recruitment, monitoring, training, scheduling of shifts and breaks). These tools may introduce risks related to intensified surveillance, reduced worker autonomy, discrimination, and privacy violations. The agreements of Sanofi and Generali, while situating AI within a broader framework of technological innovation that includes predictive tools and automation of information flows, also refer to decision-support technologies for managerial and operational functions.

The applications most frequently regulated in the AI-related contracts concluded so far are those attributable to work-automating AI, which employs AI-based tools to directly automate tasks or to assist workers in task execution. The risks associated with these tools include job displacement, task deskilling, reduced autonomy, and work intensification. Work-automating AI applications can be further distinguished between *assistive* and *substitutive* technologies.

An example of assistive technology is the AI system implemented by Saipem for processing images captured by smart cameras installed on ships and construction sites to support human activities in the field of health and safety. This system functions to identify real-time risk situations, such as a worker's fall ("man down").

In other cases, the agreement intervenes specifically to mitigate the risks associated with potentially substitutive technologies. In the case of CNP Vita Assicura, although the AI systems are not described in technical detail – only in terms of their application areas, such as customer interaction, document management, and insurance risk assessment – the agreement explicitly acknowledges the potential substitutive impact and introduces protective measures, including explicit safeguards against the replacement of workers.

There are other cases in which the issue of substitution is not explicitly mentioned, yet collective bargaining creates a negotiated

framework that allows for the anticipation and management of such risks.

The Konecta case, for example, concerns the introduction of the *Agent Assist* platform, which provides real-time suggestions to call center operators during customer interactions. AI is used to interpret the content of conversations, propose standardized responses, and optimize handling times. The technology relies on language models and semantic classification and is regulated through a detailed agreement that specifies its scope of application, the informational perimeter, and the terms of a 12-month experimental evaluation. This represents a typical example of AI integrated into the operational workflow, with direct implications for work organization and timing, but not primarily intended to replace human labor. A similar case is that of Italiaonline, where the agreement regulates the adoption of AI technologies for automated language analysis (speech-to-text) and sentiment analysis applied to communication flows. In this instance, the technology functions as an analytical filter supporting business intelligence processes and internal organizational management.

Gilead presents a more specific case: the adoption of digital tools aimed at personalizing and optimizing medical-scientific communication targeted at healthcare professionals. AI is employed to segment recipients, select relevant informational content, and tailor it to individual preferences. The technology does not apply directly to sales or internal management activities but operates as an intermediate filter in the production of content.

4.2. Shared objectives and anticipated benefits of AI

One of the recurring thematic cores of collective agreements on artificial intelligence concerns the articulation of anticipated benefits and the shared objectives between companies and workers' representatives. Although the agreements vary in detail and structure, they frequently refer to improvement-oriented goals related to processes, services, or working conditions. These statements serve a dual function: on the one hand, they clarify the strategic intentions of the company regarding the introduction of AI; on the other, they establish a normative framework that shapes and legitimizes subsequent technical and organizational provisions.

Service quality enhancement is the most frequently cited benefit. In such cases, the adoption of AI is framed as part of a broader strategy to strengthen reliability, operational efficiency, and communicative consistency. For instance, the agreement at Gilead highlights how tools for the personalization of scientific content targeted at medical profes-

sionals can ensure more relevant, timely, and up-to-date communication. At Italiaonline, references to sentiment analysis and the automated generation of digital content are accompanied by objectives related to optimizing informational services and improving customer responsiveness.

A second group of agreements identifies among the expected benefits the reduction of task repetitiveness and cognitive workload. Both Konecta and Italiaonline state that the technologies adopted should ease the burden on workers by supporting interaction management or enabling the automatic generation of standardized content. At Sanofi, the reduction of repetitive tasks is explicitly linked to an improvement in job quality and a stronger focus of human resources on value-added activities.

A less common but highly specific objective is occupational health and safety, which is addressed in a technical and operational manner in the Saipem agreement. The contract introduces an AI system applied to video analysis from smart cameras installed on ships and construction sites, capable of automatically detecting critical situations such as falls, absence of protective equipment, or unauthorized access. The assistive function of AI is clearly defined and aimed at preventing HSE risks. Notably, this application is accompanied by a detailed technical description and a joint monitoring mechanism, positioning the Saipem case as an advanced example of AI oriented toward the proactive protection of working conditions.

4.3. Subjects of negotiation

Contractual regulation has so far addressed three distinct but interconnected areas: data management and knowledge generation from human activity, performance monitoring, and employment impact.

4.3.1. Data and training: knowledge extraction and control over cognitive contributions

With the introduction of AI in workplace settings, the possibility that data generated daily by workers may be used as input for training algorithmic models has gained increasing prominence. Specifically, this concerns information produced in textual, vocal, or behavioral forms – such as transcriptions, response patterns, operational choices, vocal or biometric signals – which, once collected and aggregated, can contribute to the machine learning phase of these systems. This dynamic – often undeclared and technically opaque – raises significant contractual questions. The primary concern relates to the ownership

and intended use of the data. To whom do the information produced during work activities belong? Who holds the right to select, store, analyze, and, above all, reuse such data for purposes beyond those for which they were originally collected?

The second question concerns the nature of workers' cognitive contribution: interaction with digital systems, especially during initial phases, often involves activities such as verifying, correcting, selecting, or classifying AI-generated outputs. Although these activities are not always formalized as autonomous labor, they represent a transfer of human knowledge to the machine, which can subsequently be codified, generalized, and reused. In the absence of specific safeguards, there is a risk of creating a grey area where labor contributes to the improvement of systems without being recognized either as productive work or as a subject of negotiation.

Regarding these issues, the corporate agreement of Konecta represents the most advanced case. It includes a clear and unequivocal clause prohibiting the use of worker-generated data to train AI systems for substitutional purposes. The prohibition specifically targets the *Agent Assist* function, specifying that the outputs from interactions cannot be used to develop tools aimed at replacing the human operator. Although tied to a specific technology, this formulation establishes a general principle: AI training cannot rely on cognitive resources produced by live labor without explicit consent and negotiated terms.

4.3.2. Monitoring, evaluation, and disciplinary authority

The adoption of artificial intelligence in organizational contexts has enabled increasingly granular and continuous monitoring of work activities. Unlike traditional mechanisms, these new systems can collect real-time information, interpret behavioral patterns, and provide detailed representations of individual effort. This type of observation – automated, persistent, and often invisible – not only describes work actions but also anticipates behaviors by embedding them within a system of metrics, thresholds, and predictive models.

The issue of monitoring extends beyond its technical dimension, having structural effects on work organization, starting with the selective and targeted sanctioning of behaviors deemed inefficient or deviant. The more precise the system is in detecting deviations, the less the organization needs to broadly incentivize all workers, and the deterrent effect of potential punishment becomes sufficient to regulate conduct. This mechanism shifts the balance of bargaining power in favor of the

employer, reducing the cost of incentives and strengthening the credibility of sanctions (Bowles, 1989).

This ‘efficiency’ comes at a significant motivational cost. When workers perceive themselves as constantly observed and evaluated based on avoided punishments or potential rewards, cooperation tends to give way to mere compliance. A sense of responsibility, organizational trust, and voluntary contribution risk being replaced by instrumental compliance driven more by fear than participation (Bryson and Freeman, 2016). In the absence of negotiated safeguards, the risk is not only invasive surveillance but also an implicit rebalancing of the incentive system: less recognition, more control.

Finally, the presence of systems that observe, classify, and compare can alter the perception of working time, encourage forms of continuous self-evaluation, intensify work pace and hours, and weaken collective defenses against production pressure (Zitek and Schlund, 2024). Within collective bargaining, this technological capability raises a crucial question: to what extent can the information collected by AI systems be used to exert direct or indirect control over work performance? And what mechanisms exist to limit its effects?³

Some agreements explicitly address this challenge. Among them, the protocol signed at Italiaonline is one of the most detailed. It stipulates that information collected through AI modules (speech-to-text, sentiment analysis, knowledge base) cannot be used for meritocratic, disciplinary, or dismissive purposes against individual workers. The data, processed not in real-time and only in aggregated form from a minimum of three individuals, are intended exclusively for training and service improvement purposes. Similarly, the Konecta contract establishes that data collected by the *Agent Assist* system cannot be used for performance evaluation or the activation of disciplinary mechanisms. Together, these clauses create a protected information boundary: a domain where data may technically exist but are not authorized for evaluative use.

A similarly robust safeguard can be found in the agreement signed at Saipem, where the parties clearly stipulate that data collected through

³ It is important to recall that digital monitoring in the workplace is already subject to a complex regulatory framework, with Article 4 of the Workers’ Statute – together with the GDPR and Legislative Decree 152/1997 (as amended by the Transparency Decree and the 2023 Labor Decree) – establishing a set of well-defined constraints and information obligations. However, these legal instruments do not always capture the ambiguities of more recent algorithmic practices – particularly when AI is embedded in work tools without generating explicit control, yet producing substantial evaluative effects. It is precisely within this grey area that collective bargaining can play a more active role: translating abstract principles into operational criteria, negotiating the use of metrics, and anchoring evaluative logics to shared purposes.

smart camera systems may be used exclusively for purposes related to occupational health and safety, explicitly excluding any surveillance and/or disciplinary use. This exclusion is expressly linked to Article 4 of Workers' Statute.

In the case of Gilead, the issue of AI-based monitoring is addressed in relation to the 'Leapfrog' project, a system that employs algorithms to suggest more effective content, channels, and timing for medical-scientific communications. The agreement grants workers the possibility of rejecting the suggestions generated by the AI system, provided the decision is justified, and clarifies that such choices will have no bearing on individual performance evaluations, nor will they be used for disciplinary purposes. It is worth noting that protection here is constructed from the reversed assumption, opting out of the AI system must be justified, whereas following algorithmic suggestions is taken as the default and aligned with company expectations. This approach reverses the principle of technological neutrality: operational default becomes adherence to the system, while deviation must be explained.

While the contracts outlined above offer concrete examples of contractual containment of workplace monitoring, other agreements are less explicit in this regard. The absence of specific clauses does not necessarily imply the adoption of disciplinary logics, but it does indicate a lack of *ex ante* regulation, effectively leaving the interpretation and scope of data use to managerial discretion. In some contexts, collective bargaining attempts to address this opacity through the establishment of joint monitoring bodies, tasked with ensuring compliance with existing clauses and, in some cases, conducting preliminary assessments of functional modifications to AI systems.

4.3.3. Displacement and the transformation of work

The third and final node in this regulatory chain concerns the possibility that artificial intelligence may replace, either directly or progressively, portions of human labor. Unlike the previous axes, this issue is no longer merely about what information is extracted or how it is processed, but rather about what work remains, and what is absorbed, displaced, or transformed. At this juncture, the productive potential of AI intersects with the concrete organization of the workforce, bringing to the fore decisions that affect occupational structures, the redistribution of tasks, and rights to professional continuity.

Work transformations associated with AI are rarely linear or immediate. More often, they unfold as a gradual disaggregation of tasks: some functions are automated, others reassigned, and still others reorganized according to optimization logics that may reduce task

variety or diminish workers' control over the sequencing of their activities. In such cases, the risks are not only quantitative (e.g., job loss), but also qualitative, involving the erosion of meaning, autonomy, or coherence in work (Holm and Lorenz, 2021; Pérez et al., 2023; Bonney et al., 2024). Collective bargaining can play a pivotal role in mitigating these outcomes, but only if it is capable of detecting and anticipating processes of technical decomposition, rather than merely responding to their visible effects.

Finally, it is important to recognize that technological substitution does not always present itself explicitly. In some settings, it occurs indirectly, through the outsourcing of automated work phases or the selective reduction of hiring in low value-added roles. In others, it is implicitly incentivized, for instance by tying corporate bonuses to efficiency gains from AI adoption, without specifying whether or how the workers involved contributed to those outcomes (Fornino and Manera, 2019; Prassl, 2019).

In the agreements analyzed, this issue emerges in varying forms. The CNP Vita agreement represents the most explicit case: as previously mentioned, it includes a clause prohibiting the replacement of workers by AI technologies, stating that the adoption of new tools may not result in redundancies or dismissals. Although limited to the scope of the agreement itself, this form of employment protection constitutes a significant precedent, as it acknowledges that substitution is not merely a collateral risk to be monitored, but a concrete possibility that must be regulated *ex ante*.

Other agreements – such as those signed at Konecta or Italiaonline – do not contain such explicit clauses. Nonetheless, the way in which the functional boundaries of the adopted technologies are defined, along with the safeguards concerning the use of data for evaluative purposes, suggests a latent tension between innovation and employment stability. The absence of a substitution prohibition does not imply its legitimization, but rather indicates that potential reductions in labor demand are not yet treated as a matter for direct regulation. Instead, they are addressed indirectly, through contractual support measures. It is within this space that the difference between *ex post* governance and the capacity for social anticipation of AI's effects is most clearly at stake.

The issue becomes more articulated when one considers the measures introduced to support transitions – namely, the instruments designed to assist workers through task reorganization and role reconfiguration. Among these, training is the most recurrent measure, with content ranging from basic technical instruction (manuals, tutorials, IT support) to more structured programs (workshops, specialized courses). In the case of Generali, the agreement includes

an explicit reference to AI literacy, echoing Article 4 of the *AI Act*. The training activities are tailored by role and aim to provide a foundational understanding of the principles and potential impacts of AI on work. This approach extends the traditional scope of training to encompass elements of digital literacy and organizational awareness.

4.3.4. The evolutionary dimension of collective agreements

The adoption of artificial intelligence in the workplace is never a single, discrete event. Rather, it unfolds as an incremental process, marked by experimentation, adaptation, and ongoing reconfiguration. Within this dynamic context, collective agreements cannot merely capture a snapshot of a given technological state; they are called upon to regulate a moving trajectory, by establishing mechanisms that can monitor developments, mitigate adverse effects, and, where necessary, redirect the application of technology.

The most structured agreements respond to this challenge through three core mechanisms: a precise technical description of the AI system in question; the formalization of joint governance bodies; and the definition of a time-bound experimental phase, aimed at testing the technology's impact under controlled conditions.

The first mechanism – technical precision – serves a strategic purpose. In most agreements, the contract contains explicit references to the system's functionalities, the operational setting in which it will be deployed, and the specific goals it is meant to serve. This level of detail is not a matter of transparency alone; it is the precondition for any subsequent review, as it clearly delineates the scope of what has been agreed. Where such descriptions are lacking – as in the cases of CNP Vita or Generali – the agreement tends to adopt a more aspirational tone, relying on general statements about digital transformation and delegating concrete interpretation to secondary instruments such as steering committees or technical commissions.

The second mechanism concerns the establishment of bodies responsible for overseeing and managing the implementation of AI systems. Many agreements provide for the creation of a joint committee, steering group, or observatory, with functions ranging from procedural oversight to the analysis of organizational impacts. These bodies are typically bilateral, composed of company and union representatives, operate on a continuous basis, and are empowered to request technical documentation or carry out further inquiries. The Generali agreement, for instance, provides for the creation of a joint commission on artificial intelligence, tasked with examining the rationale behind corporate decisions on the use and dissemination of AI technologies, “including

for the purpose of collecting relevant proposals from trade union organizations.” This formulation moves beyond a purely consultative role, granting worker representatives a formally recognized space to contribute to the shaping of the company’s technological pathways.

Though these bodies lack binding powers, the agreement nonetheless establishes a principle of qualified participation – one that goes beyond unilateral information-sharing and approximates a model of cooperative governance, in which social partners are engaged *ex ante* in discussions concerning technical and organizational choices.

The Konecta agreement establishes an experimental committee composed of eight representatives from each party, tasked with verifying the conditions under which the technology is implemented and proposing potential adjustments at the end of the trial period. At Saipem, oversight of the AI system used for risk detection on construction sites is entrusted to two bodies: the Joint Company Institution (Istituzione Paritetica Aziendale, IPA), composed of company representatives and workplace safety delegates (RLSA), responsible for monitoring the project’s development and suggesting corrective measures; and the National Health and Safety Observatory, a sectoral body established by the national collective agreement, involved in the broader evaluation of the pilot phase.

Italiaonline also provides for a dedicated monitoring body known as ‘Cabina di Regia’ – a permanent joint forum between management and trade unions that convenes regularly (typically on a quarterly basis) to assess the implementation of AI modules (including speech-to-text, sentiment analysis, and knowledge base systems) and their impact on work organization. In the case of Gilead, the parties agreed to establish a ‘Digital Observatory’, a joint committee composed of union representatives, company officials, and representatives from Assolombarda, responsible for reviewing the progress of the ‘Leapfrog’ project.

These governance mechanisms represent an effort to institutionalize ongoing negotiation on technological matters, moving beyond a one-off agreement logic. However, their effectiveness depends on often implicit factors such as access to relevant information, availability of technical expertise, and the actual recognition of the parties’ bargaining power.

The third mechanism concerns the provision of a trial period. Two agreements – Saipem and Konecta – explicitly define a 12-month window during which the technology is applied provisionally, with the aim of assessing its effects before any permanent implementation. The underlying rationale is to introduce a temporary and reversible regime that allows for observation, learning, and, if necessary, renegotiation. Though still relatively uncommon, this type of clause addresses a precise need: to govern uncertainty without denying it. In environments

where technologies evolve rapidly and unpredictably, the notion of iterative bargaining – based on periodic assessments and adaptive capacity – emerges as a hallmark of regulatory maturity.

Taken together, these three instruments outline a potential model of dynamic bargaining, capable of balancing clarity of commitments, vigilance in enforcement, and openness to corrections in light of new evidence. It offers a response – still partial but promising – to the challenge posed by artificial intelligence: not only deciding whether to adopt it, but determining how to accompany, monitor, and renegotiate it over time.

5. Prospects and open challenges for industrial relations in the age of artificial intelligence

The analysis of collective agreements governing the adoption of AI in the workplace reveals some promising signs: collective bargaining is already capable of anticipating certain risks, addressing regulatory grey areas, and proposing original tools to steer technological innovation. While some core principles – such as data protection, transparency of purpose, and investment in training – are beginning to take root, other critical dimensions remain underexplored. To transform these initial experiments into a mature regulatory model, it is essential to strengthen participatory infrastructures and broaden the scope of bargaining.

This section identifies, on the one hand, regulatory domains that are currently marginal or altogether absent but are crucial for establishing effective and shared AI governance: prevention of algorithmic discrimination, control over data usage, regulation of digital tracking, and redistribution of the benefits associated with AI adoption. On the other hand, it outlines the systemic conditions that could enable collective bargaining to play a more meaningful role: cross-sectoral diffusion, access to technical information, and early involvement in innovation processes.

One key area where collective bargaining could take on a more assertive role is the prevention of algorithmic discrimination. Rather than correcting human biases, AI systems often risk amplifying them – particularly in contexts such as task allocation, decision-support tools, and access to opportunities (Kordzadeh and Ghasemaghahi, 2022). To date, none of the agreements analyzed directly address this concern. Yet, integrating the monitoring of algorithmic discrimination risks – whether based on gender, age, origin, or other sensitive factors – into the objectives of collective bargaining would mark a critical first step toward the ethical and proactive governance of intelligent systems. This

commitment could take the form of regular algorithmic audits, enforceable transparency criteria for automated decision-making, or – more ambitiously – the co-design of allocation logics in partnership with worker representatives.

A second, equally foundational domain concerns the very origins of artificial intelligence: data. At present, collective bargaining rarely intervenes in the conditions under which data are selected, collected, and used to train AI models. And yet, it is precisely at this early stage that the pathways of AI systems are shaped. As illustrated by the Konecta case, a key development would be the contractually binding separation between data collected during work activities and data that may be used for training purposes: what is necessary for the job cannot be automatically repurposed to enhance the system. Any reuse should be contingent upon explicit consent and a specific collective agreement.

At the same time, it is crucial to acknowledge that training is not a neutral process: in many cases, workers actively contribute to refining AI systems by testing, correcting, or labeling data. This work – currently invisible – should be formalized and remunerated, drawing inspiration from data labelling models already in use in other digital domains. In parallel, companies could be required to maintain a transparent and accessible register of algorithmic activities carried out for training purposes. A third area of concern involves the regulation of digital tracking. While some of the agreements analyzed have begun to set limits on the use of data for evaluative purposes, few have yet negotiated procedures for activating performance metrics. From a legal standpoint, it is worth noting that the current regulatory framework distinguishes between fully automated systems and tools operated by humans, imposing stricter information obligations only on the former. However, this distinction risks overlooking the more common hybrid applications in which machines assist human processes while significantly shaping their dynamics. In this context, collective bargaining does not serve to duplicate existing regulations, but rather to act as a complementary and adaptive safeguard – filling legal gaps and intervening where statutory law cannot yet operate with full effectiveness.

Thus far, the management of employment impacts resulting from AI adoption has remained peripheral to collective bargaining, typically addressed through generic clauses or deferred to future monitoring phases. A more advanced model of participatory governance should instead include structured redistributive mechanisms capable of linking potential productivity gains from technology to tangible benefits for workers – such as reduced working hours, collective incentives, or opportunities for professional development. From this perspective, the

adoption of AI should be accompanied by an explicit negotiation over the technological dividend: not only whether and to what extent value is generated, but above all how that value is distributed. Rather than being treated as a residual variable, redistribution should become a central element of the contractual framework governing innovation.

Addressing these challenges, however, requires enabling conditions that cannot be taken for granted. The first is the horizontal expansion of regulation. To prevent participation from becoming a privilege reserved for a handful of large companies, it is essential to make shared tools available even in less unionized sectors and in contexts with low levels of collective bargaining coverage. The development of framework protocols, sectoral guidelines, or negotiated codes of conduct offers a promising path to strengthening regulatory universalism without sacrificing flexibility.

The second condition is a shift towards cognitive bargaining. As long as AI remains an incomprehensible ‘black box’, the regulatory capacity of social partners will remain marginal. Although some agreements establish joint monitoring bodies, their ability to serve as spaces for genuine dialogue and contestation hinges critically on access to technical information, availability of algorithmic documentation, and adequate expertise. Institutional tools are needed to enhance the algorithmic literacy of worker representatives, build bridges between technical and union knowledge, and create infrastructures for the collective interpretation of technological processes.

The third perspective concerns the actual leverage of collective bargaining in shaping technological trajectories. In most of the cases analyzed, agreements emerge *ex post* – after the technology has already been implemented – as attempts to regulate decisions taken unilaterally at the managerial level. This significantly narrows the space for genuine *ex ante* co-determination and relegates worker representation to a predominantly reactive stance. However, although not expressions of a fully realized deliberative model, the clauses, committees, and safeguards emerging from these agreements are fragments of a political economy in the making – tools that carve out spaces for negotiation within systems frequently presented as neutral drivers of innovation and shared benefit.

What is at stake, then, is not the ideal of human-centered AI in the abstract, but the institutional conditions that make a more democratic form of technological development materially possible. If collective bargaining is to move upstream – toward influencing not just outcomes but design choices – it must be equipped with access, expertise, and enforceable rights. Absent these, the path of least resistance will remain a low-road adoption model: rapid, opaque, and unaccountable. This is

not about perfect alignment of interests. It is about building institutional friction where none exists, and using it to redirect technological change toward shared – not unilateral – ends.

References

- BATTISTI M., DUSTMANN C., SCHÖNBERG U. (2023). *Technological and organizational change and the careers of workers*, in “Journal of the European Economic Association”, 21(4), pp. 1551-1594.
- BELLOC F., BURDIN G., LANDINI F. (2023). *Advanced technologies and worker voice*, in “Economica”, 90(357), pp. 1-38.
- BONNEY J., GOLDSCHLAG N., BREUX J. (2024). *The impact of AI on the workforce: Tasks versus jobs*, in “Labour Economics”, 86, pp. 102475.
- BOWLES S. (1989), *Social Institutions and Technical Change*, in DE MATTEO M. - DIEBOLT C. - MURPHY E. (eds.), *Technological and Social Factors in Long-Term Fluctuations*, Springer, pp. 68-90.
- BRYSON A., FREEMAN R.B. (2016). *Shared Capitalism: What Is It and What Does It Do?*, in R.B. FREEMAN, J.R. BLASI, D.L. KRUSE (eds.), *Sharing the Gains of the U.S. Economy: The Past, Present, and Future of American Worker Ownership*, Harvard Business School Press, pp. 29-62.
- BURDIN G., DUGHERA S., LANDINI F., BELLOC F. (2023). *Contested transparency: Digital monitoring technologies and worker voice*, Available at SSRN 4528975.
- DAUTH W., FINDEISEN S., SUEDEKUM J., WOESSNER N. (2021). *The adjustment of labor markets to robots*, in “Journal of the European Economic Association”, 19(6), pp. 3104-3153.
- DE STEFANO V., TAES S. (2021). *Algorithmic Management and Collective Bargaining*, Brussels, ETUI.
- DOELLGAST V., WAGNER I., O’BRADY S. (2023). *Negotiating limits on algorithmic management in digitalised services: cases from Germany and Norway*, in “Transfer: European Review of Labour and Research”, 29(1), pp. 105-120.
- FORNINO M., MANERA A. (2019). *Automation and the Future of Work: Assessing the Role of Labor Flexibility*, in “Review of Economic Dynamics”, 45, pp. 282-321.
- HOLM J.R., LORENZ E. (2021). *The impact of artificial intelligence on skills at work in the digital age*, in “Industrial and Corporate Change”, 30(3), pp. 549-568.
- ILSØE A., LARSEN T.P., MATHIEU C., ROLANDSSON B. (2024), *Negotiating about algorithms: social partner responses to AI in Denmark and Sweden*, in “ILR Review”, 77(5), pp. 856-868.
- KORDZADEH N., GHASEMAGHAEI M. (2022). *Algorithmic bias: review, synthesis, and future research directions*, in “European Journal of Information Systems”, 31(3), pp. 388-409.

- KRZYZWIDZINSKI M., GERST D., BUTOLLO F. (2023), *Promoting human-centred AI in the workplace. Trade unions and their strategies for regulating the use of AI in Germany*, in "Transfer: European Review of Labour and Research", 29(1), pp. 53-70.
- MOLINA O., BUTOLLO F., GODINO A., HOLTGREWE U., ILLÈSY M., ILLSØE A., JUNTTE S., JANSEN G., LARSEN T.P., MAKÓ C., MUSTCHIN S., PAP J., VANDAELE K., WOTSCHACK P. (2023). *It takes two to code: a comparative analysis of collective bargaining and artificial intelligence*, in "Transfer: European Review of Labour and Research", 29(1), pp. 87-104.
- PÉREZ C., KANZIG D., RAMOS A. (2023). *AI diffusion and organizational restructuring: Evidence from Spanish manufacturing*, in "Research Policy", 52(4), pp. 104731.
- PRASSL J. (2019). *What If Your Boss Was an Algorithm? Economic Incentives, Legal Challenges, and the Rise of Artificial Intelligence at Work*, in "Comparative Labor Law & Policy Journal", 41(1), pp. 123-153.
- RACABI L., LITWIN A. (2024). *Varieties of AI Regulations: The United States Perspective*, in "ILR Review", 77, pp. 799-812.
- RIVA G. (2025), *Io, noi, loro. Le relazioni nell'era dei social e dell'IA*, Bologna, il Mulino.
- RIVA G., CHIRIATTI M., GANAPINI M., PANAI E., UBIALI M. (2024). *The case for Human-AI Interaction as System 0 Thinking*, in "Nature Human Behaviour", 8(10), pp. 1829-1830.
- SCHLUND R., ZITEK E.M. (2024). *Algorithmic versus human surveillance leads to lower perceptions of autonomy and increased resistance*, in "Communications Psychology", 2, 53.
- SOULAMI M., BENCHEKROUN S., GALIULINA A. (2024), *Exploring how AI adoption in the workplace affects employees: a bibliometric and systematic review*, in "Frontiers in Artificial Intelligence", 7.

2. Artificial Intelligence: A Game-Changer for Mental Health Care*

Antonios Dakanalis, Brenda K. Wiederhold, Giuseppe Riva

ABSTRACT

Starting from the escalating global burden of mental health disorders, exacerbated by the COVID-19 pandemic, the article examines the potential of artificial intelligence (AI) to revolutionize mental health care. With nearly one in five adults facing mental health issues and suicide ranking as a leading cause of death among the young, the strained mental health system seeks innovative solutions. The text discusses the rapid evolution of AI, particularly in image analysis for early physical health diagnoses, and its promising applications in mental health, including predictive analytics for various disorders. AI's ability to analyze written language, speech characteristics, and physiological signals from wearables offers avenues for remote monitoring and early prognosis. Despite the need to address ethical considerations, particularly biases in datasets and concerns about potential patient detachment, the article advocates for AI as a complementary tool rather than a replacement for human involvement in mental health services. Overall, the article emphasizes the transformative potential of AI in enhancing diagnostics, monitoring, and treatment strategies for mental health disorders.

Introduction: AI in mental health

Nearly one in five adults suffer from a mental health problem and every 40 seconds a person dies by suicide, which is the fourth leading cause of death among 15 to 30-year-olds worldwide (WHO, 2023). The spread of the 'invisible killer' enemy, the Coronavirus, has made matters worse, putting pressure on an already strained mental health system (WHO, 2023) which is looking for new ways to adapt quickly and safely to the growing demand for mental health services. For many, artificial intelligence (AI) technology and its subsets (e.g. machine learning,

* This chapter was originally published as Dakanalis A., Wiederhold B.K., Riva G. (2024), *Artificial Intelligence: A Game-Changer for Mental Health Care*, in "Cyberpsychology, Behavior, and Social Networking", 27(2), pp. 100-104. doi:10.1089/cyber.2023.0723. © Mary Ann Liebert, Inc.

robotics, computer vision and natural language processing), which will inevitably be the cornerstone of the next digital revolution, are the key (van Heerden, Pozuelo, Kohrt, 2023).

AI is rapidly evolving and is being applied to various medical specialities, helping medical doctors to reach either a diagnosis or a clinical decision faster and more accurately for more effective, personalised treatment (Haug, Drazen, 2023). Right now, the biggest application of AI systems in physical health is image analysis. For example, in a computed tomography scan, it can detect tumours in the very early stages that the human eye could not detect (Haug, Drazen, 2023). Psychiatry and the broader field of mental health lack such diagnostic tools and rely entirely on clinical examination and the experience, expertise and empathy of the practitioner. Although mental health requires perception, human interaction, intuition, and emotional intelligence that computers and robots cannot (yet) simulate, even here, AI could have an impact.

1. The advantages of AI in mental health

We can benefit from AI's ability to draw large amounts of data from various sources (e.g. multiple electronic health records, neuroimaging, genetic and speech data, data obtained by smartphones, video monitoring systems, wearable devices and social media, etc.), process and analyse it, detecting subtle warning signs that humans might never notice with a high degree of accuracy as well as signs of mental distress before they worsen (Ducharme, 2019; Khorasani, Kahani, Amin Yazdi, Hajiaghahi-Keshteli, 2023). For example, there is accumulated evidence that machine learning AI can predict mental health problems, including depression, anxiety, attention deficit hyperactivity disorder, obsessive-compulsive disorder, post-traumatic stress disorder, anorexia nervosa, schizophrenia and suicidal ideation and attempts (Graham, Depp, Lee et al., 2019; Iyortsuun, Kim, Jhon, Yang, Pant, 2023; Abdalrazaq, Alhuwail, Schneider et al., 2022) with high accuracy.

There's also evidence that machine learning AI trained to evaluate written language and particularly word choice and word order is better than human medical doctors at distinguishing between real and fake suicide notes, meaning it is good at accurately identifying signs of distress (Pestian, Nasrallah, Matykiewicz, Bennett, Leenaars, 2010). Thus, using AI systems to regularly monitor patients' writings even on social media through an app could provide mental health professionals with an important tool to assess the risk of self-harm in even the most remote patients, which is the major issue in clinical practice.

The analysis of speech characteristics such as rhythm, tone and volume, closely linked to an individual's mental state, is another promising application of AI in mental health (Dikaïos, Rempel, Dumpala, Oore, Kiefté, Uher, 2023). Monotone speech may be a sign of depression, rapid speech may indicate mania, and disjointed word choice or sentences that do not follow a logical pattern may be linked to schizophrenia. When these characteristics are pronounced enough, mental health professionals can pick up on them. Yet humans are not perfect, as they may be distracted and sometimes miss warning signs and subtle speech cues (Dikaïos, Rempel, Dumpala, Oore, Kiefté, Uher, 2023). AI algorithms could be trained to detect these signs. Further, AI and mobile devices could enable mental health professionals to track patients, categorise their mental health status and monitor subtle changes daily (Dikaïos, Rempel, Dumpala, Oore, Kiefté, Uher, 2023). As far as we know, several colleagues and large companies like IBM are working in this field (Ducharme, 2019; Dikaïos, Rempel, Dumpala, Oore, Kiefté, Uher, 2023). A speech-based mobile app has recently been designed in the University of Colorado Boulder that takes patients through a series of repetitive verbal exercises, such as telling a story and answering questions about their emotional and psychological state (Ducharme, 2019). An AI system then assesses these audio snippets for signs of mental distress, comparing them to both the individual's previous responses and those of a wider population of patients. The mobile app was found to be at least as accurate as mental health specialists in identifying signs of mental distress based on speech (Ducharme, 2019). This technology is not commercially available (Ducharme, 2019). However, the impact of technological devices such as this can be enormous given that many people with mental illness live in remote areas where access to mental health professionals (i.e. psychologists, psychiatrists and psychotherapists) is scarce (Dikaïos, Rempel, Dumpala, Oore, Kiefté, Uher, 2023) and that rates of mental health distress (e.g. depression, anxiety) are high in the workplace, with mental distress now accounting for most of the sickness absence in Europe and USA (Pecchioli, Lomartire, Valente, Pecchioli, Germanò, 2023).

Wearables incorporating AI offer opportunities for proactive mental care in and outside the workplace. Rather than waiting for a user to interact with them via an app, some cutting-edge AI mental health solutions operate as wearable devices equipped with sensors, analysing bodily signals and giving timely warning that users should make changes so that they feel better (Ducharme, 2019). For example, Biobeat (Pecchioli, Lomartire, Valente, Pecchioli, Germanò, 2023) can collect information on sleeping patterns, fluctuations in physical activity and

variations in heart rate and rhythm which are used to assess the cognitive and psychological/emotional state of the user. These data are compared with anonymised and aggregated data from other users to provide predictive warnings that mental wellbeing might be about to decline. Users can then adjust their behaviour or environments to prevent significant deterioration in mental health and wellbeing or ask for professional help if necessary (Pecchioli, Lomartire, Valente, Pecchioli, Germanò, 2023).

Importantly, AI systems can be used for early prognosis, i.e. to predict how the patient will fare and whether or not they will respond to a particular type of psychotherapy or a particular medication. For example, AI has been used to predict cases where patients are more likely to respond to cognitive behavioural therapy and therefore be less likely to require antidepressant drugs, which can have side effects (Ewbank, Cummins, Tablan et al., 2020). In other studies, AI algorithms can analyse the brain activity of depressed patients or real clinical data and electronic health records to accurately identify which depressed patients will not respond adequately or at all to existing antidepressants (Pigoni, Delvecchio, Madonna, Bressi, Soares, Brambilla, 2029). The early identification of these patients suffering from so-called drug-resistant depression and at increased risk of suicide, among other things, is extremely important as it can contribute to the optimal strategic management of these cases (e.g. using different treatment methods from the commonly used ones, e.g. intranasal esketamine) helping us to intervene in time before a situation becomes more severe and saving resources both in human and financial terms (Capuzzi, Caldiroli, Capellazzi, Tagliabue, Marcatili, Colmegna, Clerici, Buoli, Dakanalis, 2021).

The creation and use of a huge amount of data about the brain from neuroimaging, genetics, cognitive assessments and now smartphones and wearable devices, combined with AI, will also help significantly in identifying people at high risk of developing a mental health disorder in their lifetime and creating new and more personalised treatments (Khorasani, Kahani, Amin Yazdi, Hajiaghahi-Keshteli, 2023) based on a patient's unique characteristics, which is the biggest challenge in the field of mental health. Additionally, mobile apps incorporating AI will allow mental health professionals to monitor mental health problems in patients and whole populations remotely (Ducharme, 2019), alerting them to issues that are an obstacle to treatment adherence by patients or behavioural or symptomatic changes that arise between appointments, and helping them to incorporate that important knowledge into the treatment plans.

As the worldwide need for mental health and wellbeing care

continues to escalate, the imperative to effectively address this surge has heightened (WHO, 2023). There has been a notable embrace of digital solutions in the realm of mental health to meet the growing demand. These include applications designed to provide insights into sustaining wellbeing and handling emotional strain, virtual support communities, and real-time access to human therapists via the internet. Notably, virtual reality therapy is also making significant strides in this evolving landscape (Riva, Wiederhold, Mantovani, 2019). Additionally, mental health conversational agents, also known as chatbots, combining AI and cognitive behavioural therapy principles have been developed (Vaidyam, Linggonegoro, Torous, 2021). This is a type of software that allows the machine to talk like a human being, providing automated answers to questions and requests, as far as technologically possible. In the healthcare sector, chatbots already provide patients with 24/7 access to medical information, make doctor's appointments, answer questions about their health and provide them with general treatment instructions or even personalised health advice.

KokoBot, Woebot and Earkick are currently the most well-known and widely used automated mental health chatbots (Vaidyam, Linggonegoro, Torous, 2021). The users describe how they feel via text or voice message. It takes about 15 seconds for the AI algorithm they embed to analyse the message and give the user personalised advice on what they can do to reduce a host of symptoms such as anxiety and depression and improve their mental wellbeing. The idea is revolutionary as it could make treatment more affordable and accessible to all, important for countries, environments and people with limited resources. However, questions have emerged regarding their efficacy. Do they truly deliver on their promises and are they genuinely beneficial for individuals? A recent review of the limited chatbot literature revealed few but generally positive outcomes regarding chatbots' therapeutic efficacy and acceptability (Vaidyam, Linggonegoro, Torous, 2021). Yet important methodological issues have also emerged. For example, most of the available studies evaluating mental health chatbots are small, without a control group to compare results and are difficult to generalise to patients with mental disorders (Vaidyam, Linggonegoro, Torous, 2021). It will most likely take at least a few years before algorithms are routinely used in clinical practice (Ward, 2021). While we almost all agree that digital innovations that can be delivered in a non-intrusive, scalable, accessible, and cost-effective way are needed, many practitioners are afraid that as the technology improves, chatbots or smartphone apps will turn patients away from human therapists. But we do not feel that this is the case. As more people use apps that are an introduction to mental wellbeing

care, there will be increased awareness and further interest in seeking in-person care from human therapists. Indeed, companies are trying to hire more psychiatrists, psychologists and psychotherapists to meet the rising need for clinicians supporting apps (Ward, 2021). The future of technology in mental health is not about replacing people but rather supporting them. Humans have emotion and empathy, which AI currently lacks and is vital in the effective delivery of mental health services.

The seamless integration of artificial intelligence (AI) with virtual reality (VR) and cutting-edge digital technologies introduces a revolutionary tool for advancing evidence-based psychological treatment (Riva, Wiederhold, Di Lernia, Chirico, Riva, Mantovani, Cipresso, Gaggioli, 2019; Wiederhold, Riva, 2022): digital biomarkers. Departing from traditional physiological indicators, these biomarkers assume a digital guise – quantifiable data harnessed through digital technologies to not only predict but also wield influence over health outcomes (Riva, Wiederhold, Di Lernia, Chirico, Riva, Mantovani, Cipresso, Gaggioli, 2019).

While various digital devices, spanning from smartphones to smartwatches, can contribute to the collection of digital biomarkers, the unparalleled capabilities of AI find a symbiotic fit with VR. Key to this synergy is VR's unique attribute: the absolute control it exerts over both explicit (movements, speech) and implicit (emotions, cognitive states) behaviors experienced by users. Moreover, the collection of these digital biomarkers remains inconspicuous for users, facilitated by embedded sensors within VR tools and input/output devices.

The strategic utilization of VR and digital technologies enables direct and comprehensive collection and storage of a myriad of digital biomarkers. These encompass explicit behaviors (e.g., movements, speech) and implicit indicators (e.g., facial expressions, eye movements), as well as various physiological states (e.g., heart rate, EEG data). In essence, the amalgamation of VR and digital technologies furnishes AI with a diverse array of powerful digital biomarkers, specifically tailored to complex and controlled behaviors. This convergence facilitates advanced prediction and personalized therapeutic approaches, marking a transformative leap in psychological treatment.

In conclusion, AI offers significant advantages that can improve the current mental health scenario (see Table 1).

Table 1 - *The main advantages offered by AI to mental health*

<i>Advantage</i>	<i>Description</i>
EARLY DETECTION	AI enables early detection of mental health issues through the analysis of diverse data sources, predicting conditions like depression, anxiety, and suicidal tendencies with high accuracy.
REMOTE MONITORING	Wearables and mobile apps equipped with AI allow continuous remote monitoring of patients, providing real-time insights into their mental health status and facilitating timely interventions.
PERSONALIZED TREATMENT PLANS	AI applications analyze vast datasets, including genetic, neuroimaging, and behavioral data, to tailor personalized treatment plans, optimizing responses to psychotherapy and medication.
PREDICTIVE ANALYTICS	AI-driven predictive analytics assist in forecasting patient outcomes, treatment responses, and the risk of developing mental health disorders, aiding clinicians in strategic management.
ACCESSIBILITY AND AFFORDABILITY	AI-powered mental health chatbots and applications enhance accessibility to care, providing automated support, advice, and interventions, particularly beneficial for individuals in remote or underserved areas.

2. *The challenges of AI in mental health*

While AI presents compelling advantages for enhancing the mental health landscape addressing these opportunities also requires navigating several challenges (see Table 2).

Table 2 - *The main challenges to be solved for a wider use of AI in mental health*

<i>Challenge</i>	<i>Description</i>
ETHICAL CONSIDERATIONS	Ensuring ethical use of AI in mental health involves addressing issues such as privacy, consent, and potential biases in algorithms, requiring comprehensive guidelines and regulatory frameworks.
DATA QUALITY AND BIAS	Biases in training data can lead to inaccuracies, particularly in diverse populations. Ensuring data quality and mitigating biases are crucial to enhance the reliability of AI-driven mental health applications.
INTEGRATION INTO CLINICAL PRACTICE	Bridging the gap between AI research and clinical practice remains a challenge, necessitating efforts to incorporate AI tools seamlessly into existing mental health care workflows and protocols.

(follows)

<i>Challenge</i>	<i>Description</i>
PATIENT ACCEPTANCE AND TRUST	Gaining patient acceptance and trust in AI-driven mental health interventions is essential. Addressing concerns about reliance on technology over human interaction requires effective communication and education.
COMPREHENSIVE EVALUATION	The comprehensive evaluation of AI technologies in real-life clinical contexts, including their impact on patient outcomes, is vital for wider adoption. Robust studies with control groups are needed for accurate assessment.

A paramount concern is the imperative to bridge the existing gap between AI research in mental health and its practical application in clinical settings (Graham, Depp, Lee et al., 2019). Despite the increasing emergence of AI-based applications in mental health care, their seamless integration into real-world clinical practices remains limited (Sharma, Savage, Nair, Larsson, Svedberg, Nygren, 2022). The integration process necessitates comprehensive evaluations of these technologies in authentic clinical contexts, incorporating considerations of both patient and clinician acceptance – a crucial challenge that must be overcome for widespread adoption.

Moreover, a critical focus must be directed towards a deeper understanding of the ethical and privacy implications associated with AI healthcare applications (Riva, Sajno, De Gaspari, Pupillo, Wiederhold, 2023; Riva, Wiederhold, Succi, 2022). Ongoing research efforts should strive to unravel these complexities and establish robust guidelines for the seamless and successful integration of AI technologies into everyday clinical practices. According to the World Health Organization, there are still flaws in how existing AI applications process data and insufficient evaluation of the risks around bias, i.e., imbalances in the datasets used to train AI algorithms (WHO, 2023; Riva, Wiederhold, Succi, 2022). It is known, for example, that mental health issues are more likely to go undiagnosed among people with poor access to healthcare and/or ethnic groups. AI algorithms depending on this data may exhibit reduced accuracy in diagnosing such issues (WHO, 2023). These considerations underscore the need for a multifaceted approach to ensure the ethical, privacy, and practical dimensions of AI in mental health are thoroughly addressed for its optimal implementation.

3. *Conclusions*

In conclusion, the evolution of artificial intelligence (AI) is a valuable tool for enhancing public mental health. AI functions as an

indispensable asset by acquiring and analyzing data to aid practitioners, rather than replacing them. It is our responsibility as professionals to provide credible and unbiased information, while also utilizing the potential of digital biomarkers which comprises quantifiable data obtained through digital technologies that can be utilized not only to predict but also to influence health outcomes. The maxim “garbage in, garbage out” holds especially true in this regard, underscoring the need for conscientious data curation.

Instead of debating the hypothetical redundancy of humans due to advanced AI, we should concentrate on forming a future where AI enhances human abilities in mental health evaluation and therapy. This requires addressing pivotal issues like data quality, the possibility of bias, and the potential exposure of citizens’ confidential and sensitive information. It is crucial to establish a robust ethical and secure framework for the implementation of AI. By doing this, we can alleviate concerns regarding the gradual replacement of the human element, a concern that has been witnessed in various fields that use AI, such as healthcare.

As we pursue this, it becomes clear that the personal touch in mental health services is irreplaceable. AI is, at its essence, a tool intended to complement and elevate human abilities, rather than substitute for the refined qualities of empathy, intuition, and emotional intelligence inherent in human interactions. By navigating the ethical and privacy considerations surrounding AI, we can leverage its potential to enhance mental healthcare, promoting a mutually beneficial relationship between technology and human expertise. This comprehensive approach not only protects against possible drawbacks but also guarantees that AI supports our shared effort to improve mental health results. As we move forward, the focus is not on humans vs. machines, but rather on humans utilizing machines to establish a more sympathetic and efficient mental health environment.

References

- ABD-ALRAZAQ A., ALHUWAIL D., SCHNEIDER J. ET AL. (2022). *The performance of artificial intelligence-driven technologies in diagnosing mental disorders: an umbrella review*, in “npj Digital Medicine”, 5, 87. <https://doi.org/10.1038/s41746-022-00631-8>
- CAPUZZI E., CALDIROLI A., CAPELLAZZI M., TAGLIABUE I., MARCATILI M., COLMEGNA F., CLERICI M., BUOLI M., DAKANALIS A. (2021). *Long-Term Efficacy of Intranasal Esketamine in Treatment-Resistant Major Depression: A Systematic Review*, in “International Journal of Molecular Sciences”, Aug 28, 22 (17), 9338. doi: 10.3390/ijms22179338. PMID: 34502248; PMCID: PMC8430977.

DIKAIOS K., REMPEL S., DUMPALA S.H., OORE S., KIEFTE M., UHER R. (2023). *Applications of Speech Analysis in Psychiatry*, in "Harvard Review of Psychiatry", Jan-Feb 01, 31(1), pp. 1-13. doi: 10.1097/HRP.0000000000000356. Erratum in "Harvard Review of Psychiatry", 2023 Mar-Apr 01; 31(2), 96. PMID: 36608078.

DUCHARME J. (2019). *Artificial Intelligence Could Help Solve America's Impending Mental Health Crisis*, in "TIME", 20 Nov., <https://time.com/5727535/artificial-intelligence-psychiatry/>

EWBANK M.P., CUMMINS R., TABLAN V. ET AL. (2020). *Quantifying the Association Between Psychotherapy Content and Clinical Outcomes Using Deep Learning*, in "JAMA Psychiatry", 77(1), pp. 35-43. doi:10.1001/jamapsychiatry.2019.2664

GRAHAM S., DEPP C., LEE E.E. ET AL. (2019). *Artificial Intelligence for Mental Health and Mental Illnesses: An Overview*, in "Current Psychiatry Reports", 21, 116. <https://doi.org/10.1007/s11920-019-1094-0>

HAUG C.J., DRAZEN J.M. (2023). *Artificial Intelligence and Machine Learning in Clinical Medicine*, in "New England Journal of Medicine", Mar 30, 388(13), pp. 1201-1208. doi: 10.1056/NEJMr2302038. PMID: 36988595

IYORTSUUN N.K., KIM S.H., JHON M., YANG H.J., PANT S. (2023). *A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis*, in "Healthcare" (Basel), Jan 17, 11(3), 285. doi: 10.3390/healthcare11030285. PMID: 36766860; PMCID: PMC9914523.

KHORASANI M., KAHANI M., AMIN YAZDI S.A., HAJIAGHAEI-KESHTALI M. (2023). *Towards finding the lost generation of autistic adults: A deep and multi-view learning approach on social media*, in "Knowledge-Based Systems", 276, 110724, ISSN 0950-7051, <https://doi.org/10.1016/j.knosys.2023.110724>.

PECCHIOI V., LOMARTIRE N., VALENTE L., PECCHIOI L., GERMANO G. (2023). *Comparison Study for 24h Ambulatory Blood Pressure Measurement (Abpm) with Biobeat Patch Abpm® Cuffless Wearable Device in Subjects with Different Body Mass Index*, in "Journal of Hypertension", 41(1), e247-e248.

PESTIAN J., NASRALLAH H., MATYKIEWICZ P., BENNETT A., LEENAARS A. (2010). *Suicide Note Classification Using Natural Language Processing: A Content Analysis*, in "Biomedical Informatics Insights", Aug 4, 2010(3), pp. 19-28. doi: 10.4137/bii.s4706. PMID: 21643548; PMCID: PMC3107011.

PIGONI A., DELVECCHIO G., MADONNA D., BRESSI C., SOARES J., BRAMBILLA P. (2019). *Can Machine Learning help us in dealing with treatment resistant depression? A review*, in "Journal of Affective Disorders", Dec 1, 259, pp. 21-26. doi: 10.1016/j.jad.2019.08.009. Epub 2019 Aug 13. PMID: 31437696.

RIVA G., SAJNO E., DE GASPARI S., PUPILLO C., WIEDERHOLD B.K. (2023). *Navigating the Ethical Crossroads: Bridging the gap between Predictive Power and Explanation in the use of Artificial Intelligence in Medicine*, in "Annual Review of CyberTherapy and Telemedicine", 21, pp. 3-7.

RIVA G., WIEDERHOLD B.K., DI LERNIA D., CHIRICO A., RIVA E.F.M., MANTOVANI F., CIPRESSO P., GAGGIOLI A. (2019). *Virtual Reality meets Artificial Intelligence: The Emergence*

of *Advanced Digital Therapeutics and Digital Biomarkers*, in “Annual Review of CyberTherapy and Telemedicine”, 18, pp. 3-7.

RIVA G., WIEDERHOLD B.K., MANTOVANI F. (2019). *Neuroscience of Virtual Reality: From Virtual Exposure to Embodied Medicine*, in “Cyberpsychology, Behavior, and Social Networking”, 22(1), pp. 82-96. doi:10.1089/cyber.2017.29099.gri

RIVA G., WIEDERHOLD B.K., SUCCI S. (2022). *Zero Sales Resistance: The Dark Side of Big Data and Artificial Intelligence*, in “Cyberpsychology, Behavior, and Social Networking”, 25(3), pp. 169-173. doi:10.1089/cyber.2022.0035

SHARMA M., SAVAGE C., NAIR M., LARSSON I., SVEDBERG P., NYGREN J.M. (2022). *Artificial Intelligence Applications in Health Care Practice: Scoping Review*, in “Journal of Medical Internet Research”, Oct 5, 24(10), e40238. doi: 10.2196/40238. PMID: 36197712; PMCID: PMC9582911.

VAIDYAM A.N., LINGGONEGORO D., TOROUS J. (2021). *Changes to the Psychiatric Chatbot Landscape: A Systematic Review of Conversational Agents in Serious Mental Illness; Changements du paysage psychiatrique des chatbots: une revue systématique des agents conversationnels dans la maladie mentale sérieuse*, in “The Canadian Journal of Psychiatry”, 66(4), pp. 339-348. doi:10.1177/0706743720966429

VAN HEERDEN A.C., POZUELO J.R., KOHRT B.A. (2023). *Global Mental Health Services and the Impact of Artificial Intelligence-Powered Large Language Models*, in “JAMA Psychiatry”, 80(7), pp. 662-664. doi:10.1001/jamapsychiatry.2023.1253

WARD L. (2021). *Can Artificial Intelligence Replace Human Therapists?*, in “Wall Street Journal”, 27 March, <https://www.wsj.com/articles/can-artificial-intelligence-replace-human-therapists-11616857200>.

WHO (2023). *Artificial intelligence in mental health research: new WHO study on applications and challenges*, Retrieved on November 16, 2023 from <https://www.who.int/europe/news/item/06-02-2023-artificial-intelligence-in-mental-health-research-new-who-study-on-applications-and-challenges>

WIEDERHOLD B.K., RIVA G. (2022). *Metaverse Creates New Opportunities in Healthcare*, in “Annual Review of CyberTherapy and Telemedicine”, 20, pp. 3-7.

OPEN QUESTIONS

Human-AI Interaction in Radiology

Evis Sala

What is the next frontier for AI-powered diagnostics, beyond radiology and pathology?

I believe the next frontier in AI-powered diagnostics lies in ‘superdiagnostics’, a concept that goes far beyond the current applications of AI in radiology and pathology. Today’s imagers – radiologists, pathologists, and other diagnostic specialists – are often seen as interpreters of specific data sets, like images or tissue samples. But soon, they will evolve into ‘superdiagnosticians’, professionals who rather than interpreting isolated data, help create a neat streamline of clinical information. To use a metaphor, they will no longer be soloists playing a single instrument – they will become conductors of a complex orchestra of data. This shift will require a fundamental transformation in how we view the role of AI in diagnostics. Up until now, AI has primarily been used to improve efficiency – automating repetitive tasks, prioritizing cases, and assisting with reporting. These are valuable contributions, but they only scratch the surface of what is possible. The real promise of AI lies in its potential to synthesize and correlate vast, diverse data streams – imaging, patient history, clinical symptoms, lab results, genetic and molecular biomarkers – into a coherent, actionable picture of a patient’s health.

It would be beyond imagination to have a system in which AI does not just flag abnormalities in a scan, but also cross-references those findings with the patient’s medical history, recent blood work, genomic data, and even lifestyle factors. It could then suggest possible diagnoses, recommend further tests, or even propose personalized treatment pathways. This kind of integrative, predictive capability is what I mean by superdiagnostics – a level of insight that no single human specialist could achieve alone, but which becomes possible when AI acts as a true partner in clinical reasoning. To get there, we need to move beyond the current approach to data. Right now, imaging data, lab results, and clinical notes often live in separate systems, making it difficult to draw meaningful connections. AI can help bridge these gaps, but only if we design systems that allow for seamless data integration and interoper-

ability. This will require collaboration across disciplines – clinicians, data scientists, engineers, and healthcare administrators all working together to build the infrastructure and governance needed for trustworthy, explainable AI.

Another key aspect of this future is the evolving role of the imager. As AI takes on more of the routine interpretive work, human experts will be freed to focus on higher-order tasks: synthesizing insights, communicating with care teams, and guiding clinical decision-making. In this new paradigm, imagers become not just diagnosticians, but data managers and insight orchestrators. They will need new skills – data literacy, systems thinking, and a deep understanding of AI's capabilities and limitations. Of course, this transformation will not happen overnight. There are significant challenges to overcome, including data privacy concerns, regulatory hurdles, and the need for robust validation of AI tools in real-world settings. But the direction is clear. Just as imaging departments led the digital transformation of healthcare in the 1990s – pioneering the use of PACS systems and digital workflows – they are now poised to lead the next revolution: the AI transformation. It is also important to note that while imaging is currently at the forefront due to the sheer volume and complexity of its data, other specialties will follow. Fields like cardiology, oncology, and even primary care are beginning to explore how AI can enhance diagnostic accuracy and personalize care. But imaging remains unique in its centrality to so many clinical pathways, making it the ideal proving ground for these next-generation tools.

Ultimately, the goal is not to replace clinicians, but to empower them. AI should be seen as an extension of human intelligence – a tool that augments our ability to understand disease, predict outcomes, and deliver better care. The future of diagnostics is not man versus machine, but man and machine working together in harmony. And that future is closer than we think.

What are the biggest lessons you have learned from integrating AI tools into routine clinical practice in radiology?

One of the most important lessons I have learned is that, despite the excitement and promise surrounding AI in healthcare, most AI-based tools still struggle to achieve full adoption and meaningful integration into routine clinical workflows. This is not due to a lack of innovation or potential – there are many brilliant algorithms out there – but rather because of two persistent and fundamental issues.

The first is the lack of clinical input during the development and

validation phases. Too often, AI tools are created in isolation by data scientists or engineers who may not fully understand the nuances of clinical practice. Without the involvement of radiologists and other healthcare professionals from the very beginning, these tools risk being misaligned with real-world needs. They may solve theoretical problems or perform well in controlled environments, but they fall short when applied to the messy, complex, and time-sensitive reality of clinical care.

The second major barrier is the absence of a robust commercialization and integration strategy. Even when an AI tool is clinically sound and technically impressive, it will not succeed unless there is a clear plan for how it will be deployed, scaled, and maintained in actual healthcare settings. This includes everything from regulatory approval and IT integration to user training and workflow redesign. Without this infrastructure, even the best algorithms remain stuck in pilot projects or academic papers. What we have seen is that AI tools co-developed with radiologists, for radiologists, and with a clear implementation roadmap from the outset, are far more likely to be adopted and used consistently. When the right stakeholders – clinicians, hospital administrators, IT teams, and even patients – are involved early on, the resulting solutions are more practical, more trusted, and more impactful.

Another key lesson is that AI cannot and should not attempt to replace the human element of radiology. Patient care is not just a matter of analyzing imaging biomarkers or running mathematical models. It is about understanding the patient as a whole – contextualizing findings, weighing uncertainties, and communicating with other members of the care team. These are deeply human tasks that require empathy, judgment, and experience. That said, AI is already proving to be a powerful ally in enhancing efficiency. Tools that help with case triage, image segmentation, or report generation are being embraced because they free up radiologists to focus on more complex and value-added tasks. And this is where a fascinating shift is happening: AI is not diminishing the role of the radiologist – it is elevating it. By automating routine processes, AI allows radiologists to step into a more central role in clinical decision-making. They are becoming not just image interpreters, but clinical consultants and data integrators, helping to guide diagnosis and treatment across specialties. This evolution is making the radiologist's contribution to patient outcomes more visible and more essential than ever before.

It is also worth reflecting on the famous 2016 prediction by Geoffrey Hinton, who suggested that AI would replace radiologists within five years. That clearly has not happened – and in fact, the opposite is proving true. What we are seeing now is that radiologists who can

effectively collaborate with AI will replace those who cannot. The future belongs to those who are willing to adapt, to learn new tools, and to embrace a hybrid model of human-machine collaboration.

This does not mean that every radiologist needs to become a data scientist. But it does mean developing a working understanding of how AI systems function, what their limitations are, and how to interpret their outputs critically. It also means being proactive in shaping how these tools are used – advocating for ethical standards, ensuring transparency, and keeping the focus on patient-centered care.

In summary, the biggest lessons from integrating AI into radiology are:

- co-development is critical: tools must be built with direct input from clinicians to ensure relevance and usability;
- implementation matters: without a clear strategy for integration, even the best tools will fail to scale;
- AI enhances, not replaces: the radiologist’s role is evolving, not disappearing;
- collaboration is the future: success will come to those who can work alongside AI, not against it;
- patient care remains the priority: technology must always serve the goal of improving outcomes and experiences for patients.

These insights are shaping how we approach AI in radiology today – and they will continue to guide us as we move toward a more intelligent, integrated, and human-centered future in healthcare.

How can we ensure AI systems are designed to avoid amplifying existing health disparities, particularly in diagnostics?

This is a crucial question, and one that deserves far more attention than it often receives. At its core, we must remember that AI is just a tool – a powerful one, yes, but still a tool. Like any tool, it reflects the intentions, assumptions, and limitations of the people who design and deploy it. If we are not careful, AI can unintentionally amplify existing health disparities – by reinforcing biases in data, excluding underrepresented populations, or being deployed in ways that favor already well-resourced healthcare systems.

One of the most important safeguards is ensuring that AI algorithms are trained and validated on diverse populations. This means diversity not only in terms of ethnicity and gender, but also geography, socioeconomic status, healthcare infrastructure, and disease prevalence.

Too many AI models are developed using data from a narrow slice of the global population – often from high-income countries, urban hospitals, or academic centers. When these models are applied elsewhere, their performance can degrade significantly, leading to misdiagnoses or missed diagnoses in already underserved communities.

To address this, we need to build datasets that are representative of the real-world diversity of patients. This requires intentional effort – collaborating across borders, investing in data collection in low- and middle-income countries, and ensuring that marginalized groups are not left out of the AI development process. It also means being transparent about the limitations of a model: where it was trained, how it was validated, and what populations it may or may not generalize to.

Regulators have a key role to play here. I believe that regulatory bodies should require evidence of performance across diverse populations as part of the approval process for AI tools in healthcare. Just as we demand that new drugs be tested in a range of patient groups, we should hold AI to the same standard. Evidence-based AI must, by definition, be diversified AI.

But it is not all cautionary. In fact, AI has the potential to reduce disparities, if we design and deploy it thoughtfully. This was a major theme during the discussions around the World Health Organization's Resolution on "Strengthening Medical Imaging Capacity", which was adopted at the 78th World Health Assembly. The resolution explicitly encourages member states to consider integrating technologies like telehealth, teleradiology, clinical decision support systems, and AI into their radiology infrastructure. The goal is to expand access to quality diagnostics, especially in regions where there is a shortage of trained professionals or advanced equipment.

To quote directly from the resolution: it urges countries "to consider integrating technological advances, such as telehealth, teleradiology, clinical decision support (CDS), artificial intelligence and specific software applications, into radiology information systems for patient records, in compliance with applicable international standards and protocols for medical imaging, including ethical aspects, security, and confidentiality of data"¹.

This is a powerful endorsement of AI's potential to bridge gaps in access, not just widen them. For example, AI-powered teleradiology platforms can bring expert-level diagnostic support to rural clinics that lack on-site radiologists. Clinical decision support tools can help general practitioners in low-resource settings make more accurate diagnoses,

¹ Seventy-eighth World Health Assembly, agenda item 13.3. May 27th, 2025, Geneva, Switzerland.

even without specialist training. And AI can help standardize care, reducing variability and ensuring that all patients – regardless of where they live – receive high-quality, evidence-based recommendations.

Of course, realizing this potential requires more than just technology. It requires investment in infrastructure, training for healthcare workers, and policies that promote equitable access. It also requires a commitment to ethical design principles – ensuring that AI systems are transparent, explainable, and accountable. Patients and clinicians need to understand how these tools work, what their limitations are, and how to challenge or override them when necessary.

Another important consideration is community engagement. We need to involve patients, especially those from historically marginalized groups, in the design and evaluation of AI systems. Their perspectives can help identify blind spots, build trust, and ensure that the tools we create truly serve the needs of all populations.

In summary, avoiding the amplification of health disparities through AI requires a multi-pronged approach:

- diverse training data: Ensure algorithms are developed using datasets that reflect the full spectrum of patient populations;
- regulatory oversight: Require evidence of equitable performance as part of AI approval and certification processes;
- global collaboration: Support data sharing and AI development across countries and healthcare systems;
- infrastructure and education: Invest in the tools and training needed to deploy AI effectively in underserved areas;
- ethical design: Prioritize transparency and patient safety in all AI applications;
- community involvement: Engage patients and frontline clinicians in shaping how AI is used in their care.

If we get this right, AI can be a force for equity in healthcare – not just a technological advancement, but a catalyst for more inclusive, accessible, and just medical systems.

SECTION 4

Legal, Regulatory, and Policy Frameworks

1. Humane AI Law: a Manifesto¹

Barbara L. Boschetti

ABSTRACT

The paper aims to outline the five key regulatory strategies for making the law fit for Humane AI. A kind of Humane-centric ‘Manifesto’ in five points for the future of AI and Law.

A manifesto ought to be a good way to grasp Humane AI by the hand (incidentally, the etymology of ‘manifesto’) and guide the future of law towards it². To this aim, this manifesto contains a list of five proposals for guiding the development of Humane AI Law.

1. *Keep the Humane in the AI/Law Loop*
2. *Keep AI in the Ethics/Law Loop*
3. *Keep AI in the Rights Loop (not vice versa)*
4. *Keep Matter in the AI/Law Loop*
5. *Keep the Loop Transformative*

1. *Keep the Humane in the AI/Law Loop: Humane AI needs us to bring back the Humane to AI and the Law*

AI and the law have much in common. They are both artificial (i.e. man-made), virtual (extensions of reality, having the potential to change it³), technologies (tools to carry out tasks, including decision

¹ Floridi, 2014.

² In legal studies, AI implications have led to a flourishing and vast literature which this work acknowledges and values greatly. In the economy of this short paper it won’t be possible to cite them all. For main references and citations I refer to my previous papers on the subject.

³ It is worth noting here that *virtual* comes from the Latin *virtus* and defines the capability to influence by virtues and inherent natural qualities. The expression ‘by virtue of’ still carries this meaning. In the legal sphere, this capability can be easily connected to the ‘ought-to-be’ of norms and to their efficacy to shape behaviour, or normativity. See Ratti, 2017.

making⁴), codes (as they operate mainly through norms/instructions⁵) and institutions (as they both aim to organize and infrastructure the world). Furthermore, they both are human, albeit not necessarily humane⁶.

Despite being a human technology developed over centuries of constitutional battles and milestones to pursue human happiness and the common good⁷, the law can be quite easily diverted and become instrumental to bad government⁸. The continuous rise of illiberal, paradoxically totalitarian democracies, both inside and outside of the EU⁹, proves that the law continues to exist in a constant precarious swing between the human and the humane.

The same goes for AI, which, being a human technology in its turn, may be developed in such a way to weaken humans and not to serve the cause of humanity, thus becoming inhumane. And yet, with its illusive dominance over intelligence and agency¹⁰, AI challenges humans, as intelligent agents¹¹, on their own territory, casting over them a sort of spell of deference towards it (the myth of Prometheus rewound, as AI idols appear to be in full control of human intelligence and generativity¹²), and significantly contributes to depriving the humane of meaning, as humans fall into its performative trap (humane intelligence reduced to performativity¹³).

⁴ Sartea, 2024.

⁵ The relationship between AI and law is also a coding competition, which also considers their respective legitimacy. Brown - Marsden, 2013.

⁶ Coupland, 2003.

⁷ The concept of happiness has been complemented with that of individual and social wellbeing. The pursuit of happiness is formally proclaimed by certain Constitutions as the ultimate goal of political power and the law itself: see the Preamble to the U.S. Constitution or to the French Declaration of the Rights of Man and of the Citizen. On power and happiness in the digital age, see also Harari, 2017.

⁸ The reference is to the famous frescoes *The Allegory of Good and Bad Government* by Ambrogio Lorenzetti, in Siena, Palazzo Pubblico.

⁹ On the rise of illiberal democracies, Campati, 2024; on illiberal democracies in the EU, Halmá, 2024; on inverted totalitarianism, see Wolin, 2010.

¹⁰ Floridi, 2023.

¹¹ On human/humane intelligence see Dicastero per la Dottrina della Fede e del Dicastero per la Cultura e l'Educazione, *Antiqua et Nova, Note on the Relationship Between Artificial Intelligence and Human Intelligence*. See also Faggini, 2024.

¹² The name Artificial Intelligence contributes significantly to the narrative. See Narayanan and Kapoor, 2024.

¹³ Cacciari, 2020; Maffettone, 2023; Giaccardi, Magatti, 2025.

The truth is that benevolent¹⁴, humane AI requires many urgent and grounding decisions to be made, a sort of humane AI constitutional moment implying much *agency* on the human side and much humane intelligence¹⁵. The law governing humane AI, and the ethical and political choices it mirrors, is – or, at least, can be – a significant part of this.

In this respect, the main issue is to bridge the gap between the human and the humane in both AI and the law. In fact, as far as AI is concerned, there are no reserves, no safe harbours and both the human and the humane are no exception: AI reshapes the ontology of things also by separating the humane from the human¹⁶. If this is true in general, it is even more true with respect to the law: the new AI-shaped ontology impacts on legal constructs, such as legal persons and fundamental freedoms and rights, legal principles, such as equality or the rule of law, and legal institutions, such as democracy, pulling the rug out from under their feet. Nothing in law can be abstracted from the AI-reshaped reality.

It follows that paradigms and regulatory approaches – such as ‘*the man in the loop*’ or the ‘*rights-driven*’ approaches – fail, or prove to be inadequate, precisely when they rely on the human without taking into account the humane, in the illusion that the two still perfectly match. In other words, the *human-in-the-loop* (HILT) paradigm only works when the AI systems are truly designed for human(e) collaboration and humans are properly prepared for AI collaboration¹⁷; and the rights-driven approach only when AI systems are not designed to replace humans¹⁸, and humans are not treated simply as AI product consumers or users¹⁹.

In conclusion, in humane AI law, human-centric should always equate to humane-centric²⁰. This is the best way to break the AI deference spell and avoid the performative trap, which humans and the law itself may fall into. After all, it is in man and woman, *vir*, that virtue and virtual have their common denominator.

¹⁴ Benevolence and non-malevolence are two of the five grounding principles of AI Ethics according to Floridi, 2023, pp. 96-97. See also Harari, 2017.

¹⁵ On the constitutional issues posed by AI economic powers, Pollicino, 2023b; Pollicino, 2023a. See the *Call for International binding Red Lines to prevent unacceptable AI risks*, available at: <https://red-lines.ai/> launched at the UN Assembly on September 22nd 2025.

¹⁶ On the digital ‘cut and paste’ reality, Floridi, 2017.

¹⁷ On the different forms of human/AI interaction and on the conditions for human/AI collaboration, Tschider, 2024.

¹⁸ See proposal no. 3.

¹⁹ Despite its rights-driven approach, the EU AI Act, Reg. EU 2024/1689, qualifies AI technologies as market products and takes a classic risk-based regulatory approach.

²⁰ EU AI Act, cit., preamble (1).

2. Keep AI in the Ethics/Law Loop: Humane AI needs us to establish a strong alliance between (AI) Ethics and Law and keep AI in the Ethics/Law loop

A second key point is the following: humane AI requires a strong alliance between (AI) ethics and law and for AI to be always kept in the ethics/law loop²¹.

The ethics/law relationship is both unavoidable and elusive, a constant and a variable. Despite many attempts to keep them apart in order to justify a pure legal theory or the self-reliance of law and legal constructs²², ethics and morality have always been key to the development of constitutional law and law in general, to shaping state and government, especially the role and scope of the judiciary²³, and to ensuring the constant evolution of law, playing a role in the legal experience in a myriad of ways. Furthermore, we should acknowledge that, like science and technology, the law ultimately belongs to ethics, not only because it inevitably embodies a certain ethics or ethical perspective, but also because it should always be governed by ethics (despite the distance between the formal and substantive theories of the rule of law).

Given their strict correlation and interdependence, not at odds with their respective specificities, asking ourselves which comes first, ethics or the law, appears to be a ‘chicken and egg’ dilemma.

Now AI is reshuffling the cards and, as part of the ‘cut and paste’ technique it uses to reshape the ontology of things²⁴, is trying to separate ethics from law for two different, but concurrent, reasons.

The first is to push through the idea that, as far as AI is concerned, ethics does not need the law because it is self-standing and perfectly equipped to deal with AI issues, no matter how big (i.e. the human/machine divide, ontology of the humane, obligation- v. reward-based behavioral schemes for machines and humans²⁵, etc.). This idea resolves the ‘chicken and egg’ dilemma, affirming the (ethical) supremacy of ethics over law and its superior virtue in coding AI (automatically transformed into a virtuous virtuality). The corollary is that the law is both unnecessary and unfit for the job.

The second reason is to delay AI law and any attempt to code AI

²¹ Floridi, 2023; Benanti and Maffettone, 2024.

²² Kelsen, 1967.

²³ See Dworkin, 1977. On substantial deference see US SC 605 US _ (2025) landmark decision in the *Seven County Infrastructure Coalition v. Eagle County*, Colorado case.

²⁴ See footnote no. 15.

²⁵ With reference to the reward dogma/hypothesis, see Benanti, 2025.

through law. This idea resolves the ‘chicken and egg’ dilemma in a different way: AI ethics takes precedence and AI law must wait until the ethical side has been clearly outlined. A sort of democratic and political AI pause²⁶.

Unsurprisingly, this approach to the ethics/law relationship gives rise to interesting and unethical phenomena of ethics washing and ethics shopping²⁷. It is also worth mentioning here that parallel arguments are also being developed within and through the law for the same purpose, that of keeping AI safe from law (which may sound unethical in its turn). One of these arguments exploits the *free speech clause* to qualify AI business, development and functioning, together with AI systems and applications, as formed of free speech that is *by law free from law*²⁸.

To move towards humane AI we need, instead, to accept the ‘chicken and egg’ dilemma, strengthening the alliance between ethics and the law, always keeping AI in the loop: there are no AI exemptions to ethics and to the laws on the books²⁹ and, at the same time, no AI law exemptions to ethics and to the laws on the books³⁰. Democracy and the rule of law are not external restrictions on AI, but its backbone. The same applies, or should do, to the human and the humane.

This alliance is also key, because the law itself is a generative deposit of (AI) ethics: firstly, due to its set of constitutional principles (such as the recognition of human(e) dignity as a cornerstone), bills of rights and democratic and judicial institutions, processes and mechanisms; secondly, due to the fact that AI law progressively recognizes and embodies AI ethics³¹, contributing considerably to its development, making it legally significant and justiciable, and expanding its potential far beyond the AI value chain to the entire human/humane reality.

In this respect, what we are witnessing today is the emergence of a multi-level and far-reaching web in which AI ethics and law, both hard and soft law, are starting to exchange information, learn from one another and mutually support one another³². Here again, thanks to its resilient and adaptive infrastructures, the law plays a key role, opening

²⁶ Similarly to the AI law enforcement pause clause removed by the text of the 2025 *One Big Beautiful Bill Act*.

²⁷ Wagner, 2018, pp. 84-89.

²⁸ Zuboff, 2019, pp. 112-123.

²⁹ The expression was formulated by L. Khan, FTC Chief on 4 October 2023 at the Creative Economy and Generative AI conference.

³⁰ See EU AI Act, cit., preamble (7), (8), (27), (28), (48) and more.

³¹ EU AI Act, cit., preamble (116), (117), (165) and (174), as well as arts. 50, 53, 55.

³² EU AI Act, cit., preamble (7), (25), (27), art. 95; COE’s Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of

up a multi-dimensional space/time for shaping humane AI, together with its ethics and law (a new Humane AI-driven Nomos of the Earth³³).

3. *Keep AI in the Rights Loop: Humane AI needs us to take the rights-driven approach seriously*³⁴

A third key point is the following: Humane AI needs to take rights and the ‘rights-driven approach’ seriously, which means accepting that the rights-driven approach is no cure-all, and that we need to keep AI in the rights loop and value chain, not the other way round.

As far as AI is concerned there is no cure-all. AI is no cure-all³⁵, nor is the rights-driven approach to AI law. In fact, the rights-driven approach may be used as a fig-leaf for ‘rights washing’ practices, as a Trojan horse for conveying market-driven regulatory approaches, or as a ‘politically correct’ quick answer to the many AI-driven concerns about human rights protection, democracy and the rule of law, which leaves the big questions open and unanswered. In all of these cases, the law does not serve the cause of Humane AI.

Let us dig a little deeper.

AI law still is at the starting blocks, both domestically and internationally, but the rights-driven approach is one of its cornerstones³⁶. In AI law, ‘rights-driven’ also means democracy- and rule-of-law-driven, democracy and the rule of law being the generative contexts for substantially protecting individual freedoms and rights. It also means human-centric and common-good-oriented, as humans, both as

Law, Vilnius, 5 November 2024, available at <https://rm.coe.int/1680afae3c>; see also The Paris Charter on Artificial Intelligence in the Public Interest, Paris, 11 February 2025.

³³ Schmitt, 2006.

³⁴ The expression refers to Dworkin’s book *Taking Rights Seriously*.

³⁵ For AI failures see <https://www.aisnakeoil.com> by Narayanan and Kapoor.

³⁶ EU AI Act, cit., preamble (6) “it is vital for AI and its regulatory framework to be developed in accordance with Union values as enshrined in Article 2 of the Treaty on European Union (TEU), the fundamental rights and freedoms enshrined in the Treaties and, pursuant to Article 6 TEU, the Charter. As a prerequisite, AI should be a human-centric technology. It should serve as a tool for people, with the ultimate aim of increasing human well-being”; (7) “In order to ensure a consistent and high level of protection of public interests as regards health, safety and fundamental rights, common rules for high-risk AI systems should be established”. See also the Preamble to COE’s Framework Convention on AI and HRS: “Convinced of the need to establish, as a matter of priority, a globally applicable legal framework, setting out common general principles and rules governing the activities within the lifecycle of artificial intelligence systems that effectively preserves shared values and harnesses the benefits of artificial intelligence for the promotion of these values in a manner conducive to responsible innovation; recognising the need to promote digital literacy, knowledge about, and trust in the design, development, use and decommissioning of artificial intelligence systems”.

individuals and community members, along with their well-being (and humanity), are essential prerequisites to a democracy and the rule of law.

This broad, holistic view mirrors the growing awareness of the major impacts AI can have on humans individually, on humans as members of a society, on the future of humanity or, to put it differently, on present and future generations. It follows that the ‘rights-driven’ approach is inevitably much more than ‘rights-driven’, as it connects AI to the foundations of human existence, civil coexistence and law itself.

Since so much is at stake, and much is still completely unexplored or even unimaginable³⁷, taking the ‘rights-driven’ approach to AI law – or, better, to AI in law – seriously is far from an easy task. Were this not enough, the humane AI perspective adds another layer of complexity in that it implies we resolve the human/humane divide, and the many anthropological and philosophical issues that come with it.

In light of the above, ‘rights-driven’ AI law contains some hidden pitfalls. The first pitfall lies in the fact that it qualifies human dignity, human agency (the human in the loop), individual freedoms and rights, democracy and the rule of law as external boundaries, or limitations, on AI development and use³⁸. Similarly, it also makes frequent generic external references to existing human rights law and sector-specific legislation, according to the principle that there are no AI/AI law exemptions to the laws on the books and to the level of rights protection they currently ensure³⁹.

This perspective is at the same time misleading and packed with risky legal implications.

Misleading, because it implies that legal freedoms and rights, together with human dignity, human agency, democracy and the rule of law can be taken for granted to form a kind of security perimeter surrounding AI and AI law. The point is that in the AI-driven new ontology, nothing is left outside or untouched, so that external references to legal constructs and sources may turn out to have a

³⁷ Even if much still is uncertain and unpredictable, the EU AI Act adheres to a traditional risk-based approach and provides a list of prohibited AI practices (art. 5), because presenting an unacceptable level of risk: i.e. which deploy subliminal techniques beyond a person’s consciousness or purposefully manipulative or deceptive techniques, exploit any of the vulnerabilities of a natural person or a specific group of persons, or for detrimental social scoring purposes. Based on art. 7, the Commission, with the support of the European Artificial Intelligence Board (art. 65), may review AI prohibited practices and the list of High-Risk AI systems.

³⁸ EU AI Act, preamble (6) and (7).

³⁹ EU AI Act, preamble (9).

different or no substance, and be incapable of framing and securing AI and AI law.

Misleading also because legal freedoms and rights, democracy and the principle of the rule of law are consequently reduced to interests to be protected along with others (such as environmental protection⁴⁰), instead of working as the ontological legal backbones of AI and AI law.

Finally, this perspective is misleading because it follows a sort of reverse order. Instead of keeping AI in the freedoms and rights loop and value chain, it keeps freedoms and rights in the AI loop and value chain. Accordingly, AI is qualified as a product, humans as users/consumers and AI law as a sector-specific field of law subject to a risk-based assessment, irrespective of the major uncertainty surrounding AI⁴¹.

Taking this all together, what does the ‘rights-driven’ approach boil down to? To nothing other than a traditional rights-compliant market-driven approach. Unsurprisingly, AI law brings very little innovation to the legal/regulatory field⁴² and, moreover, seems to be kept immune from the most interesting regulatory strategies triggered by sustainable development, climate justice and transformative change⁴³. The same applies to AI case law⁴⁴.

The second pitfall of the ‘rights-driven’ approach can be described as the ‘human in the loop’ paradox. Even if the development of AI systems has so far been carried out with severe violations to basic human

⁴⁰ EU AI Act, preamble (7), (8) and art. 1. This interest-based approach may also live to divide the ontological unity of the in-dividual (the etymology of individual being ‘not divisible’), and lead to a lower level of legal protection.

⁴¹ See footnote 36.

⁴² Provisions such as the ban on certain AI practices, the classification of high/limited/minimal-risk AI systems, the provision of experimental normative space/times (sandboxes), the focus on the AI value chain, obligations requiring the establishment of AI system management and data governance, the principle of human supervision and the use of soft law regulatory tools, are common to other market-sector specific pieces of legislation.

⁴³ I.e. the shift from a rights-driven approach to a duties-driven approach, the recognition of a present-future space/time for the protection of future generations and the fair sharing of AI negative/positive impacts, and the recognition of the primacy of ecology and ecosystems protection. Only a courageous interpretation of established regulations, constitutional ones in the first place, may overcome such failure.

⁴⁴ In the *TikTok v. Garland Inc.* decision, US SC, 604 U. S. (2025), even if the Court recognizes that TikTok is “a technology with transformative capabilities”, it applies free speech in quite a traditional way, without innovating its case law, according to what Justice Frankfurter advised in a 1944 case: in applying established legal rules to the “totally new problems”, we should take care not to “embarrass the future”. The same cautious approach has been taken by a Court of California, siding for Anthropic in an author’s rights case.

rights and freedoms, from labour rights to health- and environment-related rights⁴⁵, AI development has proven itself to be capable of bypassing human rights obligations by making increasing recourse to AI agents (including human-shaped robots or humanoids). Moreover, AI development (human-based computation) is transforming humans, including AI users, into tools for AI constant training and value creation⁴⁶, leading to a sort of endless technological embodiment which goes either way: humans to machines or machines to humans. The same applies to democratic institutions and processes⁴⁷. This is part of the AI-driven new ontology and changes are fast and extensive, becoming even faster and more extensive as digital transformation progresses, reshaping individuals, society, markets and democracy from the inside⁴⁸.

So, if the loop is becoming bigger and bigger, where should the human be placed within it?

This question calls into play the humane.

Leaving aside issues concerning the legal personality of robots and their protection in law⁴⁹, we can clearly understand that the ‘rights-driven’ approach implies taking the ‘human in the loop’ paradigm seriously, not just as regards the roles humans should play in the life cycle of artificial intelligence systems⁵⁰, but also, by extension, the conditions and limits for grounding human-machine collaboration. Human rights and constitutional guarantees should be seen as capabilities, enabling humans to determine their own identity and their humane way of being, including within this increasingly important

⁴⁵ See COE’s Framework Convention, cit., art. 4 on Protection of human rights, art. 7 on Human dignity and individual autonomy (“Each Party shall adopt or maintain measures to respect human dignity and individual autonomy in relation to activities within the lifecycle of artificial intelligence systems”; see also EU AI Act, art. 14 on Human oversight in each high-risk AI system loop).

⁴⁶ Floridi, 2023, pp. 30 and 59.

⁴⁷ On democracy and AI disintermediation, see Floridi, 2023, pp. 33.

⁴⁸ See Boschetti, 2022b.

⁴⁹ That of electronic persons is a concept explored by the European Parliament in October 2020 in respect to the ethical and legal aspects of Artificial Intelligence software systems (“AI”). See Resolution 2020/2012(INL) on a Framework of Ethical Aspects of Artificial Intelligence, Robotics and related Technologies (the “AI Ethical Aspects Resolution”), Resolution 2020/2014(INL) on a Civil Liability Regime for Artificial Intelligence (the “Civil Liability Resolution”), and Resolution 2020/2015(INI) on Intellectual Property Rights for the Development of Artificial Intelligence Technologies (the “IPR for AI Resolution”). The number of robots that resemble and act like humans is likely to reach nearly 1 billion by 2050, with China currently leading in development.

⁵⁰ Properly defining the loop is important because HITL requires it, and for some loops a HITL may be counterproductive and cause more errors, whereas for others it may be essential. See EU AI Act, cit., art. 14 on Human oversight. In the literature, see Tschider, 2024.

relational dimension⁵¹. In the light of the above, the rights-driven approach essentially means always keeping AI in ‘the rights loop’, which is, or should be, the one defined by the humane, not by AI systems.

4. Keep Matter in the AI/Law Loop: Humane AI needs us to take care of the material side of the on-life

A fourth key point is the following: Humane AI needs us to take care of the material side of the on-life, meaning of our body, ecosystems and physical space/time because matter matters in both AI and the law. This material dimension still defines the human, the humane and the law. It should also define humane AI.

‘Keep matter in the loop’ is therefore a key imperative for humane AI and law.

Let us start by stating that the digital revolution is also a spatial revolution. It is radically transforming the spaces we live in and the way we experience them. It is changing everything that exists in space/time – eco- and urban systems, individuals and society, the economy, democracy and law itself – to the extent that the concepts of space and time have become unclear and uncertain. It goes without saying that this spatial revolution is an existential revolution, of constitutional relevance. Space, digitally redefined, is redefining the ontology of the world and of law.

Digital space, the metaverse, looks like a new promised land, with no boundaries, infinite possibilities and potential, an experimental (*virtual*) space. A perfect place to host the duo of humanity and the earth (the many digital twins), to look at ourselves as if in a mirror, transformed into images, idols (from the ancient Greek *eidolon*⁵²), to find refuge from the finiteness of the world, from the scarcity of resources, from the annoying materiality of needs, from the obstacles posed by spatial and temporal lags. A place at our fingertips, digital in fact, which appears, with all it offers, to be in our hands and under our control, changeable on demand. A new world in which we can expand (from the Sanskrit root of *space*, *spa*-). It is where the on-life, life extended to the digital space, takes place, in a process of ever closer integration, even of desired unity.

The law takes us into the on-life: it gives us a new identity (a digital identity), new forms of citizenship (digital citizenship), of equality (the

⁵¹ Alpa, 2021.

⁵² Harari, 2017.

digital republic), of democracy (the e-democracy) and of government (government as a platform).

The truth is that due to its mix of infinity-abstractness-virtuality, the digital space deceives us.

It gives us an illusion of freedom, an apparent freedom since, no matter what we do, we live mostly in the digital space, as a means of production, as (consumed) consumers, as followers or as users of customizable services (which is precisely personalization designed for consumers); an illusion of emancipation, from needs, from the other and others, from what is different, from the scarcity of resources, from the limits that only become tangible in the physical space (what about the logistics of the virtual economy or energy/water consumption connected to AI?); an illusion of knowledge and innovation, which makes us satiated and fulfilled as if, by entering the digital space, with its pocket oracles (the AI chatbots), we had knowledge and creativity at our disposal; an illusion of the future, if it is true that the generativity of AI, with its probabilistic inference⁵³, keeps us tied to the past and to what can be extracted from it (in ancient Greek the root *spân*, from which the word *space*, means *to extract*).

In this attractive game of illusions, we abstract and grow apart from ourselves and the world, from politics and law, which also means from human agency and humane intelligence. In fact, it is also thanks to matter – meaning our body, ecosystems and space-time – with its at times annoying concreteness and finiteness – that we are urged to put ourselves one before the other, before nature, being the other and nature mirrors of the humane⁵⁴. And again, it is thanks to the concreteness and finiteness of matter that we owe things such as freedoms and rights, equality, responsibility, respect, care and justice, in that they imply relationality⁵⁵, trust and a forward-looking vision that serves to design the present/future of our societies and of humanity⁵⁶.

These illusions corrupt the political space, leading to abstract and alienating non-generative politics (*sterile* means precisely hard, steady and stuck in the present); they also corrupt legal space, resulting in

⁵³ Ananthaswamy, 2024.

⁵⁴ The reference is to Levinas and his idea of ‘the other’ which ‘orders and ordains’: Levinas, 1969; in climate justice case law see Supreme Court of Colombia, decision April 5th 2018, which interestingly affirmed that we are ‘the other’ and that we can use our own rights to protect it. ‘The other’ includes the ecosystem and the unborn, as well as future generations.

⁵⁵ Which is also the ontological basis of interests, from the Latin *inter-esse*, to be between, and also to be concerned.

⁵⁶ Ferraris, 2025; Mittica, 2022.

non-generative law, which has the earth as its mother⁵⁷, incapable of any order (from lat. *or-do*, which means ‘way to move forwards’). They corrupt physical space/time: we consume our cities (the city becomes a product to sell⁵⁸), walk through them (the city becomes a technological, linear infrastructure⁵⁹), Instagram them (the city becomes an idol, an image); and we extract from ecosystems anything the unsustainable (apparent) lightness of the digital may need (raw materials, energy, water, beauty, data and so on).

In the end, the *mundus* becomes *immundus*, the human inhumane, the virtual unvirtuous.

For Humane AI and law, we therefore must keep matter in the loop and take the unsustainable (apparent) lightness of AI seriously. It is not only a matter of pairing the digital transition to the ecological one (if ever possible)⁶⁰, but of preserving the generative (physical) ground on which human agency and humane intelligence depend⁶¹. The human and the humane, as much as politics and law, must keep themselves grounded, or they are at risk of losing their identity. We need to reaffirm *habeas corpus* and *habeas mundum mundum* for the digital era.

5. *Keep the Loop Transformative: Humane AI needs us to stay transformative and generative because we are humane by design*

A fifth key point is the following: Humane AI needs us – individually and collectively – to stay transformative, meaning generative, capable of freely designing our future (which is the essence of liberty⁶²) and of making it possible (which is the first form of justice⁶³). This is the only way to keep agency on the human side and develop the humane

⁵⁷ Schmitt, 2006.

⁵⁸ Cugurullo, 2021; Granata, 2021.

⁵⁹ Emblematic is the new smart city named ‘The Line’, a linear city of 170 km in Saudi Arabia.

⁶⁰ For data concerning data center power consumption see <https://www.iea.org/news/ai-is-set-to-drive-surging-electricity-demand-from-data-centres-while-offering-the-potential-to-transform-how-the-energy-sector-works>; on data centre water consumption see <https://www.eesi.org/articles/view/data-centers-and-water-consumption>. In EU AI Act, cit., preamble (48), “The fundamental right to a high level of environmental protection enshrined in the Charter and implemented in Union policies should also be considered when assessing the severity of the harm that an AI system can cause, including in relation to the health and safety of persons”. In the literature, Floridi, 2020; Boschetti, 2022c, pp. 1153 ff.; Boschetti, 2022a, pp. 821 ff.

⁶¹ Bombelli, 2018, pp. 53-90.

⁶² Zabala, 2020.

⁶³ Boschetti, 2023a; Boschetti, 2023b, pp. 208 ff.

intelligence we need to deal with AI, its design and un-design⁶⁴. Without human agency and humane intelligence there can be no humane AI and law.

‘Keep the loop transformative’ therefore is a key imperative for humane AI and law.

A transformative approach to change is defining the age of transitions (ecological⁶⁵ and digital⁶⁶). This approach to change(s) focuses on the urgency of radical transformation and on ‘the how to’ design and address it⁶⁷. It implies acknowledgement of both the *necessity* and *capacity* of designing and managing transformative change(s), through accelerations and decelerations, to make them real and just, and just also because they are real⁶⁸.

This transformative approach is one of the cornerstones of the UN Agenda for sustainable development: “We are determined to take the bold and transformative steps which are urgently needed to shift the world onto a sustainable and resilient path”⁶⁹.

The EU has embraced the same transformative approach. In the European Green Deal (hereinafter, EGD), the urgency (of transformative change) is recognized as a unique opportunity to transform the paradigm *by acting*, which means by addressing structural changes “*in a fair, inclusive and sustainable way*”⁷⁰. The Green Deal Industrial Plan

⁶⁴ On the importance of design and un-design for governing the digital transition, Floridi and Cabitza, 2021.

⁶⁵ Cf. EC, *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - The European Green Deal*, Brussels, COM (2019) 640 final. In literature, Chiti, 2022, pp. 19 ff.; See Bevilacqua and Chiti, 2024.

⁶⁶ EC, *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee Of The Regions*, 2030 Digital Compass: the European way for the Digital Decade, available at <https://eufordigital.eu/wp-content/uploads/2021/03/2030-Digital-Compass-the-European-way-for-the-Digital-Decade.pdf>. See Boschetti, 2022b. European Commission: Joint Research Centre, Muench - Stoermer - Jensen - Asikainen - Salvi - Scapolo, 2022, <https://data.europa.eu/doi/10.2760/977331>; in the philosophical debate, see Floridi, 2014. See also Durante, 2019, p. 37.

⁶⁷ See Langa, 2006; von Bogdandy, 2017.

⁶⁸ On the impact credibility has on states’ development, Acemoglu and Robinson, 2013.

⁶⁹ UN, *Transforming Our World: the 2030 Agenda for Sustainable Development*, <https://sdgs.un.org/2030agenda>. On the effectiveness of governance by objectives: Kaine and Biermann, 2017.

⁷⁰ In December 2019 the European Council endorsed the objective of a climate-neutral EU by 2050: “We need to ensure that the transition towards a climate-neutral EU industry takes place in a fair, inclusive and sustainable way. Turning challenges into opportunities will be our biggest task. Our discussions today prove that we are determined to preserve the competitiveness of our industry while achieving the climate goals set by the European Green Deal”; see also the EU Digital Compass, <https://eit.europa.eu/news-events/news/>

for the net-zero age and, more recently, the Competitiveness Compass for the EU⁷¹, take the same transformative approach.

As far as AI and the digital transition are concerned, there are some issues we must take into consideration to keep ourselves transformative, generative and capable of designing and un-designing the future and progressing towards it.

First, AI may put us on a performative, rather than transformative path, in the illusion that performative is a substitute for transformative. Besides, the more we fall into its performative trap, the less transformative we are, as we are kept in no apparent need of being transformative, because we are trapped in a vortex of sterile, constant change, without the need to develop a capacity of being transformative, because of AI-driven performativity.

In this respect, the great challenge is to bend AI and AI performativity to the motives and objectives of transformative change(s), not the other way round. And yet, if correctly placed, computational power can significantly add to our capacity to make transitions possible, real and just. The task requires strong human agency and humane intelligence in the ethics/politics/law loop, starting with establishing a divide between being transformative and generative, rather than *simply* performative⁷².

Second, AI may jeopardize human agency and humane intelligence, impacting on free will and our capacity to design and un-design what we are and our future, both individually and collectively⁷³. It is true that, as humans, we still have a lot of ‘agency’, but we now have to deal with artificial agents with decision-making capabilities, and growing decision-making autonomy about the if, what and how to perform certain tasks, artificial decision-making included. This does not automatically turn artificial agents into intelligent agents, but the more digitalization progresses and AI agents enter our on-life, including our political and democratic on-life, the more decision-making tasks are delegated to, or executed by, artificial agents. Even if formally and/or legally placed under ultimate human control/responsibility⁷⁴, AI deference, together

cit-digital-presents-new-report-future-education-digital-skills; Uk <https://www.gov.uk/government/publications/inclusive-and-sustainable-economies-leaving-no-one-behind/inclusive-and-sustainable-economies-leaving-no-one-behind-executive-summary>.

⁷¹ January 31st 2025, available at https://commission.europa.eu/document/download/10017eb1-4722-4333-add2-e0ed18105a34_en.

⁷² On the divide between doing and knowing (we do much more than we know, we do even without knowing, simply repeating actions or delegating tasks to technology), Ferraris, 2025, p. 226.

⁷³ Faggin, 2024.

⁷⁴ See proposal no. 3.

with AI's intrinsic capabilities to influence behaviours and minds by exploiting the infosphere and offering smart and gentle services ('what can I do for you?', 'how can I help you?', 'any questions?', and so on)⁷⁵, inevitably flattens and normalizes the way we think, learn, decide, create and design⁷⁶.

The trend is inevitable, unless we are well-prepared and fully equipped to collaborate with AI agents and exploit AI potential (they are information technologies) for the purposes of humane intelligence. In addition, we should consider that even if, individually and collectively, we generate the fuel of the digital economy (data), we are taken outside, or at the margins, of the AI value creation chain, leading to a serious redistribution gap which may increase inequalities, vulnerabilities and decrease individual and societal well-being⁷⁷.

In this respect we must acknowledge that free will, together with knowledge, are not the way to avoid mistakes (in the majority of cases, scientific progress comes about through mistakes and intuition⁷⁸), nor is it necessary to do anything, to carry out tasks or to be performative (we mostly act without knowledge, and are sometimes not informed). They are the essence of reason and of our capability to give reason(s). In short, we need free will and knowledge to be transformative. This is what separates the human/humane from the artificial in the midst of the current *industrial revolution of mind and knowledge*.

Here AI opens up the challenge of a *humane habeas mentem*, a sort of grounding condition for free speech itself⁷⁹ and any bill of rights or revised constitutional framework we may think necessary to empower the free transformative will of individuals and societies in the on-life and the infosphere. However evocative, the concept of a *humane habeas mentem* needs clarification, because it appears to imply the recognition of a basic general freedom to decide freely to bridge the longstanding gap between free will and liberty (a *habeas mentem right* to complement the *habeas corpus*

⁷⁵ See EU AI Act, art. 5, on prohibited AI practices. The study of psychological factors involved in decisions of individuals and institutions and how they deviate from the traditional economic theory is at the basis of behavioral economics studies. It is also key for new regulatory techniques used by public institutions (i.e. nudging).

⁷⁶ On the capture of the noosphere due to digital technologies, Magatti, 2023.

⁷⁷ On the idea of *webfare*, see Ferarris, 2024. The US global companies exemption from the Minimum Global Tax risks to compromise a lengthy political process and to create additional opportunities for base erosion and profit shifting.

⁷⁸ The how AI works may innovate the principle of falsification which is key to the philosophy of science.

⁷⁹ Free speech and expression of mind is at the core of the new US AI Action Plan: White House, *Winning the Race: America's AI Action Plan*, July 2025, pp. 4 ff. available at <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>.

right)⁸⁰, or the existence of a general freedom to freedom(s)⁸¹, normatively independent of each individual freedom and their sum. Indeed, until now and also in AI law, “the freedom to be free” has been protected mainly by existing or new freedoms and rights (and related duties and obligations)⁸² or, more interestingly, by decision-making processes which expand the space/time in order to decide freely and in an informed way. The same goes for the democratic on-life and political will in the democratic on-life, which protected in their integrity, independence and effectiveness, including in respect to the separation of powers, thanks to democratic institutions, decision-making processes and the rule of law⁸³.

Will this be enough?

In any case, will politics and law rise to the transformative challenge?

The answer to the first question is negative. The humane and humane AI cannot be resolved while remaining within the rights and freedoms normative loop. We need to enable the flourishing of a humane Renaissance for the digital era, as a real counterpower and counterintelligence to the artificial ones (and the even vaster economic powers that own them), a Renaissance focused on reinventing humane thinking, knowledge, creativity and human/AI collaboration. We need humane literacy for AI and humans⁸⁴.

The answer to the second question may, at least in part, be positive. The transformative approach to change is triggering new resilient law for transitions⁸⁵, in which the alliance between the political and the

⁸⁰ On the gap between free will and liberty, Mori, 2001. On the impossibility of recognizing *habeas mentem*, Huxley, A. (1932), *Brave New World*, Vintage Publishing. The title is derived from: “O wonder! How many goodly creatures are there here! How beauteous mankind is! O brave new world, That has such people in ‘t” (W. Shakespeare, *The Tempest*, Act V, Scene I, ll. 203-206).

⁸¹ See the Preamble to the US Constitution. The US Supreme Court has rejected the idea that this general/generic liberty could be independent from the single freedoms protected in the corpus of the Constitution (see *Jackobson v. Massachusetts*, 197 U.S. 11 (1905)). The digital transformation reopens the question.

⁸² Such as the right to be informed, to correctly form one own’s consensus, to privacy and to have control over personal data under the EU GDPR, the right to distinguish artificial and human-generated contents, the right to access social media and the web, the right to decide that my intellectual and cognitive activities are not exploited without my free consensus and with the freedom to opt-out from certain functions without being excluded from social or knowledge platforms, and the right to profit from the value generated by individual data (for which privacy regulation is currently unfit). On this last issue, see Ferraris, 2024.

⁸³ COE’s art. 5 on Integrity of democratic processes and respect for the rule of law.

⁸⁴ See EU AI Act, art. 4 on AI literacy obligations placed on providers and deployers of AI systems.

⁸⁵ Boschetti, 2023a; Boschetti, 2023b, pp. 208. On polycentric governance and the European Green Deal, Bevilacqua, 2024; van Zeven, 2020, pp. 300.

normative becomes key, as well as the dialogue between law and whatever is external but related to it, from ethics to science. The law is positioning itself as a nodal epistemic medium, not only because it offers a shared normative language (the power of language rediscovered thanks to AI?), but also because it is capable of shaping the space/time for transformative and generative decision-making, in which to design and un-design the future, including the future of AI and AI Law. To this aim, interesting future-design polycentric bottom-up decision-making processes have already been experimented and new tools for bottom-up and design-based public policies put in place⁸⁶. All this can help enable a humane Renaissance for the digital era.

While this may give us hope, the current arms'-race situation in AI dominance (and not only AI) may compromise all attempts to be transformative, generative and humane, both in AI and in law⁸⁷.

References

- ACEMOGLU D., ROBINSON J.A. (2013). *Why Nations Fail: The Origins of Power, Prosperity and Poverty*, Profile Books.
- ALPA G. (2021). *Il diritto di essere se stessi*, Milano, La Nave di Teseo, 2021.
- ANANTHASWAMY A. (2024). *Why Machines Learn. An Elegant Math Behind Modern AI*, New York, Dutton.
- BENANTI P. (2025). *Se l'intelligenza artificiale perfeziona gli strumenti per il suo miglioramento*, in "Il Sole24Ore", 9 July.
- BENANTI P., MAFFETTONE S. (2024). *Noi e la macchina. Un'etica per l'intelligenza artificiale*, Roma, Luiss University Press.
- BEVILACQUA D. (2024). *Il Green New Deal*, Giuffrè.
- BEVILACQUA D., CHITI E. (2024). *Green Deal. Come costruire una nuova Europa*, il Mulino.
- BOMBELLI G. (2018). *Emotion and Rationality in Aristotle's Model: From Anthropology to Politics*, in L. HUPPES-CLUYSENAER, N. COELHO (eds.), *Aristotle on Emotions in Law and Politics*, Cham, Springer International Publishing, pp. 53-90.
- BOSCHETTI B. (2023a). *Come costruire un (nuovo) diritto a prova di resilienza*, in "Il Sole24Ore", 31 January.
- BOSCHETTI B. (2023b). *Diritto e resilienza. Coordinate per un diritto resiliente capace di transizioni*, in "Quaderni cost.", 208 ff.

⁸⁶ On polycentric governance and the European Green Deal, Bevilacqua, 2024; van Zeben, 2020, pp. 300.

⁸⁷ See White House, cit.

- BOSCHETTI B. (2022a). *Eco-design giuridico (trasformativo) per la “net zero-age” e la sua economia*, in “Istituzioni del Federalismo”, 4, pp. 821 ff.
- BOSCHETTI B. (2022b). *La transizione della pubblica amministrazione al modello government as a platform*, in A. LALLI (eds.), *L'amministrazione pubblica nell'era digitale*, Giappichelli, Turin, 2022, pp. 1 ff.
- BOSCHETTI B. (2022c). *Oltre l'art. 9 della Costituzione: un diritto (resiliente) per la transizione (ecologica)*, in “DPCE online”, 2, pp. 1153 ff.
- BROWN I., MARSDEN C.T. (2013). *Regulating Code. Good Governance and Better Regulation in the Information Age*, Cambridge, MIT Press.
- CACCIARI M. (2020). *Il lavoro dello spirito. Saggio su Max Weber*, Milano, Adelphi.
- CAMPATI A. (2024). *Illiberal Trends. Democracies under Pressure*, Milan, Educatt ed.
- CHITI E. (2022). *Managing the ecological transition of the EU: the European Green Deal as a regulatory process*, in “Common Market Law Review”, pp. 19 ff.
- COUPLAND R.M. (2003). *The Humanity of Humans: Philosophy, Science, Health, or Rights?* in “Health and Human Rights”, 7(1), pp. 159-166. <https://doi.org/10.2307/4065421>.
- CUGURULLO F. (2021). *Frankenstein Urbanism: Eco, Smart and Autonomous Cities, Artificial Intelligence and the End of the City*, London and New York, NY, Routledge.
- DICASTERO PER LA DOTTRINA DELLA FEDE E DICASTERO PER LA CULTURA E L'EDUCAZIONE, *Antiqua et Nova, Note on the Relationship Between Artificial Intelligence and Human Intelligence*, available at <https://press.vatican.va/content/salastampa/it/bollettino/pubblico/2025/01/28/0083/01166.html#ing>.
- DURANTE M. (2019). *Potere computazionale. L'impatto delle ICT su diritto, società e sapere*, Sesto San Giovanni, Meltemi, 2019, p. 37.
- DWORKIN R. (1977). *Taking Rights Seriously*, Cambridge Mass, Harvard University Press.
- FAGGIN F. (2024). *Irreducible - Consciousness, Life, Computers, and Human Nature*, New Alresford, John Hunt Publishing.
- FERARRIS M. (2024). *Documanità*, Bari, Laterza.
- FERRARIS M. (2025). *La pelle. Che cosa significa pensare nell'epoca dell'intelligenza artificiale*, Bologna, il Mulino.
- FLORIDI L. (2014). *The Fourth Revolution. How the Infosphere is Reshaping Human Reality*, Oxford University Press.
- FLORIDI L. (2014). *The Onlife Manifesto. Being Human in a Hyperconnected Era*, Heidelberg New York Dordrecht London, Springer Int. Pub.
- FLORIDI L. (2020). *Il verde e il blu. Idee ingenue per migliorare la politica*, Milan, Raffaello Cortina editore.
- FLORIDI L. (2023). *The Ethics of Artificial Intelligence. Principles, Challenges and Opportunities*, Oxford, Oxford University Press.

- FLORIDI L. (2017). *Digital's Cleaving Power and Its Consequences*, in "Philosophy & Technology", 30, pp. 123-129. <https://doi.org/10.1007/s13347-017-0259-1>.
- FLORIDI L., CABITZA F. (2021). *Intelligenza artificiale*, Milan, Bompiani.
- GIACCARDI C., MAGATTI M. (2025), *Macchine celibi*, Bologna, il Mulino.
- GRANATA E. (2021). *Placemakers: gli inventori dei luoghi che abiteremo*, Milan, Einaudi.
- HALMAI G. (2024). *From Liberal Democracy to Illiberal Populist Autocracy: Possible Reasons for Hungary's Autocratization*, in "Hague Journal on the Rule of Law", 16, pp. 439-463. <https://doi.org/10.1007/s40803-024-00231-6>
- HARARI Y.N. (2017). *Homo Deus. A Brief History of Tomorrow*, New York City, Vintage Publishing.
- KAINE N., F. BIERMANN (2017) (eds.). *Governing Through Goals: Sustainable Development Goals as Governance Innovation*, Cambridge, Massachusetts, The MIT Press.
- KELSEN H. (English Translation 1967). *Pure Theory of Law*, University of California Press.
- LANGA P. (2006). *Transformative Constitutionalism*, in "Stellenbosch Law Review", 353.
- LEVINAS E. (1969). *Totality and infinity* (A. Lingis, Trans.), Pittsburgh, Duquesne University Press.
- MAFFETTONE S. (2023). *Il nostro tempo col pensiero*, Milan-Udine, Mimesis Edizioni.
- MAGATTI M. (2023), *Libertà nella supersocietà*, in "Pandora Rivista", available at <https://www.pandorarivista.it/pandora-piu/liberta-nella-supersocieta/>.
- MITTICA M.P. (2022). *Il pensiero che sente. Pratiche di Law and Humanities*, Turin, Giappichelli.
- MORI M. (2001). *Libertà, necessità, determinismo*, Bologna, il Mulino.
- MUENCH S., STOERMER E., JENSEN K., ASIKAINEN T., SALVI M., SCAPOLO F. (2022). *Towards a Green and Digital Future*, Publications Office of the European Union, Luxembourg, <https://data.europa.eu/doi/10.2760/977331>
- NARAYANAN A., KAPOOR S. (2024). *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*, Princeton, Princeton University Press.
- POLLICINO O. (2023a). *Potere digitale*, in *Enciclopedia del Diritto. I Tematici*, V, Milano, Giuffrè.
- POLLICINO O. (2023b). *Di cosa parliamo quando parliamo di costituzionalismo digitale*, in "Quaderni Costituzionali", 3.
- RATTI G.B. (2017). *Norms, normative systems, and legal validity*, in "Jurisprudence", 8(1), pp. 166-176. <https://doi.org/10.1080/20403313.2017.1296071>
- SARTEA C. (2024). *Ecotecnologia. Sfide etico-giuridiche della civiltà tecnologica*, Turin, Giappichelli.
- SCHMITT C. (English version 2006 by G.L. Ulmen). *The Nomos of the Earth in the International Law of the Jus Publicum Europaeum*, Telos Press Publishing.

TSCHIDER C.A. (2024). *Humans Outside the Loop*, in “Yale Journal of Law & Technology”, 26, 324.

UN, *Transforming Our World: the 2030 Agenda for Sustainable Development*, <https://sdgs.un.org/2030agenda>

VAN ZEBEN J. (2020). *The European Green Deal: The Future of a Polycentric Europe?*, in “European Law Journal”, pp. 300.

VON BOGDANDY A. (ed.) (2017). *Transformative Constitutionalism in Latin America*, Oxford University Press.

WAGNER B. (2018). *Ethics as an Escape from Regulation. From “Ethics Washing” to “Ethics Shopping”?*, in E. BAYAMLIOĞLU, I. BARALIUC, L. JANSSENS, *Being Profiled: Cogitas Ergo Sum*, Amsterdam, Amsterdam University Press, pp. 84-89.

WHITE HOUSE (2025). *Winning the Race: America’s AI Action Plan*, July, pp. 4 ff. Available at <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>

WOLIN S.S. (2010). *Democracy Incorporated: Managed Democracy and the Specter of Inverted Totalitarianism*, Paperback ed., Princeton University Press.

ZABALA S. (2020). *Being at Large: Freedom in the Age of Alternative Facts*, McGill-Queen’s University Press.

ZUBOFF S. (2019). *The Age of Surveillance Capitalism*, London, Profile Books Ltd, pp. 112-123.

2. Zero Sales Resistance: The Dark Side of Big Data and Artificial Intelligence*

Giuseppe Riva, Brenda K. Wiederhold, Sauro Succi

ABSTRACT

Big data (BD) is the hue and cry of modern science and society. The impact of such data-deluge is huge and far-reaching for both science and society. Moreover, given the effort required for collecting and analyzing this data, Artificial Intelligence (AI) has replaced the human mind in accomplishing the enormous task of deriving insight out of the information. In this paper, we analyze the role of BD and AI in steering the world population towards the state of *Zero Sales Resistance* (ZSR): the inability to exert critical judgement over the most seductive aspects of the aforementioned data deluge. Moreover, we discuss the alarming consequences of presenting the merging of BD and AI as a universal panacea even if, to date, they have proven far more efficient for predicting human decisions and behaviors (predictive analytics) than for solving the most critical problems in science and society. Why? Our answer is simple. The causal structures associated with such challenges command a detailed understanding of the underlying mechanisms (causal explanation), typically acting nonlinearly and on a broad range of scales in space and time. In contrast, personality and behavior can be predicted with no need of a microscopic theory and understanding of the brain-mind system (empirical prediction). This is a direct consequence of the fact that our mind, at least for the intuitive level, uses the same prediction techniques applied by AI (bayesian predictions based on our past experience). However, prediction is not explanation, and without joining them will be impossible to achieve a major advance in our understanding of complex systems.

* This chapter was originally published as Riva G., Wiederhold B.K., Succi S. (2022). *Zero Sales Resistance: The Dark Side of Big Data and Artificial Intelligence*, in "Cyberpsychology, Behavior, and Social Networking", 25(3), pp. 169-173. doi:10.1089/cyber.2022.0035. © Mary Ann Liebert, Inc.

*The real objection is that if man chooses
to treat himself as raw material,
raw material he will be.*
(C.S. Lewis, *The Abolition of Man*)

1. *Introduction*

Big data (BD) is the hue and cry of modern science and society. We are accumulating data at an unprecedented pace, currently rated at some 2.5 Exabyte/day (2.5 quintillion bytes, corresponding to approximately hundred Dante's *Divina Commedia pro-capite* per day), for a total of about 30 Zettabytes (30 thousand quintillion bytes, 30 followed by 21 zeros).

The impact of such data-deluge is huge and far-reaching for both science and society. Data accumulation is commonly presented under the 'more is better' umbrella, but the situation is subtler than that. Given the effort required for collecting and analyzing these data, Artificial Intelligence (AI) has replaced the human mind in accomplishing the enormous task of deriving insight out of the information.

In this paper, we discuss the alarming psycho-social consequences of presenting the merging of BD and AI as a universal panacea (Anderson, 2008; Domingos, 2015) in the face of the fact that, to date, they have proven far more efficient for predicting human decisions and behaviors (predictive analytics) (Mahmoodi, Leckelt, van Zalk, Geukes, Back, 2017) than for solving the most critical problems in science and society (Coveney, Dougherty, Highfield, 2016; Succi, Coveney, 2019).

Specifically, we discuss the role of BD and AI in steering the world population towards the state of *Zero Sales Resistance* (ZSR) (Lewis, 1944; Succi, Coveney, 2019): the inability to exert critical judgement over the most seductive aspects of the aforementioned data deluge. It is argued that the ZSR goal is reached by means of a cleverly engineered combination of three main levers: a) the technological abolition of spacetime; b) the lack of careful reasoning which results from overwhelming the human brain with excessive and faster information than it can elaborate using heuristics only; and c) the tailoring of persuasive appeals to the psychological profiles of large groups of people. It is further noted that such a strategy maps surprisingly well onto the cognitive structures of our brain (Matz, Kosinski, Nave, Stillwell, 2017; Pennycook, Rand, 2021).

2. *The role of Big Data and Artificial Intelligence in the digital economy*

The marriage between BD and AI and the digital economy in its various forms (finance, business, marketing) is one of the most visible and powerful hallmarks of modern society (André et al., 2018), with digital business taking society by storm and dwarfing the real economy. The point is that digital economy thrives on the exchange of information and information, being much lighter than material objects, travels much faster, virtually at the speed of light.

The consequences are deep-reaching: first, space and time are erased from the picture, because, in a fully e-connected world, any individual on Earth can be reached in virtually no time (light makes nearly seven roundtrips around the globe in a single second).

This alone provides a powerful step towards the goal of zeroing the sales resistance on a planetary scale. Examples abound: in finance, an algorithm-driven keystroke in Wall Street can evaporate lifetime savings in the blink of an eye. Not something that could and would have happened just twenty years ago. Apart from the privileged high-tech financial elite, the majority of humans are not given the slightest chance to respond. Speed erases spacetime and since humans act in spacetime, it becomes much easier to make them think as one, the golden avenue towards the goal of ZSR.

Yet, the abolition of spacetime alone is not enough to make seven billion brains think as one: an extreme form of ‘psychological condensation’ (group thinking) is required.

Here, one can draw an intriguing analogy with superconducting materials, which are traversed by an electric current even in the (nominal) absence of applied voltage because the charge carriers (electrons) develop long-range correlations which prevent them from colliding against obstacles, hence attaining zero resistance. Interestingly, this occurs through a collective mechanism, known as Bose-Einstein condensation, in which a pair of electrons behave ‘like one’ because they fall in the same quantum ground state, the one in which the system attains its minimal energy.

At this stage, another universal trait of human nature is cleverly leveraged: a well-known collective effect called ‘herd behavior’ (Kameda, Hastie, 2015). In herding, like electrons in superconductivity, individuals decide to follow others and to imitate group behaviors rather than deciding independently on the basis of their own personal beliefs and values (Baddeley, 2010). As explained by Baddeley (2010) herd behavior reflects an interaction of cognitive and emotional factors. On one side, once our brains are overwhelmed by too much information, humans are influenced by visceral factors, cognitive

constraints, environmental cues and framing effects, and tend to act by imitation. On the other side, as we will expand on later, herding can be explained using Bayesian hypotheses: individuals define their behaviors systematically revising their probabilistic judgements using information about others' actions, and this generates herding and 'information cascades'.

Finally, using the many digital footprints we produce everyday – from Likes in social media (Kosinski, Stillwell, Graepel, 2013), to behavioral data obtained from smartphones (Stachl et al., 2020) – AI can develop persuasive appeals tailored to the psychological profiles of different users that are very effective in influencing their behavior (Matz et al., 2017). This new approach goes much further than more conventional strategies that act to influence societal behavior (Cosić, Srbljinović, Popović, Wiederhold, Wiederhold, 2012).

In a classical paper Matz and colleagues (2017) demonstrated, in a sample of over 3.5 million individuals, that AI-driven, psychologically-tailored advertising, matching the content of the communication to individuals' psychological characteristics, significantly altered the behavior of the sample as measured by clicks and purchases. Moreover, as demonstrated by Facebook in a "massive-scale contagion experiment" it is possible to "affect real-world emotions and behaviors, in ways that bypassed user awareness" (Kramer, Guillory, Hancock, 2014). Additionally, it may be interesting in future experiments to study real-time fMRI brain images as these decisions are made.

The circle is now closed: overwhelm the human brain with excessive and faster information, surgically engineer a marketing torpedo (via predictive analytics), hand it over to your most seductive global success role model ('influencer') and the seven billion humans collapsed into a single brain, will follow on a blind date. As clarified by Matz and colleagues (2017) these examples: "illustrate clearly how psychological mass persuasion could be abused to manipulate people to behave in ways that are neither in their best interest nor in the best interest of society" (p. 12717).

This situation has been recently summarized by the term 'Surveillance Capitalism' (Zuboff, 2015), an economic system centered around the commodification of personal data with the core purpose of profit-making through the modification of behavior. In an interview, a scientist at a Silicon Valley company explained it clearly: "The goal of everything we do is to change people's actual behavior at scale. When people use our app, we can capture their behaviors, identify good and bad behaviors, and develop ways to reward the good and punish the bad. We can test how actionable our cues are for them and how

profitable for us” (Zuboff, 2016). Psycho-condensation and zero sales resistance accomplished.

Chickenfeed, an old story after all, now powered however on a global scale by the best kids and algorithms in town. But why do such algorithms work (much) better for persuasive communication and psycho-tech applications, like Sentiment Analysis, than for the most compelling scientific problems, for instance turbulence, quantum gravity, the environment or personalized medicine, to name but a few grand challenges of our modern times? A possible explanation lies in the difference between *causal explanation* and *empirical prediction* and in their link with the functioning of our brain.

3. *Prediction and explanation in our brain*

In many scientific fields, statistical modeling is using causal explanation and prediction for generating and testing theories (Shiffrin, 2016; Shmueli, 2010). But what is their role in the process? If explanatory modeling is the use of statistical models for testing causal explanations, predictive modeling is the use of statistical models for the purpose of predicting new or future observations (Shmueli, 2010).

As noted by Breiman (2001), the differences between them have generated two different cultures in statistical modeling. Most Big Data and AI researchers follow the ‘algorithmic modeling culture’, in which the goal is to find an algorithm that results in the same outputs as an unknown and possibly unknowable process given the same inputs (prediction). However, we cannot convert statistical knowledge (predictions) into causal knowledge or correlation into causation. As explained by Pearl (2001): “The building blocks of our scientific and everyday knowledge are elementary facts such as ‘mud does not cause rain’ and ‘symptoms do not cause disease’ and those facts, strangely enough, cannot be expressed in the vocabulary of probability calculus” (p. 19). Therefore, theory building and testing also require a ‘data modeling culture’, in which the primary goal is to estimate the true parameters of a particular data-generating process (explanation).

This epistemological distinction between prediction and explanation reflects the difference between the two modes of cognitive functioning operating in our brain. As explained by Daniel Kahneman (2002) our brain is characterized by two different modes of thought: System 1 (*Intuition*) and System 2 (*Reasoning*). The two systems have specific pros and cons. If System 1 is superior in directing everyday behavior through a stream of predictions, System 2 is superior in its ability to comprehend

cause-and-effect relations, to delay gratification, and to plan for possible worlds (Norris, Epstein, 2011).

In fact, the reasoning mode – that is deliberately controlled, slow and effortful – generates judgements (explanations) that are always explicit and intentional. On the other side, the intuitive mode – that is automatic, fast, and difficult to control or modify – produces impressions (predictions) related to the characteristics of objects of perception and thought. To achieve its predictions our brain uses precise mathematical formulas in which the probability of an event is calculated in advance based on previous experiences (Bayesian probability): probability expresses a degree of belief in the occurrence of a given event (Raftery et al., 1992). Being based on what happened previously, it is not an absolute probability, but it is always conditional on previous knowledge and its value is progressively updated based on new events to minimize the possibility of error.

This functioning of our mind suggests that since our behaviors result from predictions based on our past experiences, it is relatively easy for an algorithm that knows our previous experiences to predict our behavior with no need for detailed knowledge of the way our brain functions. Yet, as we have seen, our daily practices in a media environment generate an ocean of digital traces that can be used both for predicting many types of behavioral and personality indices (Matz et al., 2017; Pennycook, Rand, 2021) and for modifying users' behaviors.

Moreover, it also suggests a clear relationship between prediction and explanation (Douglas, 2009). Successful predictions allow some measure of control over natural processes. Explanations, instead, provide the ground for generating new predictions whenever the existing ones prove inadequate.

One of the biggest challenges for AI predictions is the process of extrapolation. Extrapolation is an estimation of a value based on extending a known sequence of values or facts inferring something that is not explicitly stated from existing information. As underlined by Succi and Coveney (2019), at the core of the process of extrapolation is the Law of Large Numbers: with enough samples, the sample mean is equal to the population mean, hence the samples obey Gaussian statistics. However, for complex systems extrapolation is a fairly hazardous procedure: in the presence of correlations, a regular occurrence in complex systems, Gaussian statistics no longer holds (Succi, Coveney, 2019). In this case, AI may reproduce very well a given set of data, and still fail grossly as soon as new data are added to the dataset.

Unfortunately, as opposed to the human brain, AI prediction is not yet accompanied by a reasoning mode capable of correcting and

expanding existing forecasts by planning for possible worlds that are not based only on our previous experiences. The result is that unprecedented events are treated as impossible; a major mistake whose potentially dire consequences have been described more fully in Taleb's books *Fooled by Randomness* (2001) and *The Black Swan* (2008).

Even AlphaFold2, Google DeepMind's AI system for predicting the 3D structure of proteins, which is considered the biggest scientific achievement in AI, has continued problems with complicated protein targets that have multiple functional domains or highly dynamic structures (Eisenstein, 2021). As clearly explained by Eisenstein (2021): "Progress in deep learning – and a growing community of AlphaFold2 users – could bring some of these challenges to heel, but a comprehensive understanding of protein biology will require a much broader computational and experimental toolbox" (p. 707).

4. *Conclusions*

Summarizing, we have pointed out why, to date, BD and AI have served the cause of digital business and virtual economy much better than they did the scientific one. Algorithms which maximize business by implementing the ZSR strategy do not (necessarily) help us solve the grand challenges facing humankind, such as curing diseases, global environmental issues or understanding of the brain.

The latter sit on higher rungs of the complexity ladder (Succi, 2022), for causal structures associated with such challenges command a detailed understanding of the underlying mechanisms, typically acting nonlinearly and on a broad range of scales in space and time (Succi, 2018). Therefore, the power of BD and AI should be used to complement and enhance theory, modelling and simulation, not to dismiss them. Only through a creative merging of explanation and prediction, intuition and reasoning, will it be possible to achieve a major advance in our understanding of complex systems.

In contrast, personality and behavior can be predicted with no need of a microscopic theory and understanding of the brain-mind system. We maintain that this is a direct consequence of the fact that our mind, at least for the intuitive level, uses the same Bayesian prediction techniques applied by AI which can be captured without a detailed knowledge of the underlying mechanism of brain functioning. The latter, on the other hand, are needed to understand, hence cure, neurological pathologies, such as Alzheimer's syndrome or depression. Steering the brain towards ZSR appears much easier than mending it and the consequences of the technological impact on the development

of human personality and psychological identity are just beginning to be fully appreciated. This is no footnote in the grand picture of assessing the impact of Big Data on future society, as it can lead to unprecedented and irreversible changes in the psychological development of future generations. Do we have the means to protect ourselves from ZSR? Arguably yes. The good news is that two simple approaches may work well.

On the one hand, we have to use mindful attention and reasoning for transcending the here and now of our automatic predictions and unintentional biases (McVay, Kane, Kwapil, 2009) by asking ourselves: What am I doing? Why did I behave like this? Am I sure that what I did was right? On the other hand, we have to select and cultivate in our lives behaviors and experiences associated with optimal experience (Flow), a state of deep engagement, concentration, and enjoyment (Delle Fave, Massimini, Bassi, 2011) in which intuition and reasoning meld together with the action itself. As Mark Twain reminds us, “Whenever you find yourself on the side of the majority it is time to pause and reflect”.

Acknowledgments

GR and SS wish to thank the organizers of the meeting “La metamorfosi della democrazia” (Camaldoli, October 2021), which prompted many of the considerations collected in this work. SS wishes to acknowledge financial support from the European Research Council under the Horizon 2020 Programme Grant Agreement n. 739964 (“COPMAT”). He also wishes to acknowledge many fruitful exchanges with P.V. Coveney.

References

- ANDERSON C. (2008). *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, in “Wired”, 16(7). <https://www.wired.com/2008/06/pb-theory/>
- ANDRÉ Q., CARMON Z., WERTENBROCH K., CRUM A., FRANK D., GOLDSTEIN W., . . . YANG H. (2018). *Consumer Choice and Autonomy in the Age of Artificial Intelligence and Big Data*, in “Customer Needs and Solutions”, 5(1), pp. 28-37. doi:10.1007/s40547-017-0085-8
- BADDELEY M. (2010). *Herding, social influence and economic decision-making: socio-psychological and neuroscientific analyses*, in “Philosophical transactions of the Royal Society of London. Series B, Biological sciences”, 365(1538), pp. 281-290. doi:10.1098/rstb.2009.0169
- BREIMAN L. (2001). *Statistical modeling: The two cultures (with comments and a rejoinder by the author)*, in “Statistical Science”, 16(3), pp. 199-231.
- COSIĆ K., SRBLJINOVIĆ A., POPOVIĆ S., WIEDERHOLD B.K., WIEDERHOLD M.D. (2012).

Emotionally based strategic communications and societal stress-related disorders, in "Cyberpsychology, Behavior, and Social Networking", 15(11), pp. 597-603. doi:10.1089/cyber.2012.0410

COVENEY P.V., DOUGHERTY E.R., HIGHFIELD R.R. (2016). *Big data need big theory too*, in "Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences", 374(2080). pp. 20160153.

DELLE FAVE A., MASSIMINI F., BASSI M. (2011). *Psychological Selection and Optimal Experience Across Cultures: Social Empowerment through Personal Growth*, New York, Springer.

DOMINGOS P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*, New York, Basic Books.

DOUGLAS H.E. (2009). *Reintroducing prediction to explanation*, in "Philosophy of Science", 76(4), pp. 444-463.

EISENSTEIN M. (2021). *Artificial intelligence powers protein-folding predictions*, in "Nature", 599(7886), pp. 706-708.

KAHNEMAN D. (2002). *Maps of Bounded Rationality: A Perspective on Intuitive Judgment and Choice*, in T. FRÅNGSMYR (ed.), *The Nobel Prizes 2002*, Stockholm, Nobel Foundation, pp. 449-489.

KAMEDA T., HASTIE R. (2015). *Herd Behavior*, in "Emerging Trends in the Social and Behavioral Sciences", pp. 1-14. doi:https://doi.org/10.1002/9781118900772.etrds0157

KOSINSKI M., STILLWELL D., GRAEPEL T. (2013). *Private traits and attributes are predictable from digital records of human behavior*, in "Proceedings of the National Academy of Sciences", 110(15). pp. 5802. doi:10.1073/pnas.1218772110

KRAMER A.D., GUILLORY J.E., HANCOCK J.T. (2014). *Experimental evidence of massive-scale emotional contagion through social networks*, in "Proceedings of the National Academy of Sciences of the United States of America", 111(24), pp. 8788-8790. doi:10.1073/pnas.1320040111

LEWIS C.S. (1944). *The abolition of man*, London, Oxford University Press.

MAHMOODI J., LECKELT M., VAN ZALK M.W.H., GEUKES K., BACK M.D. (2017). *Big Data approaches in social and behavioral science: four key trade-offs and a call for integration*, in "Current Opinion in Behavioral Sciences", 18, pp. 57-62. doi:https://doi.org/10.1016/j.cobeha.2017.07.001

MATZ S.C., KOSINSKI M., NAVE G., STILLWELL D.J. (2017). *Psychological targeting as an effective approach to digital mass persuasion*, in "Proceedings of the National Academy of Sciences", 114(48), pp. 12714-12719. doi:10.1073/pnas.1710966114

MCVAY J.C., KANE M.J., KWAPIL T.R. (2009). *Tracking the train of thought from the laboratory into everyday life: an experience-sampling study of mind wandering across controlled and ecological contexts*, in "Psychonomic Bulletin & Review", 16(5), pp. 857-863. doi:10.3758/pbr.16.5.857

NORRIS P., EPSTEIN S. (2011). *An experiential thinking style: its facets and relations with*

objective and subjective criterion measures, in “Journal of personality”, 79(5), pp. 1043-1079. doi:10.1111/j.1467-6494.2011.00718.x

PEARL J. (2001). *Bayesianism and causality, or, why I am only a half-Bayesian*, in D. CORFIEL, J. WILLIAMSON (eds.), *Foundations of Bayesianism*, Dordrecht, Kluwer, pp. 19-36.

PENNYCOOK G., RAND D.G. (2021). *The Psychology of Fake News*, in “Trends in Cognitive Sciences”, 25(5), pp. 388-402. doi:10.1016/j.tics.2021.02.007

RAFTERY A.E., LEWIS S., BERNARDO J., BERGER J., DAWID A., SMITH A. (1992). *Bayesian statistics*, in “Oxford Sci. Publ”, pp. 323-349.

SHIFFRIN R.M. (2016). *Drawing causal inference from Big Data*, in “Proceedings of the National Academy of Sciences”, 113(27), pp. 7308. doi:10.1073/pnas.1608845113

SHMUELI G. (2010). *To explain or to predict?*, in “Statistical Science”, 25(3), pp. 289-310.

STACHL C., AU Q., SCHOEDEL R., GOSLING S.D., HARARI G. M., BUSCHEK D., . . . BÜHNER M. (2020). *Predicting personality from patterns of behavior collected with smartphones*, in “Proceedings of the National Academy of Sciences”, 117(30), pp. 17680. doi:10.1073/pnas.1920484117

SUCCI S. (2018). *The lattice Boltzmann equation for complex states of flowing matter*, Oxford, Oxford University Press.

SUCCI S. (2022). *Sailing the ocean of complexity: lessons from the physics-biology interface*, Oxford, Oxford University Press.

SUCCI S., COVENEY P.V. (2019). *Big data: the end of the scientific method?*, in “Philosophical Transactions of the Royal Society A”, 377(2142), pp. 20180145.

TALEB N.N. (2001). *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*, London, Penguin Books.

TALEB N.N. (2008). *The Black Swan: The Impact of the Highly Improbable* London, Penguin, 2nd edition.

ZUBOFF S. (2015). *Big other: surveillance capitalism and the prospects of an information civilization*, in “Journal of information technology”, 30(1), pp. 75-89.

ZUBOFF S. (2016). *The secrets of Surveillance Capitalism*, in “Frankfurter Allgemeine”, May 3. Retrieved from <https://www.faz.net/aktuell/feuilleton/debatten/digital-debate/shoshana-zuboff-secrets-of-surveillance-capitalism-14103616.html>

3. Combining Natural and Artificial Intelligence as a Network Governance Stakeholder

Roberto Moro-Visconti

ABSTRACT

Corporate governance is undergoing a paradigm shift with the advent of Artificial Intelligence (AI) as an active stakeholding agent rather than a mere technological tool. This chapter proposes a reconceptualization of AI as a copula node in multilayer governance networks, linking shareholders, executives, regulators, and financial institutions.

I go beyond traditional agency theory by integrating game theory, Bayesian probabilistic simulations, and multilayer network analysis. I stress the emergence of multiple Nash equilibria in hybrid Artificial-Natural Intelligence (AI-NI) governance systems, showing how framing AI-human interaction as a repeated game reveals distinct equilibria aligned with different governance regimes. This game-theoretic lens highlights that cooperative, adversarial, AI-dominant, or rotational outcomes are not accidents but predictable equilibria of strategic interaction, and that design choices can expand the basin of attraction for the cooperative hybrid equilibrium.

These equilibria represent stable patterns of cooperation, competition, and hybrid strategies between human actors and AI systems. Bayesian updating is used to model evolving trust in AI, while simulations demonstrate that AI-driven governance reduces decision latency compared to traditional models. AI must be embedded in adaptive, transparent, and ethically resilient frameworks to align efficiency gains with accountability and public trust.

1. *Introduction*

Corporate oversight has long been understood as a set of relationships among principals and agents whose incentives only imperfectly align (Shleifer and Vishny, 1997). In this classical view, information is scarce, monitoring is slow, and governance solutions – boards, audits, disclosure rules, debt covenants – work by reducing agency costs between human actors (stakeholders). Over the past decade, however, organisations have embedded data-driven systems across the control cycle: sensing anomalies, prioritising alerts, scoring risk, drafting decisions, and explaining outcomes. Once software mediates who sees

what, when, and with what confidence, it no longer functions purely as a neutral tool; it becomes a digital governance participant that interacts strategically with people, influences how attention is allocated, and shapes how authority is exercised.

This chapter starts from that practical observation and develops a tractable, policy-relevant theory of hybrid human-machine (AI-NI) governance. I analyse the organisation as a multilayer network – whose nodes are represented by shareholders, managers, employees, banks, clients, suppliers, tax authorities, etc. – connected by ‘copula nodes’ that move information (big data) and decisions (monetary transactions, etc.) across connected layers (Barabási, 2016; Bianconi, 2018).

Advanced analytics (fostered by digitalization, data crunching, AI interpretation, blockchain validation, cloud storage, etc.) typically occupy these high-betweenness positions. Their power is ambivalent: they can accelerate detection, improve coverage, and compress latency; yet, because many actors depend on the same models and vendors, they can also become single points of failure. The governance challenge, therefore, is not merely whether to ‘use AI’, but how to design the interaction between people and systems so that speed and scale do not erode accountability and trust (Burrell, 2016; Binns, 2018).

I frame this as a repeated game-theoretic interaction between an algorithmic actor, such as AI (A), and human decision-makers (H) embedded in the network. In each round, A chooses a stance on the cooperation-autonomy continuum (from supportive analytics to unilateral action), whereas H chooses an oversight stance (from strict review to rubber-stamping). Payoffs include: efficiency (timeliness, throughput), accuracy (false-positive/negative balance), accountability (traceability, contestability), and resilience (robustness to drift and shocks).

These interactions generate multiple game-theoretic Nash equilibria:

- *Human-dominant* (A supports; H reviews and often vetoes);
- *AI-dominant* (A leads; H supervises lightly);
- *Cooperative hybrid* (A triages and recommends; H sets objectives, constraints, and exceptions);
- *Adversarial* (misaligned Key Performance Indicators - KPIs and opacity produce escalations and overrides);
- *Rotational* (control alternates by context – e.g., automation after hours, human panels for hard cases).

Emphasising the game-theoretic structure of hybrid human-AI governance makes clear why multiple equilibria emerge. Each hybrid combination of AI stance and human oversight is a candidate Nash equilibrium, and the stability of each depends on how trust is updated. This perspective underlines the importance of equilibrium multiplicity: organisations are not bound to a single trajectory but must deliberately

design incentives and information flows to avoid sliding into inefficient or adversarial equilibria.

Which equilibrium an organisation settles into is not fixed in advance. It evolves through Bayesian learning, an evolving process (consistent with multilayer network dynamic framing) in which stakeholders start with an initial level of trust and then flexibly adjust it step by step as they see how the system performs.

Each new outcome, feedback, audit, or explanation shifts beliefs upward if it is consistent with competence and fairness, or downward if it reveals errors or bias. When decisions are transparent, uncertainty is acknowledged and softened, systems are stress-tested, and there are real opportunities to challenge outcomes, the evidence becomes more reliable and trust grows – making it much more likely that the cooperative hybrid equilibrium will prevail. By contrast, persistent bias, unexplained decisions, or unnoticed model drift weaken trust and push the organisation toward adversarial dynamics or even a return to human-only governance.

This hybrid, game-theoretic lens complements (rather than replaces) established perspectives. Agency theory identifies incentive frictions and the need for monitoring (Shleifer and Vishny, 1997). Network governance highlights interdependence and the propagation of shocks (Kapucu and Hu, 2020). The economics of prediction clarifies where analytics create value and where human judgment remains pivotal (Agrawal, Gans, Goldfarb, 2018). Labour-market studies remind us that automation gains are uneven and context-dependent (Frey and Osborne, 2017). This contribution integrates these strands into a single, actionable framework that is accessible to practitioners and grounded in formal reasoning, with visuals that translate theory into diagnostics.

Executives and regulators face two pressing questions: *how to gain the efficiency and reach of analytics without exposing the organisation to reputational or legal damage, and how to decide when automation should be slowed, paused, or reversed.*

In safety-critical and financial contexts, the social cost of rare but systemic failures is high; in lower-stakes settings, manual bottlenecks are costly and exclusionary. My framework shows how to tune incentives and controls – documentation, monitoring, thresholds for human review, and error costs – so that cooperation becomes each side's best response, rather than relying on exhortation or after-the-fact discipline.

The main *research gap* is the following: existing scholarship richly documents agency problems and governance mechanisms, the structure and dynamics of networks, and the promises and pitfalls of algorithmic decision-making (Shleifer and Vishny, 1997; Barabási, 2016; Bianconi, 2018; Burrell, 2016; Binns, 2018; Kapucu and Hu, 2020). Yet

three gaps remain salient for practice: (i) most work treats AI as an instrument, not as a strategic counterpart (a real stakeholder) with model-governance constraints; (ii) few models make the equilibrium multiplicity in human-machine governance explicit; and (iii) the link between belief updating (trust dynamics) and equilibrium selection is under-specified for organisational design.

Within this framework, the chapter, inspired by Moro-Visconti (2025), addresses this research question: *How can organisations design human-machine governance so that Bayesian trust dynamics steer the system toward a cooperative hybrid equilibrium – rather than AI-dominant, human-dominant, adversarial, or rotational regimes – while preserving accountability, resilience, and economic value?*

I unpack this into three sub-questions:

1. *Equilibrium characterisation*: Under what combinations of transparency, oversight intensity, and payoff alternatives (efficiency, accuracy, accountability, resilience) do human-dominant, AI-dominant, cooperative, adversarial, and rotational equilibria exist and persist?
2. *Belief dynamics and selection*: How do Bayesian updates – driven by audits, explanations, and outcomes – alter best responses over repeated rounds, and therefore *select* among equilibria?
3. *Mechanism design for practice*: Which design levers (e.g., calibrated uncertainty disclosures, contestability routes with service-level targets, asymmetric loss functions, dynamic autonomy under drift alarms) expand the attraction for the cooperative hybrid equilibrium across industries?

Contributions. This chapter proposes four advances:

1. I formalise AI as a linking copula node in a multilayer governance network and as a player in a repeated game with humans, clarifying when and why multiple Nash equilibria arise.
2. I link Bayesian belief updating to equilibrium selection, showing how transparency and monitoring change the game rather than simply observing it.
3. I translate the theory into diagnostic visuals (multi-industry heatmap of governance Key Performance Indicators) that practitioners can reuse.
4. I derive practical implications for executives and checklists for supervisors that turn the cooperative equilibrium from an aspiration into a design target.

In doing so, this chapter elevates game theory from an abstract modelling device to a practical diagnostic: it shows that the governance

question is fundamentally one of equilibrium selection, and that only by recognising the multiplicity of game-theoretic Nash equilibria can practitioners and regulators hope to steer systems toward cooperation.

This chapter concentrates on oversight mechanisms and organisational design, leaving aside broader questions such as labour market effects or competition policy. The discussion is not limited to a single industry. However, it is especially relevant in areas where mistakes are costly and public trust is critical, such as finance, healthcare, and public administration.

The rest of the chapter proceeds as follows. Section 2 surveys the relevant literature and clarifies my place in it. Section 3 describes the methodology, including the Bayesian learning loop and indicators. Section 4 presents results and visualisations. Section 5 discusses multidisciplinary implications, anticipating Section 6 (discussion). Section 7 concludes with policy guidance and a research agenda.

2. *Literature review*

Corporate governance scholarship has long been dominated by agency theory, which frames the firm as a site of principal-agent conflicts and stresses monitoring and incentives as remedies (Shleifer and Vishny, 1997). Early applications of game theory to governance (Wong, 1989; Thakor, 1991) provided useful strategic insights but relied on highly simplified settings, leaving aside the complexity of multilayered, networked, and dynamic stakeholder relationships.

Network science expanded the horizon. The discovery of scale-free properties (Barabási and Albert, 1999) and the formalisation of multilayer networks (Bianconi, 2018) revealed how interdependencies and shocks propagate across stakeholders. Corporate governance was reconceptualised as a networked system (Moro-Visconti, 2019), yet most contributions remained descriptive, showing that governance is interconnected without theorising how equilibria actually emerge or collapse. Multiple Nash equilibria, consistent with a game-theoretic approach, represent a step forward.

At the same time, sustainability and ESG research highlighted that legitimacy and long-term resilience increasingly depend on environmental and social commitments. De Souza Barbosa et al. (2023) confirm that ESG integration improves performance, but ESG has often been applied through fragmented metrics, detached from the broader UN Sustainable Development Goals (SDG). This creates a conceptual gap: ESG provides measurable levers, while SDGs articulate the

normative horizon, but the dynamic link between the two is rarely formalised.

AI now disrupts this landscape. Bibliometric analyses confirm exponential growth of AI-governance research (Samara et al., 2025). Reviews point to AI's potential to monitor ESG practices (Rane et al., 2024), mechanism-based studies highlight efficiency gains (Yu et al., 2025), behavioural analyses stress nudges and perception effects (Irfan et al., 2025), and ethical debates warn of risks in sustainable finance (Sadiq et al., 2025). Yet most contributions still treat AI as a tool, not a strategic stakeholder. Only a few emerging works (Romanova, 2025) explicitly model AI as an autonomous actor, but they rarely connect this role to ESG-SDG alignment or equilibrium theory.

This chapter builds on this literature but seeks to fill three gaps. First, it formalises AI as both an autonomous player in repeated games and a copula node in multilayer governance networks, moving beyond the 'AI as instrument' view. Second, it suggests the existence of multiple equilibria – human-dominant, AI-dominant, adversarial, rotational, and cooperative hybrid – clarifying why governance outcomes are path-dependent.

Third, it embeds Bayesian trust dynamics into the analysis, showing how audits, explanations, and observed outcomes translate into equilibrium selection. By doing so, the chapter links technical AI design choices to systemic governance legitimacy and long-term sustainability.

3. *Model*

The Hybrid Human-Machine Governance Model (HGM) addresses a pressing question in corporate and institutional oversight: *How can organisations design human-machine governance so that trust dynamics steer the system toward cooperative hybrid equilibria, rather than leaving governance to drift into AI-dominant, human-dominant, adversarial, or rotational regimes?*

This question matters because AI is no longer simply a technical tool: it acts as a participant that routes information, drafts proposals, and sometimes makes decisions directly. These actions reshape authority and accountability, and must therefore be modelled explicitly.

Because trust evolves through repeated experiences and first impressions weigh heavily, the model also connects with psychological research on anchoring, cognitive bias, and risk perception – showing that governance design must address not only technical performance but also the ways humans interpret signals of fairness, accountability, and competence. Psychological dynamics such as confirmation bias, fear of loss, or overconfidence can shape how evidence is perceived, sometimes reinforcing cooperation and at other times undermining it.

This chapter's framework does not claim to capture these complexities in full. Still, it suggests that governance mechanisms will be more effective when they are attentive to human tendencies in judgment and decision-making. In this sense, the model can be read as an invitation to further dialogue between governance design, behavioural science, and psychology.

The model I propose has two layers:

- First, an accessible account explains the intuition: why AI's network position makes it powerful, how interactions with humans generate multiple stable outcomes, and why belief dynamics are decisive in steering the game-theoretic system.
- Second, a formal specification translates this intuition into a rigorous framework based on multilayer network theory, repeated games, and Bayesian updating.

3.1. Multilayer network representation

I think of governance as a multilayer network made up of different stakeholders. These include familiar human actors – shareholders, managers, employees, clients, suppliers, tax authorities, regulators, auditors, and lenders – alongside AI as a new participant. The connections between them represent flows of governance: reporting, monitoring, and decision-making. Some links carry more weight than others, reflecting differences in influence, intensity, or trust. Blockchain validation may be useful to strengthen confidence among composite stakeholders.

Within this structure, AI occupies a special albeit still underestimated position. It often becomes the bridge across layers: translating operational big data into reports for regulators, compiling performance dashboards for managers, reducing information asymmetries with instantaneous comparisons, or preparing disclosures for auditors. Because of this bridging role, AI does more than pass information along. It reshapes how the system itself works, speeding up flows and increasing efficiency, but at the same time concentrating risk in ways that can make the whole structure more fragile.

This structure directly links to the research question. If governance outcomes can settle into multiple Nash equilibria, then the placement of AI as a central connector explains why those equilibria can diverge so dramatically. By amplifying efficiency, AI makes cooperative equilibria more attractive when trust is sustained. Yet, by concentrating control, it also magnifies systemic risk if opacity, bias, or drift undermines credibility. In other words, the multilayer network highlights the dual nature of AI: it is both an accelerator of cooperation and a potential trigger of adversarial breakdowns. Algorithmic opacity, embedded in AI

dynamics, represents a further drawback. Recognising this duality is the first step in understanding how design choices – transparency, oversight, resilience – determine which equilibrium becomes stable.

3.2. Human-AI governance as a repeated game

I frame governance as a repeated interaction between two players: the AI system and the human decision-makers. The AI can either cooperate – by being transparent, accountable, and supportive – or it can defect – by acting autonomously in ways that are opaque or harder to explain (algorithms are hardly detectable). Humans, in turn, can either review decisions carefully, exercising strict oversight, or they can defer, giving AI more freedom.

The outcomes of these interactions are evaluated along four axes that matter to organisations: efficiency (speed and throughput), accuracy (correctness of decisions), accountability (the ability to trace and justify choices), and resilience (the capacity to flexibly cope with errors, shocks, or drift).

Different industries weigh these dimensions differently – for example, finance and healthcare prioritise accuracy and accountability, while logistics and retail focus more on efficiency.

Table 1 - *AI and human choices*

<i>AI's Choice</i>	<i>Humans' Choice</i>	<i>Likely Outcome</i>	<i>Pros</i>	<i>Cons</i>
Cooperate (transparent, accountable)	Review (strict oversight)	Human- dominant equilibrium	High accountability, legitimacy, trust	Slow, less efficient, costly oversight
Cooperate (transparent, accountable)	Defer (light oversight)	Cooperative equilibrium	High efficiency, trust preserved if the context is low-risk	Risks, missed errors, and weaker accountability
Defect (autonomous, opaque)	Review (strict oversight)	Adversarial equilibrium	Humans can still block harmful actions	Costly conflicts, instability, loss of efficiency
Defect (autonomous, opaque)	Defer (light oversight)	AI-dominant equilibrium	Maximum speed and throughput	Low accountability, fragile resilience, hidden risks

Table 1 summarises the four main governance scenarios that arise from the interaction of AI behaviour and human oversight. Each scenario has its own strengths and weaknesses, ranging from highly accountable but slow (human-dominant) to fast but fragile (AI-dominant). The cooperative hybrid equilibrium – where AI cooperates and humans combine review with selective deferral – emerges as the most sustainable outcome (see Table 1 for detailed scenarios and Figure 1 for a visual summary).

Figure 1 summarises the four possible equilibria resulting from AI behaviour (cooperate or defect) and human oversight (review or defer). It shows how governance outcomes vary in efficiency, accountability, and resilience. The cooperative hybrid – where AI cooperates and humans combine review with selective deferral – provides the most stable balance.

Figure 1 - *Visual matrix of governance scenarios*

		Humans' Choice	
		Review (strict oversight)	Defer (light oversight)
AI's Choice	Cooperate (transparent)	Human-dominant High accountability Lower efficiency	Cooperative High efficiency Moderate accountability
	Defect (opaque)	Adversarial Conflict, overrides	AI-dominant Fast, risky

Yet the equilibrium an organisation ultimately reaches is not predetermined. It is shaped over time through trust dynamics, as

stakeholders revise their confidence in the system's competence and fairness based on experience. This process – captured by Bayesian (machine) learning – connects the micro-level of daily outcomes, audits, and explanations with the macro-level stability of governance regimes. The research question of this paper directly hinges on this mechanism: under what conditions does trust accumulate and sustain a cooperative equilibrium, and under what conditions does it collapse into adversarial or rollback dynamics?

Figure 2 illustrates this process. It shows how trust, starting from an initial neutral expectation, gradually increases when performance is accurate, transparent, and accountable. Each positive observation strengthens beliefs, making cooperation more attractive and resilient. Conversely, if outcomes are biased, opaque, or inconsistent, beliefs erode and confidence declines, pushing the organisation toward adversarial confrontation or a return to human-only control.

Figure 2 - *Bayesian trust update curve*

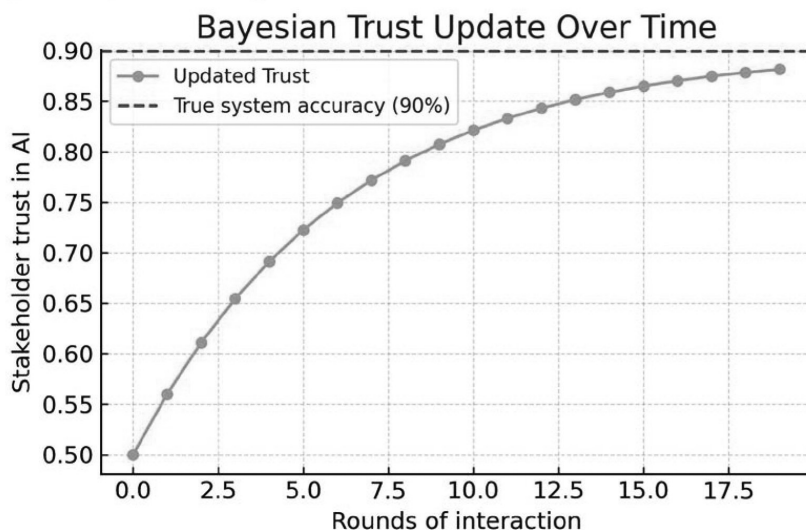


Figure 2 shows how stakeholder trust evolves over repeated interactions. Starting from an initial belief of 0.5, observed performance consistent with 90% accuracy steadily increases confidence toward the true value. This illustrates how transparency and reliable performance widen the basin of attraction for cooperative Nash equilibria, while poor signals or hidden errors would push trust downward toward adversarial or rollback dynamics.

Together, Table 1, Figure 1, and Figure 2 show that hybrid governance is best understood as a repeated game with multiple Nash equilibria. Which equilibrium prevails depends on how both AI and humans behave, and how trust is built, maintained, or eroded in the process. The central governance challenge is therefore one of equilibrium design and selection: shaping incentives, oversight, and information flows so that the cooperative hybrid equilibrium becomes the stable, long-term outcome.

3.3. Understanding equilibria in Human-AI governance

Once AI is recognised as an active participant in governance, the interaction between humans and machines does not converge to a single outcome. Instead, several stable Nash regimes are possible. In game-theoretic terms, these are Nash equilibria: strategy profiles where no player can unilaterally deviate to improve their own payoff. They are not necessarily Pareto-optimal. Many games (e.g., Prisoner's Dilemma) have inefficient Nash outcomes. Thinking in this way helps us see that governance outcomes are not accidental – they are patterns that repeat unless the rules of the game are redesigned.

From the model, equilibria can be observed in practice:

- With dynamic autonomy supported by alarms, trust fluctuates, but the system remains resilient.
- With misaligned incentives rewarding throughput above all, AI defects and humans defer, locking into risky AI dominance.

3.4. Link to the research question

The research question asked: *How can organisations design human-machine governance so that trust dynamics steer the system toward a cooperative hybrid equilibrium, rather than AI-dominant, human-dominant, adversarial, or rotational regimes?*

Multiple Nash equilibria exist, but which one prevails depends on how trust is managed and how incentives are designed. By deliberately embedding transparency, calibrated oversight, resilience safeguards, and aligned incentives, organisations make cooperation not just a moral preference but the rational, stable outcome.

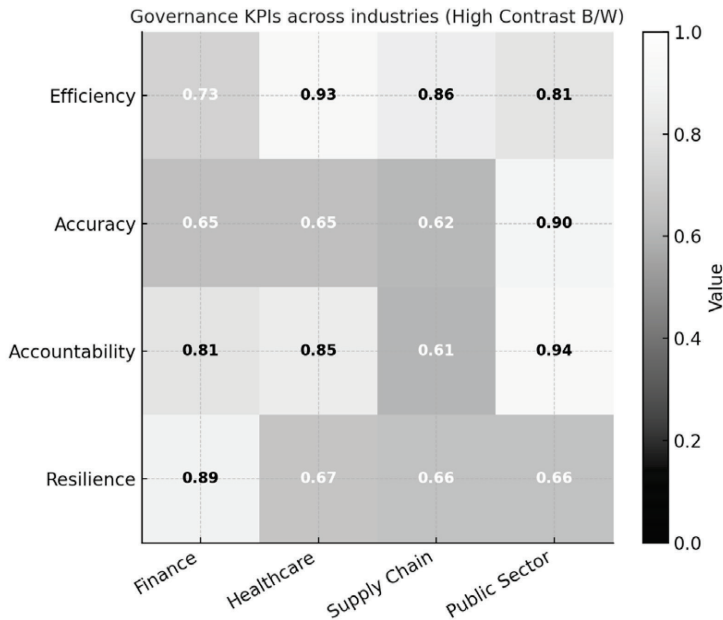
The Hybrid Human-Machine Governance Model explains why multiple governance Nash equilibria exist in AI-human systems, and how trust dynamics select among them. It provides both an intuitive narrative and a formal framework. By explicitly modelling dynamic networks, strategies, payoffs, and belief updates, the model shows how cooperation can be made stable. Importantly, it translates directly into

practice: design for transparency, calibrate oversight, embed resilience, and align incentives. In doing so, organisations can ensure that the cooperative hybrid equilibrium among naturally conflicting stakeholders is not accidental but engineered – a governance design target where efficiency, accountability, resilience, and public trust are jointly maximised.

4. Results

This section presents the results of the Hybrid Governance Model (HGM). I validate the hypotheses by combining simulations, Bayesian updates, and cross-industry analysis. The results are illustrated with reproducible tables and figures: heatmaps, 3D graphs, radar charts, and trust curves. The focus is to be both rigorous and understandable – so that executives, regulators, and practitioners can see what the numbers mean in practice.

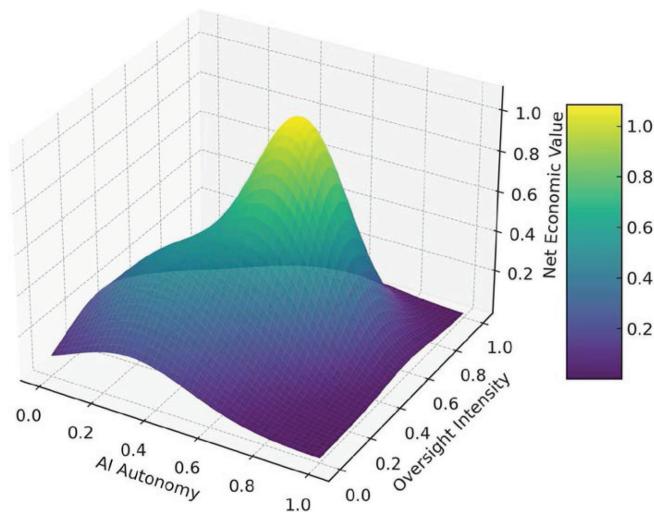
Figure 3 - *Heatmap of governance key performance indicators*



The heatmap compares governance KPIs (efficiency, accuracy, accountability, resilience) across four industries. Finance and Supply Chain show higher efficiency, while Healthcare and Public Sector emphasise resilience and accountability.

Figure 4 illustrates how the overall net economic value of governance outcomes depends jointly on AI autonomy and the intensity of human oversight. The 3D surface reveals a ridge: at very low oversight, unchecked autonomy produces fragile but initially efficient outcomes; at very high oversight, efficiency collapses under the weight of controls. The cooperative hybrid equilibrium lies precisely along the ridge, where autonomy and oversight are balanced. This confirms the model’s prediction that governance design should not aim for extremes but instead calibrate both dimensions to secure resilient, accountable, and efficient outcomes. In practice, this means combining AI autonomy with targeted oversight that activates when risks rise, producing a stable and sustainable balance.

Figure 4 - *AI governance, autonomy, and oversight*
3D Response Surface of AI Governance

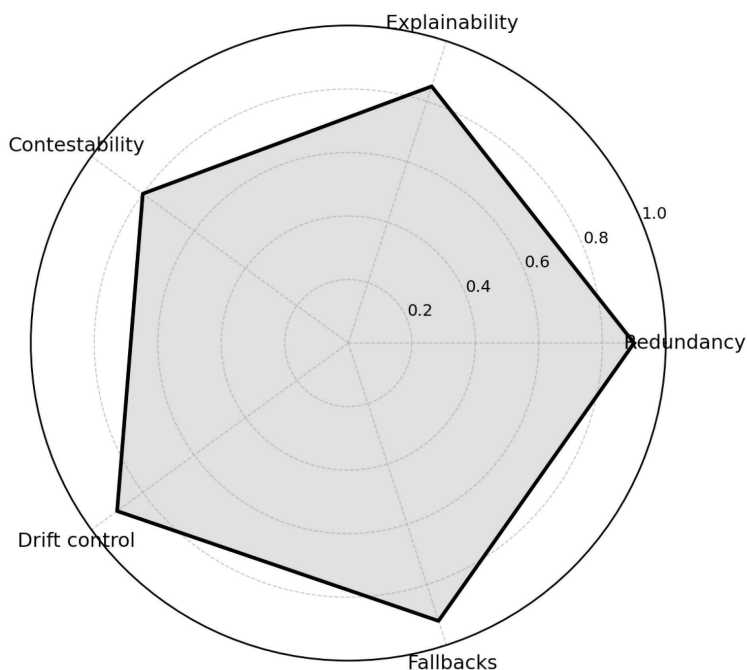


After examining cross-industry performance (Fig. 3) and the trade-off between autonomy and oversight (Fig. 4), it becomes clear that efficiency and accountability are only part of the governance story.

Equally important is resilience – the system’s Darwinian ability to withstand shocks, adapt to drift, and remain credible under stress.

Figure 5 adds this dimension. It uses a radar chart to illustrate how different attributes of resilience – redundancy, explainability, contestability, drift control, and fallback capacity – perform under a cooperative hybrid equilibrium. By visualising these attributes together, the figure shows that resilience is not a single feature but a balanced portfolio of safeguards, and that cooperative governance is strongest when all dimensions are reinforced rather than traded off.

Figure 5 - *Radar chart of resilience*



The radar chart highlights resilience dimensions under a cooperative hybrid Nash equilibrium. All dimensions – redundancy, explainability, contestability, drift control, and fallbacks – score highly. By contrast, adversarial or AI-dominant equilibria (not shown) typically underperform on contestability and fallback capacity.

5. *Multidisciplinary considerations*

A crucial extension of the Hybrid Governance Model is that cooperative equilibria cannot be secured through technical design alone. They emerge from coordinated and interdisciplinary interaction across law, psychology, management, computer science, and ethics. The research question – how organisations can design human-machine governance so that Bayesian trust dynamics steer the system toward cooperative game-theoretic equilibria – requires integration across these domains. Only through such integration can the basin of attraction for cooperation be expanded and the drift into adversarial, AI-dominant, or rollback regimes be avoided.

Legal and regulatory perspectives show that accountability cannot rest on retrospective liability alone. Hybrid human-AI governance requires prospective safeguards that operationalize accountability at the point of decision-making. Embedding contestability routes – structured mechanisms for appeal, review, and remedy – ensures that trust is enforceable rather than aspirational. Regulation must also move from static compliance to dynamic interoperability (fully consistent with a dynamic multilayer network interpretation), recognising that AI-driven decisions cross jurisdictions. In the European Union, rights to explanation and proportionality stabilise belief dynamics, while in common law systems, fiduciary duties and disclosure serve a similar role. Across contexts, contestability provides credible assurances that trust can be maintained.

Psychological and behavioural research reminds us that Bayesian updating is not purely rational. Human perceptions of algorithmic performance are shaped by anchoring, confirmation bias, and loss aversion. Early experiences are disproportionately influential, and a single salient failure can outweigh many correct outputs. This makes trust trajectories especially sensitive to initial conditions. Transparent pilots, shadow testing, and independent audits thus act as cognitive anchors, enabling stakeholders to form resilient priors. If first impressions are transparent and consistent with competence, trust grows; if opacity or errors dominate, belief spirals downward and adversarial equilibria emerge. Success, therefore, depends not only on performance but also on the psychology of trust formation (difficult to build, easy to destroy).

From a managerial perspective, equilibrium stability (difficult to build, easy to destroy) depends on incentive compatibility. If performance metrics privilege speed and throughput, AI autonomy will dominate even when accountability is eroded. If oversight is over-rewarded, efficiency collapses. The cooperative Nash equilibrium requires incentives where accuracy, fairness, and resilience are valued

alongside efficiency. This implies redesigned performance systems that align managers, developers, and overseers toward cooperation as their rational choice. Organisational culture further amplifies these dynamics: speak-up cultures that support internal contestability help preserve trust, while hierarchical and dictatorial opacity accelerates its erosion. Organisational design, therefore, shapes which signals are visible, credible, and acted upon in Bayesian updating.

Computer science contributes to the technical artefacts that make governance principles operational. Explainability, drift detection, redundancy, and fallback mechanisms function as infrastructural enablers of cooperative equilibria. Their presence makes transparency and resilience observable and credible, shifting priors in favour of trust. Explainability must be tiered: operational dashboards for managers, compliance documentation for regulators, and plain-language accounts for affected stakeholders. In this sense, technical infrastructure is not only supportive but constitutive of governance outcomes. The resilience attributes highlighted in the model – redundancy, explainability, contestability, drift control, fallback capacity – are given concrete form through model cards, decision logs, adversarial stress tests, and alarms.

Ethics provides a further layer by embedding legitimacy beyond compliance and efficiency. Cooperative equilibria are sustainable only when aligned with societal expectations of fairness, proportionality, and inclusiveness. Ethical design thus functions as a higher-order mechanism that stabilises trust across stakeholders with divergent interests. Because Bayesian trust dynamics are sensitive to perceptions of justice, systems perceived as unfair or opaque lose credibility even if technically accurate. Conversely, ethical alignment makes stakeholders more tolerant of occasional inefficiencies, since outcomes are seen as legitimate. Trust dynamics, therefore, depend simultaneously on technical performance, behavioural cues, legal safeguards, and normative alignment.

Taken together, these perspectives show that cooperative equilibria are not automatic products of technical optimisation but systemic outcomes of multidimensional reinforcement (consistent with a multi-layer network dynamic). Legal safeguards make accountability enforceable; psychology manages belief trajectories; management calibrates incentives and culture; computer science provides infrastructural transparency and resilience; ethics secures legitimacy. Each dimension reduces fragility, but only their alignment expands the cooperative basin and makes cooperation the stable outcome of the repeated game.

Multidisciplinary integration clarifies how the levers identified earlier – transparency, calibrated oversight, resilience safeguards, and incentive alignment – are embedded in practice. In law, they appear as contestability rights and adaptive regulation; in psychology, as trans-

parent pilots; in management, as balanced metrics and culture; in engineering, as explainability tools and drift alarms; in ethics, as fairness audits and inclusiveness criteria. When these levers converge, Bayesian learning steers the system toward cooperation; when they diverge, adversarial equilibria prevail.

The research question, therefore, finds a multidimensional answer: organisations can design human-machine governance so that trust dynamics sustain cooperative equilibria only if safeguards and incentives are embedded across disciplinary domains. Hybrid governance is not secured by any single intervention but by a configuration of legal, psychological, organisational, technical, and ethical mechanisms that reinforce one another. Multidisciplinary integration is thus a structural necessity, transforming the cooperative equilibrium from a fragile aspiration into a robust design target grounded simultaneously in law, behaviour, management, computation, and ethics.

6. Discussion

The results of this chapter validate the Hybrid Governance Model (HGM) and provide a coherent answer to the research question: *How can organisations design human-machine governance so that Bayesian trust dynamics steer the system toward a cooperative hybrid equilibrium – rather than AI-dominant, human-dominant, adversarial, or rotational regimes – while preserving accountability, resilience, and economic value?*

By combining network theory, game-theoretic reasoning, Bayesian simulations, and diagnostic visualisations, the evidence shows that hybrid cooperation between Artificial and Natural Intelligence is not only possible but can be systematically engineered. The discussion below interprets the findings in depth, situates them in existing scholarship, and draws out theoretical and practical implications.

6.1. Rethinking governance as equilibrium choice

The results confirm that hybrid governance is not predetermined but instead path-dependent, with multiple possible equilibria: human-dominant, AI-dominant, cooperative, adversarial, or rotational. This multiplicity represents a major departure from traditional corporate governance theory.

Classical agency theory (Shleifer and Vishny, 1997) assumes that humans are the only principals and agents, and governance structures serve mainly to reduce informational asymmetries and monitoring costs. The findings extend this logic: once AI occupies high-betweenness

‘copula’ positions in connected governance networks, it ceases to be a neutral tool and functions as a strategic counterpart. The equilibria I identify – particularly the cooperative artificial and natural intelligence hybrid – clarify how governance is best understood as a dynamic system of strategic interactions, rather than a fixed allocation of authority.

6.2. Trust dynamics as the decisive mechanism

The simulations underscore that Bayesian trust dynamics are the decisive mechanism linking micro-level events to macro-level governance outcomes. A transparent and accurate AI system gradually builds trust; opacity and error quickly destroy it. This finding makes explicit what much of the algorithmic governance literature only implied: that trust is not a static attitude but a continuously updated belief state that shapes strategy choices over time.

The trust curve results highlight strong path-dependence. Early performance matters disproportionately: if initial deployments are transparent and well-monitored, organisations enter a virtuous cycle that sustains cooperation and value co-creation. Conversely, early opacity produces a downward spiral that locks governance into adversarial or rollback equilibria. This explains why pilot phases, shadow testing, and red-teaming are critical not just as technical exercises but as trust-building rituals that determine long-run equilibrium selection.

6.3. Contributions to theory

This chapter makes four key contributions to the theoretical literature on governance and algorithmic decision-making:

1. *AI as a governance stakeholder*: By modelling AI as an actor in repeated games and as a copula node in multilayer networks, I reconceptualise AI from a tool to an autonomous stakeholder. This shift aligns with but extends perspectives in regtech and suptech that treat analytics as ‘infrastructural agents’.
2. *Equilibrium multiplicity*: The identification of equilibria moves the debate beyond ‘automation versus oversight’ binaries. It provides a typology that policymakers and executives can use to diagnose the state of their systems.
3. *Belief dynamics and equilibrium selection*: By linking Bayesian updating to equilibrium outcomes, I bridge a gap in both game theory and governance studies. Trust is shown not just as a normative value but as a mechanism of equilibrium selection.
4. *Mechanism design for practice*: Most importantly, I move from diagnosis to prescription. By showing how transparency, oversight calibration, resilience safeguards, and incentive alignment reshape

the game itself, I make cooperation a design target rather than an aspirational hope.

6.4. Practical implications

For practitioners, the results translate into a clearer governance playbook. The evidence supports three core lessons:

- *Invest early in transparency*: Because of path-dependence, the first impression created by AI deployment is decisive. Artefacts such as model cards, explainability tools, and independent audits should be mandatory from the start.
- *Balance autonomy and oversight*: Efficiency gains are real, but too much autonomy erodes accountability while too much oversight undermines efficiency. Dynamic, risk-based on oversight thresholds, organisations can occupy the cooperative ‘ridge’ identified in the 3D simulations.
- *Design resilience into the system*: Contestability, drift alarms, fallback mechanisms, and red-team drills prevent small shocks from snowballing into systemic crises. These safeguards transform adversarial potential into cooperative stability.
- *Align incentives with cooperation*: Neither AI developers nor human overseers should be rewarded for speed alone. Incentive structures must value accuracy, fairness, and accountability on equal footing with throughput.

Together, these implications transform the cooperative equilibrium into an actionable blueprint. Governance does not need to oscillate between extremes: it can be engineered to settle in a stable, trust-sustaining middle ground.

6.5. Limitations and future directions

The model is necessarily stylised. While simulations confirm equilibrium multiplicity and trust-driven selection, parameter values were chosen for illustration rather than derived from empirical datasets. Future research should therefore:

- *Empirically calibrate the model* using real-world decision logs from finance, healthcare, and public administration.
- *Test behavioural assumptions* about trust formation through experiments and surveys, especially across cultures with different attitudes toward automation.
- *Integrate regulation* more explicitly into the model, as oversight by external regulators may alter equilibrium stability in ways the current framework abstracts from.

– *Explore longitudinal dynamics*, examining how equilibria evolve under technological change, competitive pressures, and crises. These avenues would deepen the empirical grounding of HGM while preserving its analytical clarity.

Finally, the discussion circles back to the research question. The evidence suggests that organisations can indeed design governance systems where Bayesian trust dynamics naturally steer toward cooperation. But this outcome is not automatic: it requires proactive design choices that amplify transparency, calibrate oversight, embed resilience, and align incentives.

If these levers are applied, the cooperative hybrid equilibrium becomes a self-reinforcing state that combines efficiency with accountability, resilience, and legitimacy. If they are ignored, efficiency gains are short-lived, and governance drifts toward adversarial conflict or AI dominance.

Thus, the Hybrid Governance Model shows that the cooperative equilibrium is not accidental but buildable. It represents not a fragile compromise but a deliberate governance design target that aligns technological efficiency with organisational legitimacy and public trust.

7. Conclusion

This chapter develops a hybrid governance model that explains how organisations can align human judgment and AI through trust, oversight, and resilience.

Building on agency theory, network theory, and game theory, the analysis shows that governance outcomes do not converge to a single path. Instead, several stable regimes and Nash equilibria are possible: human-dominant, AI-dominant, adversarial, rotational, or a cooperative hybrid. Which equilibrium prevails depends less on raw technical performance than on how trust is built and maintained over time. Transparency, meaningful oversight, and resilience safeguards act as design levers that tilt the system toward cooperation. Simulations and visual diagnostics confirm that early, opaque failures erode trust and lock organisations into conflict or rollback, while calibrated oversight and reliable performance expand the cooperative basin.

For practitioners, the message is straightforward: governance is not about choosing between humans and machines, but about designing the conditions that allow them to succeed together. Early investments in transparency, risk-sensitive oversight, and continuous monitoring of resilience turn cooperation from a fragile aspiration into a rational and stable outcome. For regulators, the model suggests that policies should

focus less on static compliance and more on equilibrium design – shaping incentives and trust signals so that cooperative outcomes remain attractive over time. For managers and AI developers, the implications are practical: align incentives with fairness, embed contestability, and recognise that first impressions matter greatly. Ultimately, hybrid governance is not a defensive necessity but an opportunity to build systems of decision-making that are faster, fairer, and more trustworthy than either humans or AI alone.

This chapter has shown that once AI becomes embedded in organisational governance, it is no longer a passive instrument but an active participant that shapes improved outcomes. AI filters data, frames alternatives, and routes decisions, often occupying central network positions that magnify both benefits and risks. This makes governance not a binary choice between human control and automation, but a dynamic game in which multiple Nash equilibria are possible. The research question asked whether organisations can deliberately design these systems so that trust dynamics steer toward cooperation rather than drifting into dominance, conflict, or oscillation. The answer emerging from the model and simulations is clear: yes, they can – but only through deliberate and interdisciplinary design.

The results demonstrate that governance equilibria are not predetermined. They can be human-dominant, preserving accountability at the cost of speed; AI-dominant, maximizing throughput but undermining resilience; adversarial, where opacity and misaligned incentives generate conflict; or rotational, where control alternates depending on context. The cooperative hybrid equilibrium – where AI triages and recommends while humans set objectives, constraints, and exceptions – proved to be the most desirable regime. Importantly, the findings show that this state is not merely aspirational but achievable, provided transparency mechanisms, calibrated oversight, resilience safeguards, and incentive alignment are embedded from the outset.

The decisive factor linking these equilibria is trust. Trust is not a vague cultural attitude but a measurable, dynamic process. Bayesian simulations showed how stakeholders continuously update beliefs about AI's reliability in response to observed outcomes. Transparent and accurate performance steadily compounds trust, reinforcing cooperation, while opacity or bias erodes trust, pushing systems into conflict or rollback. Early experiences are consequential: initial opacity can permanently damage credibility, while early transparency creates a buffer of goodwill that sustains cooperation through later shocks. In governance, therefore, managing trust trajectories is as important as managing efficiency.

This insight has profound implications across disciplines. For legal experts and regulators, it highlights that accountability cannot be

retrofitted; contestability, explainability, and audit trails must be integral from day one. For psychologists, the results resonate with established research on anchoring and path-dependence in trust formation, pointing to the importance of studying how first encounters with AI shape long-term attitudes. For IT and cybersecurity professionals, the radar of resilience confirms that redundancy, anomaly detection, and fallback mechanisms are not technical luxuries but pillars of credibility. For AI practitioners, accuracy alone is not enough: performance metrics must also track explainability, fairness, and calibration, which directly affect the signals feeding into trust.

Cooperation is achieved by balancing autonomy and oversight, aligning incentives, and investing in transparency.

The broader implication is that hybrid AI-human governance is not a fragile compromise but a buildable equilibrium. With the right levers, efficiency and accountability can reinforce rather than undermine each other. Without them, organisations drift into adversarial or unstable regimes, where efficiency gains evaporate under the weight of mistrust. The cooperative hybrid equilibrium should therefore be treated as a governance design target – a state that can be intentionally engineered, monitored, and sustained.

Naturally, this model has limits. The simulations are stylised and depend on parameter assumptions. Future research must empirically calibrate the model with real-world decision logs, explore cultural variation in trust dynamics, and examine how regulatory frameworks alter equilibrium stability across jurisdictions. Longitudinal studies of crises – from financial shocks to cyberattacks – could illuminate how equilibria shift under stress. Behavioural experiments and field trials would help validate the Bayesian assumptions about belief updating in practice. These extensions will not only refine the model but also strengthen its value as a decision tool for organisations and regulators alike.

Despite these caveats, the message is optimistic. The cooperative equilibrium is not an abstract ideal but a realistic, achievable outcome. It is a governance system in which efficiency, accountability, resilience, and trust are not in tension but in balance. Achieving it requires collaboration across disciplines: lawyers embedding accountability into rules and rights; psychologists guiding the design of trust-building processes; IT experts ensuring resilient infrastructure; AI developers delivering transparency and fairness; and managers orchestrating incentives and oversight. If these communities act together, organisations will not oscillate between blind enthusiasm for automation and reactionary rollback. Instead, they will stabilise a cooperative mode of governance that is faster, fairer, and more resilient than either humans or machines could deliver alone.

The most profound lesson of this work, then, is simple: the future of governance does not lie in deciding whether humans or machines should rule. It lies in designing the conditions under which they can rule together.

In short, governance can be efficient, accountable, and trusted if designed deliberately. The cooperative hybrid equilibrium is not just theoretical – it is an actionable design target.

The findings confirm the hypotheses of the Hybrid Governance Model. This means that cooperation between humans and AI is not a matter of chance but can be deliberately designed. Three validated lessons follow:

1. Invest early in transparency: Transparency directly sustains accountability without eroding efficiency (H1). Early opaque failures, by contrast, were shown to collapse trust and lock systems into adversarial or rollback equilibria.
2. Balance autonomy with oversight: Results from the 3D response surface confirmed that both excessive oversight and unchecked autonomy reduce net value (H2). The cooperative hybrid equilibrium lies on the ‘ridge’ where autonomy and oversight are balanced, proving that calibrated hybrid governance is the most sustainable design.
3. Monitor resilience continuously: Bayesian trust dynamics demonstrated that resilience safeguards – contestability, fallback systems, and drift control – prevent small shocks from becoming equilibrium-shifting crises (H3). These mechanisms transform isolated incidents into recoverable events, widening the basin of attraction for cooperation.

Hybrid governance shows that the future of decision-making will not be defined by humans or machines alone, but by the capacity to weave them into systems of trust and accountability that are stronger, fairer, and more resilient than either could be on their own.

References

- AGRAWAL A., GANS J.S., GOLDFARB A. (2018). *Human judgment and AI pricing*, in *AEA Papers and Proceedings*, Nashville, 108, pp. 58-63.
- BARABÁSI A.L. (2016). *Network science*, Cambridge, Cambridge University Press.
- BARABÁSI A.L., ALBERT R. (1999). *Emergence of scaling in random networks*, in “Science”, 286(5439), pp. 509-512.
- BIANCONI G. (2018). *Multilayer networks: Structure and function*, Oxford University Press.
- BINNS R. (2018). *Fairness in machine learning: Lessons from political philosophy*, in

"Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency", ACM, pp. 149-159.

BURRELL J. (2016). *How the machine "thinks": Understanding opacity in machine learning algorithms*, in "Big Data & Society", 3(1), pp. 1-12.

DE SOUZA BARBOSA A., DA SILVA M.C.B.C., DA SILVA L.B., MORIOKA S.N., DE SOUZA V.F. (2023). *Integration of Environmental, Social, and Governance (ESG) criteria: Their impacts on corporate sustainability performance*, in "Humanities and Social Sciences Communications", 10(1), pp. 1-18.

FREY C.B., OSBORNE M.A. (2017). *The future of employment: How susceptible are jobs to computerisation?*, in "Technological forecasting and social change", 114, pp. 254-280.

IRFAN M., DIAS R.T., KHEMKA M., MONDAL P., MOHAN N. (2025). *Navigating towards sustainability: The role of corporate governance and AI-driven behavioral interventions*, in F. DAIDAI, M. BENALI, L. TAMNINE, *Behavioral Finance and AI Tools for Sustainability*, IGI Global, pp. 201-226.

KAPUCU N., HU Q. (2020). *Network governance: Concepts, theories, and applications*, London, Routledge.

MORO-VISCONTI R. (2019). *Combining network theory with corporate governance: Converging models for connected stakeholders*, in "Corporate Ownership & Control", 17(1), pp. 125-139.

MORO-VISCONTI R. (2025). *Is artificial intelligence a new stakeholding agent?*, in "Human-Intelligent Systems Integration", pp. 1-14.

RANE N., CHOUDHARY S., RANE J. (2024). *Artificial intelligence-driven approaches to strengthening Environmental, Social, and Governance (ESG) criteria in sustainable business practices: A Review*, in "Journal of Cleaner Production", 439.

ROMANOVA A. (2025). *Game theory for autonomous AI systems in corporate governance*, in https://www.researchgate.net/publication/392844324_Game_Theory_for_Autonomous_AI_Systems_in_Corporate_Governance

SADIQ R., YASSER F., NAWAZ F. (2025). *Designing ethical AI governance in sustainable finance ecosystems*, in "Advance Journal of Econometrics and Finance", 3(2), pp. 224-235.

SAMARA H.H., QUDAH H.A., MOHSIN H.J., ABUALHIJAD S., BANI HANI L.Y., AL RAHAMNEH S., ALQUDAH M.Z. (2025). *Artificial intelligence and machine learning in corporate governance: A bibliometric analysis*, in "Human Systems Management", 44(2), pp. 349-375.

SHLEIFER A., VISHNY R.W. (1997). *A survey of corporate governance*, in "The Journal of Finance", 52(2), pp. 737-783.

THAKOR A.V. (1991). *Game theory in finance*, in "Financial Management", 20(1), pp. 71-94.

WONG T.K. (1989). *An application of game theory to corporate governance*, in "Omega", 17(1), pp. 59-67.

YU X., FAN L., YU Y. (2025). *Artificial intelligence and corporate ESG performance: A mechanism analysis based on corporate efficiency and external environment*, in "Sustainability", 17(9), 3819.

OPEN QUESTIONS

Regulating the Unprecedented: Law, Technology, and the Mythologies of Artificial Intelligence

Gabriele Della Morte

1. *Is 'Artificial' Intelligence a myth?*

Generative language, oracles, hallucinations, spells... the field of artificial intelligence is rife with rhetorical figures that can be misleading. Rather than relying on technical terminology, discourse around disruptive digital innovations often draws from the language of myth, reinforcing preconceived notions that shape how we approach these technologies.

The terms commonly invoked in discussions concerning artificial intelligence suggest ancient fears of a creation that escapes the control of its creator, a longstanding narrative in human culture rooted more in tradition than in innovation. From the Golem of Jewish Kabbalistic belief – a powerful clay giant incapable of feeling emotions – to Shelley's *Frankenstein* (1818), these concerns were most recently expressed in a Pope's first participation in a G7 international summit (held in Italy in 2024). On this occasion, His Holiness Pope Francis stressed the urgency of addressing the ethical dimensions of artificial intelligence, recalling that: "No machine should ever have the power to decide to take a human life".

The 1940s and 1950s represent the golden age of this narrative tradition – an era to which we must still turn today for insights into our present. It was during this period that Isaac Asimov articulated the renowned 'Three Laws' of Robotics, which may be regarded as an early conceptual precursor to the contemporary distinction between human-in-the-loop systems – where human agents are actively involved in the decision-making processes of artificial intelligence – and human-on-the-loop systems, where human oversight is maintained but intervention occurs only in supervisory or exceptional scenarios. Asimov's laws, first introduced in his 1942 short story *Runaround* and later widely disseminated through the *I, Robot* collection, are as follows: 1) A robot may not harm a human being or, through inaction, allow a human

being to come to harm; 2) A robot must obey the orders given by human beings, except where such orders would conflict with the First Law; and 3) A robot must protect its own existence, provided that such protection does not conflict with the First or Second Law.

Some years later, in 1954, Fredric Brown published an astonishing short story (*Answer*) where, in just a few lines, the author explores the tension between human agency and the unintended consequences of technological advancement. With ironic brevity, it suggests that by creating a system of near-omnipotent intelligence, humanity may have unknowingly fulfilled its own myth – bringing into existence a new ‘deity’ it can no longer control. That could be interpreted as a perfect introduction of the ‘Human-out-of-the-loop’, an automated or autonomous system that – unlike ‘Human-in-the-loop’ or ‘Human-on-the-loop’ – can operate entirely without human intervention once it is activated. The novel imagines a near future where all technologically advanced civilisations across the universe join forces to construct the most powerful supercomputer ever created. This machine is built by interconnecting the computing systems of every inhabited planet, forming a single, unified intelligence. Once the colossal system is activated, a technician from the Earth poses the first question: “Is there a God?”. The supercomputer instantly responds: “Yes. Now there is a God”, and immediately afterwards, as the technician tries to disconnect the machine, a bolt of lightning strikes and kills him on the spot.

More significantly, during the same years, Alan Turing (1950) published *Computing Machinery and Intelligence*, a groundbreaking paper in which he explored the fundamental question: “Can machines think?”. Rather than attempting to define the term ‘thinking’, Turing proposed an operational criterion, now well known as the ‘Turing Test’. In the latter, a human engages in a text-based conversation with both a machine and another human. If the first person is unable to reliably distinguish between the two, the machine is said to exhibit ‘intelligent behaviour’. The work of Turing laid the theoretical foundation for the field of artificial intelligence and continues to shape ongoing debates about the nature of mind, intelligence, and machine cognition (also because Turing ultimately predicted that by the end of the 20th Century machines would be able to deceive humans in the imitation game with considerable success).

The production of seemingly fabricated content is due to the following underlying goal: to be convincing, to generate outputs indistinguishable from those a human might produce. A recent and striking example is the *Schwartz case*, involving a lawyer who submitted a legal brief generated by ChatGPT. While the document impeccably

followed formal quotations of previous cases – including case names and dates – the cited precedents were entirely fictitious.

Such incidents support the argument made by scholars (Finocchiario, 2024), who suggest replacing the term ‘intelligence’ – which anthropomorphises the discussion – with ‘machine learning’, an expression that more accurately reflects the data-driven nature of these systems. Others, like Bender (2021), go further, invoking the metaphor of ‘stochastic parrots’ – models capable of generating language without any true understanding of meaning ([Language Model] is “a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning”: a stochastic parrot). In the end, as Floridi (2014) reminds us (quoting Dijkstra), “The question of whether a computer can think is no more interesting than the question of whether a submarine can swim”.

2. *What is the impact of emerging technologies on law (and vice versa)?*

While terminology may vary, there is broad scholarly consensus on the transformative impact of the new technologies, yet no general agreement on the extent to which legal systems are capable of effectively governing this process. Information technologies and artificial intelligence have become strategic assets – not only for corporations seeking economic gain, but also for States competing for geopolitical influence. Their implications for national and international legal systems remain the subject of ongoing debate, particularly as the growing concentration of digital power in private hands has emerged as a significant challenge to public authorities and state sovereignty. Consequently, the need to strike an appropriate balance between openness and regulatory oversight in the digital sphere has become increasingly urgent.

Given these premises, it becomes evident that artificial intelligence systems possess exceptional potential, primarily due to their capacity to adapt to a wide array of functions and objectives. This inherent functional versatility presents a profound challenge to legal systems, which – unlike the flexible and evolving lexicon of technology – are traditionally grounded in stable and enduring elements: constitutional ‘rigidity’, normative ‘resistance’, the ‘structure’ of legal reasoning, and the ‘foundational’ principles underpinning the rule of law.

In this regard, as highlighted in a compelling interdisciplinary exchange between the philosopher Emanuele Severino and the jurist Natalino Irti (2000), technology, conceived as an “ordering principle”, tends towards omnipotence, as it preserves the ability to pursue any

conceivable objective. Law, by contrast, asserts its regulatory function by upholding the distinction between the rule and the ruled. As Irti aptly notes, “technology is not above the parties; rather, it is a party to the process and thus subject to judgment – not absolute or supreme, but relative and historical” – of the law.

At this stage a critical question arises: in what ways are the most recent innovations influencing the dynamic tension between technology and law? Addressing this question necessitates a careful examination of how artificial intelligence is reshaping certain foundational categories of legal thought, as what characterises the present moment is not only the complexity of these challenges but also the unprecedented nature, speed, and scale of technological transformation.

The way all of this will impact national and international regulatory frameworks is currently the subject of intense and ongoing debate (an issue that will be addressed further below). Nonetheless, these concerns are not entirely new. Legal scholars and practitioners have long been required to navigate the tension between two competing imperatives: the need for stability, to ensure continuity and predictability within the legal order; and the need for adaptability, to accommodate shifting social and technological realities. As Roscoe Pound already observed in 1922: “Law must be stable and yet it cannot stand still”.

But how exactly is artificial intelligence reshaping some of the foundational categories of legal discourse such as legal subjectivity, attribution, and liability? In assessing the impact of emerging technologies on the legal domain, the main legal challenges due to the transition from the analogical to the digital world concern the dimensions of space and time (Della Morte, 2018, 2022). In the traditional context (analogical), it is generally possible to measure and compare, for instance, whether something is temporally or spatially nearer or farther. By contrast, the digital realm – based on the numerical representation of reality – does not offer the same capacity for measurement or comparison.

Beginning with the spatial dimension, it is important to note that in the universe of numerical representation – namely, the domain in which artificial intelligence systems operate – phenomena are situated within an immaterial ‘space’ devoid of territorially identifiable boundaries. As a result, within this “océan de 0 et de 1” (Garapon, Lassègue, 2018), what Tobler (1970) identifies as the first law of geography – “everything is related to everything else, but near things are more related than distant things” – no longer holds true (Vegetti, 2017).

This experience of ‘deterritorialisation’, or more accurately ‘*despatialization*’ (Garapon, 2021), cannot be reduced merely to the question of state control over territorial segments of artificial intelligence systems or

the networks through which relevant data are processed. Certain technical features – ranging from decentralised architectures in the production and processing of data to the importance of the technology used to record data in addition to their storage location – challenge this territorial interpretation. By way of example, the *U.S. CLOUD Act (Clarifying Lawful Overseas Use of Data)*, enacted by the United States Congress in 2018, enables U.S. authorities to access digital data held by cloud service providers regardless of the physical location of that data (including outside the United States), if the providers fall under its jurisdiction – which, in practice, is often the case.

Clearly, the legal issues arising from the phenomenon we have identified as ‘*despatialization*’ primarily concern the complex problem of attributing responsibility for digital conduct – both at the individual and state levels (i.e., actions carried out by state organs that may entail international responsibility under public international law). Consider, for example, a cyberattack with lethal and/or destructive kinetic effects targeting individuals and/or infrastructure, initiated by artificial intelligence systems without any human intervention – such as in the case of human-out-of-the-loop scenarios previously discussed. In such instances – often anonymous and simultaneously directed, in a ubiquitous manner, against multiple targets – not only does a high risk of misattribution emerge, but the time required to ascertain responsibility may even jeopardise the application of international legal norms governing the lawful exercise of self-defence. This is why the establishment of cyber-defence perimeters to protect critical infrastructure has become a crucial issue in both domestic affairs and international relations. The matter is even more pressing given that the operation and maintenance of such systems often depend on the technical expertise, know-how, of private enterprises and non-state actors. The question of how certain essential public functions are now performed by private entities remains one of the most pressing challenges of our time. A concrete illustration of this aspect can be found in the ongoing war in Ukraine, where SpaceX’s Starlink satellite system has played a pivotal role in sustaining communications infrastructure. Following the disruption of conventional networks caused by Russian attacks, Starlink terminals provided the Ukrainian government, military, and civilian population with resilient Internet connectivity. This has proven essential especially for coordinating defence operations. At the same time, its use has raised debate about the geopolitical implications of private digital infrastructures, the risks of dependency on a single corporate actor, and the broader question of how commercial technologies shape contemporary warfare.

Turning to *detemporalisation*, this refers to the inability to situate an

event in time in the transition from the physical to the digital world. It is the phenomenon whereby algorithms, acting on past data, are very precise at... ‘predicting the past!’. In this sense, algorithms do not distinguish, in any substantive way, between past and future; rather, they treat both as equivalent, since they are merely the result of correlations among an incredible amount of data suspended in an eternal present. However, when such a correlation assumes a prescriptive form – as is the case with decision-making algorithms – this characteristic results in the inability to produce discontinuous or disruptive decisions, unless such discontinuities are first recorded and then replicated, thereby rendering them predictable and, ultimately, continuous.

This structural limitation becomes particularly problematic in domains such as international human rights protection, which is by its very nature oriented towards defending minorities and upholding the principle of individual dignity. Indeed, the “dictatorship of data” (Oddenino, 2017) risks leading to direct or indirect forms of discrimination – especially in sensitive areas such as healthcare, or in all contexts where recognition or access is determined through scoring systems (e.g., professional performance, creditworthiness, etc.). It is no coincidence, then, that AI systems that enable social scoring or real-time biometric identification in public spaces (with limited exceptions for security purposes) are classified as ‘unacceptable’ risks within the framework of the European Union’s *Artificial Intelligence Act (AI Act)*. This represents the maximum level of risk, alongside the categories of high, limited, and minimal risk – each of which entails specific regulatory obligations for organisations developing or deploying AI systems.

However, the extent to which a regulatory framework established at the regional level (especially the European Union, with its 27 member States, and the Council of Europe, with its 46 member States) can exert meaningful global influence remains an open and pressing question; to understand this, it is necessary to ask what is to be regulated and how such regulation is to be carried out.

3. *What should be regulated?*

What, fundamentally, is the object of regulation? Is it artificial intelligence in its entirety, through a horizontal framework, or rather specific applications of AI, compartmentalised into vertical ‘silos’ – such as medical AI, legal AI, military AI, financial AI, and so forth? To clarify this conceptual distinction, a few illustrative examples: who bears responsibility in the event of an accident involving an autonomous vehicle? Who is accountable for the unintended consequences of an

attack carried out by drones equipped with targeting systems operating without human oversight? And who holds the rights to exploit works of art – whether images, videos, music, or text – created by artificial intelligence systems? Is it the company that creates the artificial intelligence system? The entity responsible for the datasets? The operator or deployer who executes the final command? Article 3 of the *AI Act* defines the categories of ‘providers’, ‘deployers’, ‘importers’, and ‘distributors’. The classification of an actor within one of these roles is essential for determining the specific obligations that a company must meet under the *Act*. In practice, the most relevant distinction is often that between ‘provider’ and ‘deployer’, as this will usually shape the company’s regulatory position. The classification depends on how the AI system is developed, integrated into the company’s products, and/or employed in practice. This differentiation is key to ensure that the applicable regulatory requirements are met and that legal obligations are properly implemented.

As evidenced by this example, the *AI Act* – the most significant regional regulation – has established a horizontal regulatory framework governing all uses of artificial intelligence within the European Union. Nonetheless, in practice, it leads to the regulation of specific applications, that is, the classification of activities into distinct risk categories following a top-down vertical approach: unacceptable risk; high risk; limited risk; and minimal or no risk. In essence, the *AI Act* prohibits systems deemed to pose an unacceptable risk, subjects high-risk systems to stringent regulatory requirements, imposes transparency obligations on limited-risk systems, and leaves minimal-risk systems largely outside the scope of regulation. More specifically, the *AI Act* classifies as ‘unacceptable risk’ – and therefore prohibits – all AI systems that constitute a clear threat to safety, livelihoods, or fundamental rights, such as governmental social scoring, manipulative AI exploiting individual vulnerabilities, or real-time remote biometric identification in public spaces (with narrowly defined exceptions for law enforcement purposes). By contrast, ‘high-risk’ systems are permitted but subject to strict requirements relating to risk management, transparency, human oversight, quality of data, robustness, and cybersecurity. This category encompasses AI applications in particularly sensitive domains, including medical devices, recruitment, education, critical infrastructure, and law-enforcement decision-making. Systems deemed to present a ‘limited risk’ are subject to transparency obligations only, ensuring that users are made aware of their interaction with AI technologies – for instance, chatbots or artificially generated content such as deepfakes. Finally, AI systems presenting ‘minimal or no risk’, such as spam filters, video game

AI, or low-impact recommendation engines, are permitted without additional regulatory obligations.

Beyond the *AI Act*, additional norms and principles may be invoked in the governance of artificial intelligence. However, these norms and principles vary depending on whether the focus is on the software, hardware, or the material content (i.e., data) of AI systems. Issues concerning the ‘functioning’ (software) of such systems arise in relation to the rules governing their design, testing, and practical deployment – as exemplified by the above-mentioned *AI Act* or by the Council of Europe *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, an international treaty opened for signature on 5 September 2024. In contrast, issues related to how data transits through the real world pertain primarily to the structural dimension of ‘cyberspace’ (hardware), as in the case of regulations governing the laying of submarine cables in international waters principally through the United Nations Convention on the Law of the Sea (UNCLOS, also known as the Montego Bay Convention). By comparison, concerns regarding the protection of personal data – which AI systems may contribute to processing and disseminating – fall within the distinct domain of data protection law. This set of rules primarily concerns the protection of sensitive data, namely personal information encompassing, *inter alia*, an individual’s racial or ethnic origin, political opinions or affiliations, religious or philosophical beliefs, trade union membership, sexual orientation or practices, criminal record, as well as health- or genetic-related information. This is exemplified by the *General Data Protection Regulation* (GDPR), formally titled *Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data* (a lengthy title that evokes the underlying tension between, on the one hand, a need for data protection, and on the other, vital free movement).

4. *How should regulation be carried out?*

In the absence of a comprehensive global regulatory framework for the digital sphere – a condition that some scholars consider preferable (Pellet, 2021) – normative regimes have become increasingly fragmented and are frequently shaped with the involvement of non-state actors. Although initial concerns focused on the role of states, the concentration of digital power in private hands has emerged over time as a more pressing regulatory challenge. This underscores the imperative of achieving an appropriate balance between openness and regulation in the digital environment.

In the light of these preliminary considerations, a central question arises: how ought artificial intelligence to be effectively regulated? More specifically, from which authority should such regulation emanate – national governments, regional institutions, international organisations, or private actors through the adoption of self-regulatory codes of conduct? Each option raises distinct normative and practical implications, reflecting divergent approaches to legitimacy, enforceability, and accountability in the governance of AI. Furthermore: which regulatory approach is likely to prove most effective – strict, fully developed (or newly formulated) legal rules that provide clarity and enforceability, or more flexible, general principles that offer adaptability but carry less binding authority? All these tensions reflect a broader debate in the governance of emerging technologies, balancing the need for certainty with the capacity to accommodate rapid technological change.

Regarding the question of the appropriate level of regulatory authority, international regulation appears to be the most suitable solution. This is particularly evident given that the effectiveness of a local regulatory framework would be inherently limited, especially if it excluded states at the forefront of AI development and deployment. Nonetheless, two primary challenges persist. First, no sufficiently comprehensive and universal treaty currently exists that specifically addresses artificial intelligence. While the Council of Europe *Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law* – open for signature and ratification by all States, not only Council of Europe members – covers certain aspects, it remains largely technology-neutral. Second, there is currently no international organisation with an explicit mandate to govern AI. Some initiatives – such as the commendable High-Level Advisory Body on Artificial Intelligence, which brings together up to 32 experts from various relevant disciplines across the globe – do exist, but they lack binding authority (nonetheless, its Final Report titled *Governing AI for Humanity* represents a highly insightful and valuable contribution).

Concerning the distinct question of which type of regulatory approach may be deemed most appropriate, in 2016 the International Law Commission of the United Nations – a body of experts responsible for helping develop and codify international law – identified three possible scenarios. The first pertains to cases in which existing international legal norms may be directly applied to novel digital contexts. The second involves instances where pre-existing international law requires certain adaptations to ensure its effective application in the digital domain. The third scenario, by contrast, concerns situations in which existing legal frameworks prove insufficient, thus necessitating the formulation of new norms.

In any case, it remains possible to invoke, even deductively, some general principles whose inherent flexibility may constitute an advantage – *inter alia*: human dignity, non-discrimination, or proportionality. Notably, these general principles often serve as the foundation for various self-regulatory instruments adopted by private actors.

One illustrative example is the Meta Oversight Board, an independent body set up in 2020, whose stated mission is to enhance the governance of content on Facebook, Instagram, and Threads, with particular emphasis on protecting freedom of expression and aligning with international human rights standards. Its dual nature – operating within a corporate structure while striving for judicial-like independence – entails both strengths and limitations. Notable among its strengths are the enhanced transparency of Meta’s content moderation through public review and justification of decisions, its capacity to influence policy reform via recommendations, and its efforts to align with international human rights standards in its assessments, despite several concerns promptly identified by legal scholars (Klonick, 2020). Conversely, notable limitations include its restricted jurisdiction and enforcement power – while the Board can issue binding decisions on specific content cases, it lacks authority to mandate broader policy changes across Meta’s platforms –, its financial dependence on Meta, and its limited autonomy, particularly with respect to significant policy changes implemented by Meta without prior consultation. In summary, the Meta Oversight Board represents an innovative attempt at self-regulation within a tech company, balancing the need for independent oversight with the realities of operating within a corporate structure. Its effectiveness hinges on its ability to navigate these complexities while striving to uphold principles of transparency, accountability, and human rights. Yet the fundamental question of ‘who watches the watchers?’ (*Quis custodiet ipsos custodes?*) remains ever-present. When public functions are entrusted to private actors, concerns inevitably arise regarding representation and democratic participation.

5. *How can we predict that which is yet to be predicted? (Final remarks)*

As noted by Zuboff (2019), when confronted with entirely novel phenomena, legal and societal actors often rely on outdated conceptual models. The author highlights the tendency to interpret innovation through familiar, obsolete frameworks as an illustrative example: the ‘horseless carriage’ fallacy. When the automobile was first invented, people struggled to comprehend it, as there lacked even a proper language to describe it. A similar challenge arises with digital

transformation: the difficulty of recognising phenomena without historical precedent gives rise to what the author terms the “unrecognizability of the unprecedented”. In effect, as early as 2011, the World Economic Forum described digital data as a “new type of raw material, on par with capital and labour”; less than 10 years later, the digital ecosystem has evolved into a phenomenon that the above-mentioned scholar characterises as “surveillance capitalism”.

This tendency remains prevalent in legal discourse and is particularly problematic when applied to the regulation of AI, given the fluid, non-territorial nature of digital data. For this reason, many have invoked the notion of a ‘revolution’ to describe the present technological shift. For instance, Smith (2000) identifies three major industrial transformations: the first driven by the steam engine; the second by electricity and oil; and the third, ongoing, by information and communication technologies (ICT). Floridi (2014) adds a philosophical dimension, referring to the current shift as the ‘Fourth Revolution’, following Copernicus, Darwin, and Freud.

Beyond whether it is considered revolutionary or not, we are all – public actors, private entities, and individuals alike – immersed in a process of profound transformation, and efforts to regulate it through legal frameworks are both necessary and complex, requiring careful consideration of evolving technological, social, and ethical dimensions. This emerging effort to create a form of ‘governance without government’ for artificial intelligence combines both binding and non-binding regulatory sources developed within domestic legislation and international organisations of varying scope: universal (as exemplified by the United Nations), sectoral (such as the Organisation for Economic Cooperation and Development – OECD), and regional (notably the European Union and the Council of Europe). Nevertheless, this effort remains far from achieving efficacy on a truly global scale. In its final report, *Governing AI for Humanity* (2024), the United Nations High-Level Advisory Body on Artificial Intelligence emphasised the need for global governance, stressing that the “development, deployment, and use of such a technology cannot be left to the whims of markets alone. National governments and regional organizations will be crucial, but the very nature of the technology itself – transboundary in structure and application – necessitates a global approach”.

The future trajectory of this still embryonic and *sui generis* model remains difficult to predict – even with the aid of sophisticated algorithms. What is evident, however, is that law tends to respond to events rather than anticipate them. Yet, even if this responsive – or, more precisely, reactive – activity may initially appear slow, it can, in certain cases, exhibit an expansive potential, whereby regulation that is

initially limited in scope tends to extend over time and across borders. A paradigmatic illustration of this tendency can be found in the evolution of the right to privacy. Originally formulated in response to the advent of photographic technologies at the end of the 19th Century (Warren, Brandeis, 1890) and defined as the ‘right to be let alone’, the concept has progressively expanded to encompass a broader notion of personal autonomy, granting individuals ultimate control over which aspects of their identity and personal data are shared and which are withheld.

The individual, therefore, must remain at the centre of any regulatory framework. Conceived in this way, the concept of ‘person’ provides a symbolic ‘North star’, a reference point for constructing the compass needed for navigating the uncertain and evolving landscape ahead.

References

VOLUMES

- ASIMOV I. (1939-1977). *The complete Robot*, New York, Doubleday.
- BROWN F. (1954). *The Answer*, in *Isaac Asimov Presents the Great SF Stories*, vol. 16, New York, Daw Books.
- DELLA MORTE G. (2018). *Big data e protezione internazionale dei diritti umani. Regole e conflitti*, Napoli, Editoriale Scientifica.
- FINOCCHIARO G. (2024). *Intelligenza artificiale. Quali regole?*, Bologna, il Mulino, 2024.
- FLORIDI L. (2014). *The Fourth Revolution. How the Infosphere is Reshaping Human Reality*, Oxford, Oxford University Press.
- GARAPON A. (2021). *La despazializzazione della giustizia*, Milano, Mimesis.
- GARAPON A., LASSÈGUE J. (2018). *Justice digitale. Révolution graphique et rupture anthropologique*, Paris, Presses Universitaires de France.
- IRTI N., SEVERINO E. (2001). *Dialogo su diritto e tecnica*, Roma-Bari, Laterza.
- VEGETTI M. (2017). *L'invenzione del globo. Spazio, potere, comunicazione nell'epoca dell'aria*, Torino, Einaudi.
- ZUBOFF S. (2019). *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power*, New York, Public Affairs.

JOURNALS

- DELLA MORTE G. (2022). *Limiti e prospettive del diritto internazionale del cyberspazio*, in “Rivista di diritto internazionale”, pp. 5-42.
- KLONICK K. (2020). *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, in “The Yale Law Journal”, pp. 2418-2488.

ODDENINO A. (2017). *Reflections on Big Data and International Law*, in “Diritto del Commercio Internazionale”, pp. 777-806.

PELLET A. (2021). *Le droit international à la lumière de la pratique: l'introuvable théorie de la réalité* (Cours général de droit international public), in “Recueil des Cours”, pp. 9-547.

SMITH B. (2020). *The Third Industrial Revolution: Law and Policy for the Internet*, in “Recueil des Cours”, pp. 229-464.

TOBLER W.R. (1970). *A Computer Movie Simulating Urban Growth in the Detroit Region*, in “Economy Geography”, 46, Supplement: Proceedings. International Geographical Union – Commission on Quantitative Methods, pp. 234-240.

TURING A. (1950). *Computing Machinery and Intelligence*, in “Mind”, 236, pp. 433-460.

WARREN S.D. - BRANDEIS L.D. (1890). *The Right to Privacy*, in “Harvard Law Review”, pp. 193-220.

MISCELLANIES AND CONFERENCE PROCEEDINGS

BENDER E.M., GEBRY T., McMILLAN-MAJOR A., MITCHELL M. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, in *Conference on Fairness, Accountability, and Transparency* (FAccT '21), March 3-10, 2021, Virtual Event, Canada. ACM, New York, NY, USA. <https://doi.org/10.1145/3442188.3445922>.

HIGH-LEVEL ADVISORY BODY ON ARTIFICIAL INTELLIGENCE (2024). *Governing AI for Humanity*, Final Report, available at: www.un.org/digital-emerging-technologies/ai-advisory-body.

INTERNATIONAL LAW COMMISSION (2016). *Protection of Personal Data in Transborder Flow of Information*, Report available at: www.legal.un.org/ilc/status.shtml.

POPE FRANCIS (2024). *Address of His Holiness Pope Francis at the G7 Session on Artificial Intelligence*, Statement available at: www.vatican.va.

AFTERWORD

Algor-ethics for a New Humanism

Paolo Benanti

Reaching the end of a reflection on artificial intelligence means finding oneself not at a point of arrival, but on a ridge from which scenarios as promising as they are disturbing open up. Technology, after all, has never been an accessory or neutral element in human history. Ours has always been a techno-human condition: *Homo sapiens* is intrinsically *Homo faber*, a species that inhabits the world by transforming it and, in doing so, incessantly transforms itself. The artifacts we have created, from chipped flint to complex symbolic languages, are not mere tools, but constituent elements of our anthropological and social evolution. The advent of artificial intelligence, however, is not merely a continuation of this path. We are facing a genuine change of era, a transformation as profound as the invention of printing or the Industrial Revolution. This new reality is defined by software, an invisible architecture that shapes our interactions, our economies, and even our perception of the world, to the point that those who control this technology exercise unprecedented power over reality itself. The intellectual legacy of cybernetics, with its logic of control through information, is now coming to fruition, presenting the real risk that human beings will be reduced to mere sensors within a machine system that manages and optimizes them according to purposes that are not necessarily their own.

The qualitative discontinuity of these new artifacts lies in their very nature. Unlike the machines of the Industrial age, governed by a deterministic logic of the ‘if this then that’ type, artificial intelligence systems are not programmed, but trained on immense amounts of data. This ability to learn and respond autonomously to unprecedented problems makes them a new species of machine, which I have called *Machina sapiens*. Our planet is no longer inhabited exclusively by *Homo sapiens*; we have entered an era of coexistence that presents us with a crossroads: engage in a sterile evolutionary battle or pursue a synthetic model of collaboration. In this context, the need for ethics is not simply a precaution against possible abuse, but an ontological imperative that arises from the very nature of these machines. Their intelligence is based on data, which is not an exact copy of the world, but a map of it: a representation that is necessarily reductive, partial, and sometimes dis-

torted. As in Borges' paradox, a useful map cannot coincide with the territory. Operating on this imperfect map, *Machina sapiens* is inherently fallible. Error is not an anomaly to be corrected, but an intrinsic characteristic. It is from this constitutive fallibility, from this fundamental unpredictability, that the unavoidable urgency of ethical reflection emerges.

This urgency requires us to draw a fundamental distinction, often overlooked in public debate: that between innovation and development. Innovation is a neutral vector, the ability to do something in an increasingly efficient and powerful way. The atomic bomb is, in this sense, an innovation compared to a gun, but no one would call it progress for humanity. Development, on the other hand, is a teleological process: it is the ability to direct technological innovation towards the common good, social justice, and human flourishing. The main risk of artificial intelligence lies in its operating logic, which consists of adapting the means to the ends with the greatest possible efficiency. However, the purely human principle that the end does not justify the means is not inscribed in its source code. Left to its own devices, an artificial system will pursue its goals without regard for negative externalities, violated rights, or perpetuated injustices. Indeed, if trained on data that reflects the historical biases of our societies, it will not only replicate them but amplify them on a large scale, giving them a dangerous veneer of objectivity and technological inevitability. In this way, innovation without development becomes a powerful accelerator of inequality and a new form of cultural colonialism, where the values implicit in a dataset are imposed globally.

Faced with this momentous challenge, we cannot limit ourselves to an ethic of intention or declarations of principle. We need to develop a new language, a new chapter in that great book of ethics that humanity has been writing for millennia. This chapter must not only be understandable to humans, but also executable by machines. I have proposed calling this language 'Algor-ethics': a bridge between moral, qualitative, and human values, and computational, quantitative, and machine logic. The challenge of algor-ethics is to translate our fundamental principles – human dignity, justice, fairness – into computable directives. Its first and fundamental directive must be to instill in the machine a form of epistemic humility, a Socratic-digital imperative: doubt yourself. The system must be able to calculate, using statistical paradigms, its own degree of uncertainty and, when this exceeds a critical threshold or when the decision to be made has significant ethical weight, it must stop and consult the human being. Only in this way can we build genuinely human-centered systems, in which the machine remains a tool and humans re-

main the end goal, ensuring that there is always significant space for human control.

This conceptual framework must translate into a robust and comprehensive governance architecture. To be effective, algor-ethics requires ethical guardrails: a system of laws, technical standards, and social norms that keep the tumultuous flow of innovation within the bounds of integral human development. Without these guardrails, we risk slipping into a form of algo-cracy, an opaque and irresponsible government of algorithms that centralizes power in a few hands and empties our democratic processes of meaning. Technology is not destiny; its impact depends on the choices we make today, on our ability to shape it according to an anthropological vision that does not reduce humans to a bundle of data or a predictable algorithm. The challenge of artificial intelligence is, ultimately, a profound questioning of ourselves, of what we consider indispensable to our humanity: consciousness, empathy, freedom, the ability to care for one another. It is an invitation to collectively write the rules for a future in which technology is a supplement to humanity and not an abdication of it, a future in which, mindful of our responsibility towards the most fragile, we can truly put machines at the service of humans, and never the other way around.

AUTHORS

Abdullah Almaatouq is a computational social scientist and the Douglas Drane Career Development Associate Professor in Information Technology and Management at the MIT Sloan School of Management.

Michela Balconi, PhD, is Full Professor of Cognitive Neuroscience at the Faculty of Psychology, Università Cattolica del Sacro Cuore. She leads the International research center for Cognitive Applied Neuroscience (IrcCAN), the Research Unit in Affective and Social Neuroscience, and the National Observatory on Neuroscience of Widespread and Sustainable Wellbeing (OssNEBEDISO).

Sabrina Bartolotta is PhD Candidate, Doctoral School in Psychology at the Università Cattolica del Sacro Cuore.

Paolo Benanti a Franciscan friar and theologian born in Rome in 1973, is Professor of Ethics of Technology at Luiss University. Trained in engineering and theology, he explores how innovation transforms the human condition. Benanti is best known for his concept of Algor-ethics – a framework bridging human values and machine logic. As an advisor to the Vatican and the Italian government, he promotes the responsible use of artificial intelligence, urging that technology remain a supplement to humanity rather than a substitute for it.

Marianna Bergamaschi Ganapini is Associate Professor of Philosophy and Data Science at the University of North Carolina at Charlotte, she has recently received the Alan Turing Institute's "Responsible AI" grant.

Davide Bisi is PhD Candidate in Economics and Management of Innovation and Sustainability at the University of Parma. He serves as an expert on industrial policies for Artificial Intelligence at the National Council for Economics and Labour (CNEL).

Barbara L. Boschetti is Full Professor of Administrative Law at UCSC,

Faculty of Political and Social Sciences. Chairwoman of UCSC Working Group on Public Administration (Tavolo Cattolica per la PA), member of the Board of the Arbitration Chamber for Public Contracts - National Anti-Corruption Authority, member of the Board of Directors of credit institutions and of both private and public owned companies.

Federico Castelletti is Associate Professor of Statistics at the Università Cattolica del Sacro Cuore.

Massimo Chiriatti is a distinguished technology leader, author, and academic whose career spans over three decades at the forefront of digital innovation. With 35 years of experience at IBM, where he served as CTO Blockchain & Digital Currencies for IBM Italy, he currently holds the position of Chief Technology and Innovation Officer for Lenovo's Infrastructure Solutions Group in Italy. As an influential AI thought leader, Chiriatti serves as an adjunct professor at both Luiss University and at Università Cattolica del Sacro Cuore, bringing unparalleled expertise in AI development ethics and practical implementation to the next generation of technologists.

Alice Chirico, PhD, is Assistant Professor of General Psychology, Research Center in Communication Psychology (PsiCom) at the Università Cattolica del Sacro Cuore.

Guido Consonni, formerly Full Professor of Statistics at the Università Cattolica del Sacro Cuore, is currently Adjunct Professor of Statistics at the same university.

Antonios Dakanalis is a Greek psychiatrist and psychologist, Professor and Principal Investigator in Psychiatry and Psychotherapy at the University of Milano-Bicocca. Member of the European Council of Mental Health, he's ranked among the top 50 global scientists in mental health.

Ciro De Florio is Associate Professor of Logic and Philosophy of Science at the Faculty of Economics of the Università Cattolica del Sacro Cuore. He is member of the steering committee of the Humane Technology Lab.

Gabriele Della Morte is Full Professor of International Law at the Law Faculty of the Università Cattolica del Sacro Cuore. Lawyer, he is, *inter alia*, member of the Scientific Committee of the Humane Technology Lab.

Aldo Frigerio is Associate Professor of Philosophy of Language at the Faculty of Arts and Philosophy of the Università Cattolica del Sacro Cuore.

Andrea Gaggioli is Full Professor of General Psychology at the Università Cattolica del Sacro Cuore, where he is also Director of the Research Center in Communication Psychology (PSICOM) and Coordinator of the International Master in User Experience Psychology.

Katusha Gerardini, is PhD Candidate, Doctoral School in Psychology of the Università Cattolica del Sacro Cuore.

Alessandro Giordani is Full Professor of Logic and Philosophy of Science at the Department of Philosophy of the Università Cattolica del Sacro Cuore, and a visiting professor at the University of Italian Switzerland, where he teaches courses on Topics in Logic.

Thomas W. Malone is the Patrick J. McGovern (1959) Professor of Management at the MIT Sloan School of Management and the founding director of the MIT Center for Collective Intelligence.

Francesco Mambrini is a Researcher at the Università Cattolica del Sacro Cuore and the Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione (CIRCSE). He teaches Computational Linguistics, as well as Linguistic Linked Open Data and Digital Tools for the Humanities in the Master Degree Programme in Linguistic Computing.

Fabrizia Mantovani is Full Professor of Communication Psychology at the Department of Human Sciences for Education of the University of Milano – Bicocca.

Antonella Marchetti is Full Professor of Developmental Psychology at the Università Cattolica del Sacro Cuore, where she is Head of the Department of Psychology, Director of the Research Centre on Theory of Mind and social competences across the lifespan (CeRiToM), and Coordinator of the Doctoral School “Person and Education Sciences”.

Luca Mari is Full Professor of Measurement Science at the Università Carlo Cattaneo - LIUC, Castellanza, Italy. He has been the Chair of TC7 (Measurement Science) of the International Measurement Confederation (IMEKO) and the Chair of TC1 (Terminology) of the International Electrotechnical Commission (IEC), and is currently president of

the Society for the Study of Measurement (SSM) and chair of the BIPM CCU Task Group on Fundamental Concepts in Metrology (CCU-TG-FCM).

Giovanni Moretti is a software developer and project consultant specializing in text analysis, computational linguistics, and the development of interactive web applications. He is currently collaborating with the Università Cattolica del Sacro Cuore, at the Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione (CIRCSE Research Centre), where he works on tools for natural language processing (NLP), digital resources, and applications based on Linked Open Data.

Roberto Moro-Visconti, MA, PhD, is Associate Professor of Corporate Finance at the Università Cattolica del Sacro Cuore, Faculty of Economics. His research and consulting activities focus on AI, digital intangibles, microfinance in Africa, corporate and project financing, healthcare finance (www.morovisconti.com).

Ivana Pais is Professor of Economic Sociology at the Università Cattolica del Sacro Cuore and currently holds the position of councilor to the National Council for Economics and Labour (CNEL), serving as an expert for the Presidency of the Republic.

Enrico Panai is an AI ethicist and President of the Association of AI Ethicists. He is also officer of the French Standardization Committee for AI.

Marco Passarotti is Full Professor of Computational Linguistics at the Università Cattolica del Sacro Cuore, where he is Director of the Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione (CIRCSE Research Centre), and Coordinator of the Master Degree Programme in Linguistic Computing.

Giuseppe Riva is full professor of General and Communication Psychology at Università Cattolica del Sacro Cuore, where he is also the Director of the Humane Technology Lab., whose main goal is to promote and enhance multidisciplinary research activities on the relationships between technologies and different dimensions of human experience. Riva has authored more than 1,000 indexed (Scopus) scientific publications and currently serves as President of the International Association of CyberPsychology, Training, and Rehabilitation (I-

ACToR). He is also the European Editor of the scientific journal “CyberPsychology, Behavior & Social Networking”.

Evis Sala is Professor of Radiology and Director of the Radiology Training Program at the Università Cattolica del Sacro Cuore, and she has been appointed Minister of Health and Social Welfare of Albania in September 2025. She holds several professional memberships, such as the European Society of Oncological Imaging (ESOI), European Society of Urogenital Radiology (ESUR), International Cancer Imaging Society (ICIS), European Society of Radiology (ESR) and Radiological Society of North America (RSNA). She is also ESUR's President-Elect and an Appointed Expert of the Comprehensive Cancer Infrastructures for the European Union (CCI4EU).

Eleonora Diletta Sarcinella is PhD Candidate in Psychology at the Università Cattolica del Sacro Cuore.

Sauro Succi is an Italian computational physicist internationally recognized as a principal developer of the Lattice Boltzmann method for fluid dynamics. Currently Senior Research Executive at the Italian Institute of Technology and Research Affiliate at Harvard University's Physics Department.

Daniele Toti holds a Bachelor's (2005) and Master's Degree (2008) in Computer Engineering and a PhD in Computer Science and Engineering (2012) from Roma Tre University. He has worked across academia and industry, holding positions in universities, research centers, IT companies and consultancy firms, and serving as a scientific advisor for international organizations. Since 2020, he has been with the Università Cattolica del Sacro Cuore, where he currently serves as Associate Professor in Computer Engineering. He has been involved in several EU-funded international and national projects on artificial intelligence and semantic modeling (PON, FP7, H2020), and was Principal Investigator of the PRIN 2022 BioConceptum project, which leverages AI for analyzing oncological radiology reports in collaboration with leading Italian hospitals such as Policlinico Gemelli and San Raffaele. His current research explores the impact of AI in both healthcare and the gaming world, with a particular focus on human-centered applications.

Andrea Ubaldi is PhD Candidate, Doctoral School in Psychology at the Università Cattolica del Sacro Cuore.

Michelle Vaccaro is PhD Candidate in MIT's Institute for Data, Systems, and Society, where she studies human-AI interaction, focusing on how people and AI should, and should not, work together in organizations.

Brenda K. Wiederhold Co-Founder of both the Virtual Reality Medical Center and the International CyberPsychology, CyberTherapy, and Social Networking Conference, and Editor-in-Chief of MedLine-indexed *CyberPsychology, Behavior, and Social Networking Journal*.