

MEDIA, COMMUNICATION  
& SOCIO-CULTURAL PROCESSES

## Coordinator of the Series

Alessandra Micalizzi

PEGASO UNIVERSITY

## National Board

Manuela Farinosi

UNIVERSITY OF UDINE

Antonella Mascio

ALMA MATER STUDIORUM — UNIVERSITÀ DI BOLOGNA

Rosy Nardone

ALMA MATER STUDIORUM — UNIVERSITÀ DI BOLOGNA

Rebecca Paraciani

ALMA MATER STUDIORUM — UNIVERSITÀ DI BOLOGNA

Tiziana Piccioni

IUSVE UNIVERSITY OF VENICE

Rita Salvatore

UNIVERSITY OF TERAMO

Mariacristina Sciannamblo

“LA SAPIENZA” UNIVERSITY OF ROME

Rosantonieta Scramaglia

IULM UNIVERSITY

## International Board

Jelena Filipović

UNIVERSITY OF BELGRADE

Maria Koprivica Lelicanin

SAE INSTITUTE- BELGRADE

Laura Leon

UNIVERSITY OF PERU

Bojana Radenković Šošić

UNIVERSITY OF KRAGUJEVAC (SERBIA)

Eugenia Siopera

CITY UNIVERSITY OF DUBLIN

Maria Touri

UNIVERSITY OF LEICESTER

## MEDIA, COMMUNICATION & SOCIO-CULTURAL PROCESSES

*Creative media are contexts, catalysts and cultural technologies, playing a pivotal role in activating and directing contemporary phenomena that take place in our society. Communication processes and Cultural Practices book series meet the perspective of observing the social reality starting from the role of media and of communication's processes. Media, Communication and cultural processes, in fact, aims at being the publishing frame for editorial proposals, academic and with a strong attention to empirical research, that want to investigate contemporary phenomena looking at what happens concretely in our society and that involve individuals: as single person, group or community.*

*The research areas*

*Phenomenon, culture and subjectivity are the three main research points on media that guide the selection of the proposals. The starting point of the Communication processes and Cultural Practices book series' perspective is that it is not possible to identify clear and neat borders with in these three social constructs and that the richness of the contributions is represented by the contamination, contact and dialog among them. Moreover, it is the way to guarantee a multidisciplinary glance to contribute the "discover", the proposition of new analysis, enable to contribute to the dialog theories and tools of contiguous disciplines.*

I media creativi si presentano come contesti, catalizzatori e tecnologie culturali, svolgendo un ruolo centrale nell'attivazione/direzione dei fenomeni contemporanei che nella società prendono forma. Osservare la realtà sociale a partire dal contributo dei media e della comunicazione è la prospettiva che caratterizza la collana Media, Comunicazione e Processi culturali che intende fare da cornice per le proposte editoriali, di tipo accademico e con una forte attenzione alla ricerca empirica, volte a indagare fenomeni della contemporaneità a partire da ciò che accade nella società e coinvolge direttamente l'individuo: come singolo, come gruppo e come comunità.

### **Le aree di ricerca**

Fenomeno, cultura e soggetto sono i tre punti focali delle ricerche e degli studi sui media che trovano spazio all'interno della collana. Il principio di fondo è che la definizione dei margini di questi costrutti sia impossibile e che nei limen, nel contatto o intreccio, nella relazione tra di essi vi sia la ricchezza prospettica e interpretativa che possa garantire uno sguardo multidisciplinare e favorire la scoperta, la proposizione di analisi nuove, capaci di fare dialogare teorie e strumenti di discipline attigue.



# **Artificial Intelligence and Social Research: Methods, Contexts, Imaginaries**

edited by Alessandra Micalizzi





CC 4.0 International

Attribution-Non Commercial-No  
Derivatives

DOI 10.69146/55440895

Printed edition

Copyright WriteUp Books© 2025  
via Michele di Lando, 77 — Roma

ISBN 979-12-5544-089-5

ISSN 3103-554X

[www.writeupbooks.com](http://www.writeupbooks.com)

[redazione@writeupbooks.com](mailto:redazione@writeupbooks.com)

1<sup>st</sup> edition: December 2025

## Table of Context

- 9 *Artificial Agencies in Research. Uses, Dilemmas, and Epistemic Shifts. An introduction* [Alessandra MICALIZZI]
- 21 **The Role of Generative AI in Qualitative Data Analysis: Opportunities and Limitations in Supporting Dissertation Supervision in Academia** [Leonard BUSUTTI and Rosienne CAMILLERI]
- 57 **Minimum Thresholds of Relationship: AI and Autobiographical Writing** [Lara BALLERI]
- 79 **Moving Beyond Text: A Comparative Case Study of AI-assisted Audio Interviews in Relation to Textual Data from Mobile Diaries in Singapore** [Nadia OLISA and Azaleah MOHD ANIS]
- 97 **Human-Machine Feedback Loops in Qualitative Research: Co-Constructing Semi-Structured Interviews with Generative AI** [Giulia COPPO]
- 125 **Talking with AI about trust in health topic: an explorative research** [Alessandra MICALIZZI and Caterina SAPONE]
- 149 **Beyond Bias: Understanding Social Representations Embedded in Generative AI Outputs** [Elisabetta RISI]
- 175 **Framing AI in the audiovisual industries on LinkedIn** [Anouck Butraud-ASSATHIAN, Jaércio DA SILVA and Cécile MÉADEL]
- 201 **AI Ethnography: A Methodological Proposal for the Analysis of Vernacular Prompting Practices** [Gabriella TADDEO]
- 225 **AI-Augmented Anticipatory Ethnography: Envisioning, Design Fiction and Generative AI for Co-Creating Eutopias** [Agnese VELLAR and Matteo FOGLI]
- 247 *Bios*



# Artificial Agencies in Research

## Uses, Dilemmas, and Epistemic Shifts

by Alessandra MICALIZZI

### How Using AI: Roles, Practices, and Effects

Artificial Intelligence (AI) represents a pervasive and transformative force in contemporary society, reshaping sectors ranging from healthcare to finance, from security to communication. Its growing integration into everyday life makes an in-depth analysis of the dynamics governing its development strategically essential. At present, this development is defined by two parallel yet often divergent strands of research: on the one hand, the technical advancement of algorithms and computational models; on the other, the critical inquiry into their profound social, ethical, and political implications.

In this increasingly complex scenario, enriched by interdisciplinary approaches, the social and human sciences seek to make their specific contribution—generally focusing on how AI usage affects social and cultural practices. From this perspective, AI is seen as the object of research, as part of the very social phenomena under investigation. Social scientists are thus interested in understanding the cultural, ethical, and economic impacts of AI within the broader system of social relations.

Beyond, and even prior to, considering AI as an object of research, a more compelling debate concerns AI as a tool of research. Since the early 1990s, experiments involving the hybridization of AI with traditional and systematic analytical techniques have prompted reflection on the implications of this integration. On one side, the gains in efficiency—both in time and in the allocation of cognitive and computational resources—are undeniable; on the other, the researcher's control over the process tends to decrease, and the growing complexity of such systems becomes

harder to capture and interpret.

Only recently —and with increasing urgency— has AI been reconsidered by the human sciences in new socio-cultural roles: as a context of research, a space where phenomena unfold, new practices emerge, and novel interactions negotiate fresh rules; as a partner in research, a semi-autonomous assistant for testing the solidity and validity of analytical models or theoretical constructs; as an expert, capable of selecting sources and producing preliminary drafts for state-of-the-art reviews on specific topics; as a servant, to which mechanical or repetitive tasks can be delegated; as a key informant, able to provide the researcher with detailed, insider-like knowledge about the organization, values, and practices of specific social groups (Bernard, 2017); and finally, as a cultural broker, an intermediary facilitating communication and translation between diverse cultural domains within ethnographic research, thanks to its extensive —albeit synthetic— knowledge of the object of study (Hammersley & Atkinson, 2019).

Delving deeper into this complex human-machine interaction within applied research, recent scenarios have also highlighted the potential use of AI as a respondent. This technology can interrogate vast datasets—not only academic materials but, more importantly, user-generated content (UGC). Yet this possibility raises critical questions: is it methodologically sound, ethically acceptable, and epistemologically valid to treat AI as a representative of the “average profile” of a given culture? What implications does this have for the quality and authenticity of data?

The process of datafication, as described by Pop Stefanija and Pierson (2020), consists in the transformation of social actions into quantifiable data. These data constitute the base of algorithmic identities —profiles that categorize individuals based on inferences and correlations. It is essential to emphasize, as the authors do, that this is «not the personal identity of the embodied individual, but rather the actuarial or categorical profile of the collective». The algorithmic identity is therefore a statistical construct applied to a profile, rather than a reflection of their lived

identity, and it is employed to make decisions that have tangible consequences on people's lives—from credit access to content personalization.

In the multiplicity of these roles, the conversational metaphor remains the common thread defining the human–machine relationship, suggesting the effectiveness of Esposito's (2022) expression *Artificial Communication*. What binds these diverse applications together is the dialogic structure through which humans and algorithms co-construct meaning—an interaction that alternates between cooperation and delegation, understanding and automation. The researcher is therefore invited to experience this new form of machine agency as both a resource and a challenge—at times a productive temptation, at others a potential risk. In this context, the classic Van Maanen advice retains its relevance: when studying a social object, the only safeguard lies in the researcher's ability to look at it with the dual gaze of the Martian—estranged and critically distant—and the convert—immersed within the processes and practices being observed (Van Maanen, 1988)

### **Developing new perspectives: towards a real integrated model**

Despite the impressive advances in machine learning and neural network research, a purely technical approach to Artificial Intelligence faces intrinsic limitations that are not technological in nature, but rather epistemological and social. These limitations become evident when algorithmic systems interact with complex human contexts, where predictive efficiency is not the only value at stake. This section examines two of these crucial limitations: the unresolved trade-off between predictive performance and explanatory transparency, and the persistent tendency to address inherently social problems with technically intuitive but epistemologically fragile solutions.

The fundamental dilemma of contemporary AI is effectively

illustrated by Di Franco & Santurro (2020). On one hand, machine learning algorithms such as Artificial Neural Networks (ANNs) demonstrate predictive performances (apparently) far superior to traditional statistical models (e.g., linear regression) when applied to social data, as they can identify non-linear patterns and complex interactions that conventional methods often miss. On the other hand, these powerful tools operate as “black boxes.” Their critical weakness lies in their inability to contribute to the understanding of the internal relationships among the system’s individual components. While a traditional regression model provides interpretable coefficients that suggest causal relationships, a neural network produces accurate predictions without explaining the why. This opacity, although technically efficient, becomes a major obstacle to critical evaluation, contestability, and accountability: if we cannot understand how a system arrives at its conclusions, it becomes impossible to ensure that those conclusions are fair, justified, and free from illegitimate bias.

The field of Explainable AI (XAI) was established precisely to address the black box paradox. However, as Tim Miller (2019) points out in a foundational critique, this endeavor suffers from an intrinsic limitation: the technical community attempts to solve a profoundly social problem—what constitutes a good explanation for a human being—by relying primarily on the researchers’ own intuition. This approach neglects decades of well-established research in philosophy, psychology, and cognitive science on how people define, generate, select, and evaluate explanations. Studies in these disciplines have shown, for instance, that effective human explanations are often contrastive (they clarify why event X occurred instead of event Y) and selective (they emphasize the most relevant causes for the listener while omitting less significant ones). Ignoring these insights leads to the design of XAI systems that produce explanations which are technically correct but humanly irrelevant or even misleading. The case of XAI exemplifies a broader trend within the technical field to “reinvent the wheel” when confronted with social problems, re-

vealing a deep epistemological limitation that only genuine, interdisciplinary collaboration can overcome.

The social sciences do not confine themselves to a merely deconstructive critique; rather, they actively develop new conceptual and methodological frameworks to study Artificial Intelligence in a rigorous and innovative way. Lindgren & Holmström (2020) propose four “building blocks” for a social science of AI. These pillars call for:

- studying humans and machines within their sociocultural contexts;
- considering AI agents as social actors (in line with Actor–Network Theory);
- analysing AI as a social construction, shaped by discourse and power relations; and
- developing a critical reflection on research methods themselves, given their dependence on platforms and “formatted” data.

In response to the opacity of algorithmic systems, Mohammad Hossein Jarrahi (2025) proposes a qualitative framework based on the idea of “interviewing AI.” This approach adapts ethnographic techniques to explore the emergent and unpredictable behaviours of AI systems. Methods such as systematic probing —posing structured sequences of questions to test the limits and biases of a model— and temporal analysis —asking the same questions over weeks or months to observe how responses evolve— enable researchers to uncover internal patterns and logics that would remain invisible through purely quantitative analysis.

This critical and methodological toolkit developed within the social sciences thus highlights not only the external problems generated by AI but also the intrinsic limitations of a purely technical perspective.

The first step toward bridging the divide is to deconstruct the very idea that there exists a clear separation between the “technical” and the “social.” The concept of coproduction, drawn from Science and Technology Studies (STS) and highlighted by Gov-

ia (2020), provides the ideal conceptual foundation for this purpose. Coproduction posits that technologies and social orders continuously and reciprocally shape one another in an ongoing, inextricable process.

Decisions that appear to be merely “technical” —such as the choice of an algorithm, the structuring of a dataset, or the design of an interface— are in fact infused with social values, assumptions, and priorities. In turn, once implemented, technologies reconfigure social practices, identities, and power relations. Recognizing this intrinsic interconnection means acknowledging that there is no purely technical domain isolated from the social world. This awareness transforms interdisciplinary dialogue from an encounter between separate worlds into a necessary collaboration for understanding and governing a unified sociotechnical system.

The field of Explainable AI (XAI), previously discussed as an example of the limits of a purely technical approach, also represents an exemplary area of potential collaboration. As Miller (2019) argues, rather than relying on intuition, engineers can — and must— integrate decades of research from the social sciences to design systems that provide explanations genuinely effective for human understanding.

This constitutes a model of collaboration in which social research moves beyond ex-post critique to make an ex-ante contribution to design. The benefits of such an integrated approach are manifold:

- **Enhancing Effectiveness:** Integrating models of human cognition from psychology and cognitive science allows for the creation of explanations that are truly comprehensible, relevant, and useful to end users—going beyond the mere presentation of technical features.
- **Building Trust:** Designing systems whose explanations take into account social expectations and human cognitive biases fosters user adoption and trust. A “good” explanation from a human perspective can increase acceptance more effectively than a “complete” but incomprehensible one.
- **Moving Beyond Intuition:** The interdisciplinary approach re-

places developers' subjective intuition with rigorous, empirically validated methods—derived from the social sciences—for defining and evaluating what makes an explanation effective and satisfactory.

### **A step forward: the purpose of this book**

Building upon the theoretical premises outlined above, the reflections gathered in the following chapters seek to take a further step—moving from conceptual analysis to practical experimentation. Each contribution explores how artificial intelligence can operate not merely as an analytical tool—whose reliability and epistemic validity are still under scrutiny—but as a multifaceted actor within the research process itself. By leveraging its conversational capabilities and its synthetic power to reconstruct patterns from dispersed online data, AI can be a partner in inquiry, capable of supporting interpretation, comparison, and theoretical innovation.

AI is not simply a tool or object of analysis, but a social actor that co-constitutes the conditions of knowledge. It intervenes in research design, data collection, interpretation, and dissemination, while simultaneously reconfiguring our epistemic frameworks.

The studies presented here engage with this dual potential, offering empirical applications and methodological reflections that illustrate both the opportunities and the limitations of integrating AI into qualitative and hybrid research settings. While acknowledging the persistent challenges and critical resistances that accompany such integration, the volume aims to provide constructive insights for a debate that the academic community can no longer defer—a debate concerning how artificial intelligence reshapes not only our methods, but also our very ways of knowing, questioning, and representing the social world.

Understanding this co-production between human and machine requires methodological innovation and reflexivity—two guiding principles that inform all the contributions included in

this volume.

The opening chapter, written by Leonard Busuttil and Rosinette Camilleri and focused on the *Role of Generative AI in qualitative data analysis* introduces the reader to the pragmatic and ethical challenges of integrating AI into academic supervision. By examining how AI can assist both students and supervisors in conducting qualitative research, the author critically reflects on the dual nature of these technologies —as enablers of analytical depth and as potential sources of dependency and bias. This chapter sets the stage for the book by positioning AI not merely as an instrument for efficiency but as an epistemic companion whose presence redefines scholarly authority and reflexive practice.

The second contribution extends the discussion from research methodology to creative self-expression. Lara Balleri explores the dialogical dimension of human–AI co-authorship. The chapter proposes that writing with AI exposes the affective, cognitive, and ethical thresholds that define what it means to narrate oneself in the presence of an algorithmic interlocutor. Here, generative AI becomes both mirror and mediator of human subjectivity, revealing how technology participates in autobiographical meaning-making. This essay enriches the book’s methodological reflection by introducing an intimate and phenomenological perspective on co-creation.

Nadia Olisa and Azaleah Mohd Anis, move into the empirical field, examine the integration of AI tools within cross-cultural qualitative research. Through a comparative case study, the authors show how AI-assisted transcription and analysis can support multilingual, mobile ethnographic methods while raising new questions about voice, translation, and contextual nuance. Their work contributes a critical, practice-oriented dimension to the volume, demonstrating how generative AI can expand methodological possibilities without erasing the interpretive labour at the heart of social inquiry.

Giulia Coppo’s chapter offers a reflexive autoethnographic account of how researchers engage with AI in the early stages of

data collection. By documenting her own process of designing interview guides with the assistance of a large language model, Coppo conceptualises AI as a co-actor within sociotechnical systems of knowledge production. Drawing on theories of co-production (Jasanoff, 2004) and human–machine communication (Guzman & Lewis, 2020), she demonstrates how prompts and outputs create recursive loops that shape both research design and epistemic stance. This chapter deepens the book’s methodological focus by revealing the subtle negotiations of agency, authorship, and reflexivity that accompany AI-assisted research practices.

Alessandra Micalizzi and Caterina Sapone investigate the intersection of generative AI, healthcare, and qualitative research design. Through an exploratory study on the socio-cultural representations of trust in the doctor–patient relationship, the authors employ AI both as object and as instrument of inquiry. By combining visual content analysis, AI-mediated photo-elicitation interviews, and qualitative questionnaires, the chapter reflects on how artificial intelligence participates in the symbolic construction of trust and care. In doing so, it demonstrates the methodological and epistemological value of hybrid approaches that engage AI not only as a data-processing tool but also as a reflexive interlocutor within the research process. This contribution enriches the overall argument of the book by connecting theoretical reflection with empirical innovation, and by highlighting the ethical and affective implications of algorithmic mediation in sensitive human domains such as health.

Elisabetta Risi’s contribution addresses a crucial sociological question: how do AI systems reproduce, and sometimes amplify, the biases of the societies that create them? In “Beyond Bias: Understanding Social Representations Embedded in Generative AI Outputs”, she positions generative AI as both a product and a reflection of social structures. Through a rigorous synthesis of recent studies on cultural bias in large language and text-to-image models, Risi shows how social stereotypes—around gender, ethnicity, class, and geography—are not peripheral distortions

but constitutive elements of algorithmic reasoning. This chapter thus grounds the volume in a broader socio-political reflection, linking the epistemic challenges of AI-assisted research to questions of ethics, equity, and democratic accountability.

The following chapter, written by Anouck Butraud-Assathian, Jaércio da Silva and Cécile Méadel shifts the focus from research and academia to professional discourse. Based on a large corpus of LinkedIn posts and comments, the authors analyse how professionals in the cultural and creative industries narrate and negotiate the role of AI in audiovisual production. By tracing the tensions between enthusiasm, anxiety, and pragmatic adaptation, they reveal the ways in which AI is discursively framed as both a threat and an opportunity. Their work exemplifies the methodological value of digital methods and platform analysis in understanding emerging imaginaries of technological change within creative economies.

Continuing along the ethnographic thread, “Following the Prompt” proposes a new methodological horizon for studying AI: an “AI ethnography.” Gabriella Taddeo situates generative models within complex ecologies of practice where users, systems, and algorithms co-produce meaning. The chapter argues that prompts—the textual interface through which users interact with AI—constitute a crucial ethnographic site, revealing how intentions, imaginaries, and social conventions are translated into computational exchanges. By advocating for an ethnography that observes these micro-interactions, Taddeo restores the ordinary and affective dimensions of algorithmic life, positioning AI ethnography as a necessary evolution of digital methods.

The volume concludes with Agnese Vellar and Matteo Fogli’s forward-looking chapter, which expands the methodological terrain toward futures thinking and speculative design. Here, AI is reframed as a partner in envisioning preferable futures through “creative co-intelligence.” Drawing on design fiction, futures literacy, and AI fluency, the authors propose an anticipatory approach that merges ethnographic observation with imaginative world-building. Their contribution not only synthesises

the book's conceptual threads but also projects them toward the horizon of innovation, ethics, and collective imagination.

Taken together, these eight chapters articulate a plural vision of how AI reshapes research, creativity, and social understanding. The book crosses scales—from the personal to the institutional, from the methodological to the political—trying to open to the complexity the debate about pros and cons of Generative AI in Social Research: the impact of this agentic technology is not confined to automation or efficiency but extends the reflections to possible other uses, according to their synthetic power and depth of statistical analysis.

By weaving together theoretical reflection, empirical inquiry, and methodological experimentation, this volume represents a critical and constructive contribution to the ongoing redefinition of qualitative research in the age of generative intelligence. Its ambition is not merely to document change, but to invite scholars to inhabit it reflexively—to learn, with and through AI, new ways of seeing, questioning, and imagining the social world.

## References

- Bernard, H. R. 2017, *Research Methods in Anthropology: Qualitative and Quantitative Approaches*, (Rowman & Littlefield)
- Hammersley & Atkinson, 2019, *Ethnography: Principles in Practice*, (Routledge)
- Bitkina, O. V., Kim, H. K., & Park, J. (2023). Application of artificial intelligence in medical technologies: A systematic review of main trends. *Artificial Intelligence in Medicine*, 2(2), 101–119.
- Di Franco, G., & Santurro, M. (2021). Machine learning, artificial neural networks and social research. *Quality & Quantity*, 55, 1007–1025.
- Esposito, E. (2022). *Artificial communication: How algorithms produce social intelligence*. MIT Press.
- Jarrahi, M. H. (2025). Interviewing AI: Using qualitative methods to explore and capture machines' characteristics and behaviors. *Big Data & Society*, 12(3), 1–15.
- Kuzior, A., & Kwilinski, A. (2022). Cognitive Technologies and Artificial Intelligence in Social Perception. *Management Systems in*

- Production Engineering, 30(2), 109–115.
- Lagerkvist, A. (2020). Digital Limit Situations: Anticipatory Media Beyond ‘The New AI Era’. *Journal of Digital Social Research*, 2(3), 16–41.
- Ligo, A. K., Rand, K., Bassett, J., Galaitsi, S. E., Trump, B. D., Jayabalasingham, B., Collins, T., & Linkov, I. (2021). Comparing the Emergence of Technical and Social Sciences Research in Artificial Intelligence. *Frontiers in Computer Science*, 3, 653235.
- Lindgren, S., & Holmström, J. (2020). A Social Science Perspective on Artificial Intelligence: Building Blocks for a Research Agenda. *Journal of Digital Social Research*, 2(3), 1–15.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.
- Pop Stefanija, A., & Pierson, J. (2020). Practical AI Transparency: Revealing Datafication and Algorithmic Identities. *Journal of Digital Social Research*, 2(3), 84–125.
- Svensson, J., & Poveda Guillen, O. (2020). What is Data and What Can It Be Used For? Key Questions in the Age of Burgeoning Data-Essentialism. *Journal of Digital Social Research*, 2(3), 65–83.
- Van Maanen, J. (1988). *Tales of the field: On writing ethnography* (2nd ed.). University of Chicago Press.

# **The Role of Generative AI in Qualitative Data Analysis**

## **Opportunities and Limitations in Supporting Dissertation Supervision in Academia**

by Leonard BUSUTIL and Rosienne CAMILLERI

Generative AI is rapidly reshaping dissertation supervision, particularly in qualitative research where large language models (LLMs) are increasingly used for tasks such as coding, summarisation, and thematic exploration. This chapter examines how LLMs can support—but not replace—the interpretive work that underpins rigorous qualitative inquiry. Drawing on Cultural-Historical Activity Theory (CHAT) and connectivism as theoretical framework, we show how technical parameters such as tokenisation, context windows, temperature, and platform guardrails function as methodological variables that directly influence analytical outcomes. We identify key supervisory concerns including students' over-reliance on automated analysis, the erosion of interpretive competencies, documentation and transparency challenges, and inequities in access to advanced AI tools. In response, the chapter advances two guiding commitments: reflexive integration, which positions AI outputs as provisional and subject to triangulation, and digital stewardship, through which supervisors model ethical, transparent, and methodologically coherent AI use. To operationalise these commitments, we introduce the SUPERVISE framework, offering practical strategies for hybrid human–AI workflows, responsible documentation, and equitable supervisory practice. The chapter argues that supervision must shift from transmission to co-navigation, preserving interpretive ownership while leveraging AI's analytical affordances. In doing so, it aligns responsible AI use with the epistemic values and pedagogical aims of qualitative research.

## 1. Introduction

The rapid emergence of large language models (LLMs) such as ChatGPT, Gemini, and Claude has created new possibilities for dissertation supervision in higher education. In qualitative research, these tools are increasingly applied to tasks ranging from initial coding to thematic analysis (Morgan, 2023; Zhang et al., 2023), offering students support in managing extensive data, navigating deadlines, and building analytical confidence. However, efficiency gains must be weighed against potential erosion of interpretative depth, contextual sensitivity, and reflexive practice - the cornerstones of rigorous qualitative analysis (Braun & Clarke, 2020; Friedman et al., 2024).

This chapter examines LLMs as supplementary analytical tools in dissertation supervision, focusing on three critical dimensions. First, we explore how technical parameters such as tokenisation, context windows, temperature settings, and platform guardrails shape analytical outcomes, functioning as method variables rather than neutral background settings (Press et al., 2021). Second, we examine supervisory challenges including student over-reliance, erosion of interpretative competencies, and ambiguity around authorship and accountability (Cook et al., 2025; Goyanes et al., 2025). Third, we address equity concerns: differential access to advanced tools, data privacy risks, and transparency in reporting AI use (Kasperūnienė & Mažeikienė, 2024).

These dimensions intersect within the relational dynamics of dissertation supervision. Drawing on activity theory (Engeström, 1987) and learning ecologies (Siemens, 2004), we argue that LLM technical mechanisms actively mediate students' analytical development. Context window limits influence how data are segmented and interpreted, while temperature settings shape perceptions of reliability, interpretation, and creative inference. Thus, computational affordances produce pedagogical consequences that require thoughtful supervisory intervention.

In response, the chapter advances two commitments for responsible integration: reflexive integration, which positions LLM outputs as provisional and requiring triangulation, and digital stewardship, which frames supervisors as ethical guides who model informed, rigorous AI use. The chapter concludes by introducing the SUPERVISE framework, offering practical strategies for navigating this evolving supervisory landscape. While generative AI also supports quantitative and mixed-methods research, this chapter focuses on qualitative inquiry, where interpretive, contextual, and reflexive dimensions create distinct pedagogical tensions and opportunities.

## **2. Theoretical and Pedagogical Background**

The integration of artificial intelligence into qualitative research methodologies has emerged as one of the most significant developments in contemporary social science research methods. Informed by theoretical foundations that integrate CHAT with the principles of *connectivism* (Siemens, 2004), this section first portrays dissertation supervision in the 21<sup>st</sup> century as part of a socio-technical learning ecology in which knowledge, ethics, and interpretation are distributed across both human and digital nodes. It then turns to a review of the literature on dissertation supervision in higher education, situating supervisory practices within ongoing debates about research integrity and academic development. Next, the technical underpinnings of LLMs are examined, outlining how these systems process and generate data, which in turn affects their analytical capabilities. Finally, the review synthesises empirical studies published between 2023 and 2025 that investigate the application of LLMs in qualitative data analysis. The goal is to critically evaluate the current state of knowledge, identifying convergent findings, methodological variations, and persistent challenges in this emerging field.

## 2.1 Theoretical Framework: CHAT and the Principles of Connectivism

To examine how generative AI reshapes dissertation supervision, this chapter adopts a hybrid theoretical framework that integrates Cultural-Historical Activity Theory (CHAT) (Engeström, 1987; Engeström & Sannino, 2010) with key principles from Siemens' (2004) theory of connectivism. This combined lens enables an exploration of how the teaching and learning processes embedded within the supervisory relationship are transformed when generative AI functions as a mediating artefact. Through this framework, the chapter examines the evolving interactions, connections, and experiential dimensions that shape the supervisor–supervisee dynamic in an AI-mediated academic context.

Viewed through the activity-theoretical lens, dissertation supervision emerges as an activity system comprising subjects (supervisors and supervisees), tools (including generative AI), and the academic community as its sociocultural context (Engeström, 1987). CHAT provides a structural framework for interrogating how learning, development, and scholarly practice are mediated by technological tools within institutional and cultural settings. Central to this perspective is an attention to contradictions, tensions, and transformations that arise within the system — for instance, between traditional notions of intellectual work and the automation of analytic processes enabled by AI. In this sense, supervision is not merely a pedagogical transaction but a dynamic, culturally embedded practice, continually negotiated through the evolving rules, divisions of labour, and ethical boundaries. The CHAT framework sheds light on how generative AI mediates analytical and interpretive practices, functioning not as an autonomous authority but as a collaborative cognitive agent that reconfigures supervisory relationships, redistributes cognitive labour, and rearticulates the parameters of methodological rigour.

Building on this, connectivism (Siemens, 2004) extends CHAT's analytic scope and reframes it within a broader ecology of distributed cognition and networked learning. As a learning

theory attuned to the digital age, connectivism foregrounds the flow of knowledge across complex human-machine-institutional networks, constantly mediated by technological advancements. From this perspective, generative AI operates as a *node* within a larger knowledge network that learners must navigate critically and reflexively. Within such a positioning, connectivism emphasises adaptability and knowledge flow across networks, where supervisors facilitate students' metacognitive awareness of how these tools shape their learning trajectories, knowledge production, and scholarly values.

This integrated theoretical framework thus accentuates generative AI functions as both a mediating artefact and as a knowledge node within the learning network—a dynamic participant in the network through which knowledge is constructed, shared, and evaluated. Supervisors and students co-navigate this network, interpreting AI outputs, negotiating reliability, and critically engaging with algorithmic representations of data. One can easily deduce that the interactions and connections that occur in the flow among the supervisor, supervisee, and technological affordances transform supervision from a dyadic exchange into a distributed pedagogical system in which knowledge production and reproduction, methodological reflection, and ethical reasoning unfold across both human and digital agents. The fusion of CHAT and connectivism, therefore, positions supervision as a site of reflexive networked practice where supervisors become mediators of digital epistemologies, teaching students not only how to use AI tools methodologically but also how to think critically about their roles in shaping knowledge, interpreting results, and demonstrating academic integrity.

## *2.2 Dissertation Supervision in Higher Education*

Dissertation supervision plays a central role in cultivating students' commitment to research integrity and responsible scholarship. Supervisors model disciplinary norms and ethical reasoning through guidance, feedback, and relational practices (Bird,

2001; Krásničan et al., 2024). These expectations encompass critical engagement, authorship practices, and collaborative inquiry (Lofström & Pyhältö, 2020). Foundational research emphasises shared understandings, transparent expectations, and trust as essential to effective supervision (Pizzolato, 2022; Pyhältö et al., 2015). Lee's (2008) multidimensional model remains influential, describing supervisory work across functional, enculturation, critical thinking, emancipation, and relational domains. The emergence of generative AI introduces new complexities. While AI can support writing clarity, organisational structure, and preliminary feedback, allowing supervisors to focus on conceptual and methodological development (Rafi & Amjad, 2025), it also raises concerns about authorship, hallucination, and concealed reliance on automated outputs. As a result, institutions are developing new integrity frameworks (Bjelobaba et al., 2024; Eacersall et al., 2024), emphasising accountability, transparency, and prohibition of AI authorship (Gulumbe et al., 2025). Supervisors must now address not only long-standing ethical risks but also AI-specific challenges such as dataset bias, overreliance, and misplaced trust in authoritative-sounding outputs (Gao et al., 2025).

The integration of AI into dissertation supervision must be situated within broader developments concerning AI literacy and institutional readiness in higher education. Recent research reveals significant variation in both institutional policies and faculty preparedness for AI integration (Southworth et al., 2023; Chan & Hu, 2023). Whilst some universities have developed comprehensive frameworks including training programs and ethical guidelines, others have adopted reactive, prohibition-focused approaches that leave supervisors uncertain about appropriate engagement. Studies of faculty attitudes reveal tensions between recognising pedagogical potential and concerns about academic integrity, workload implications, and adequacy of institutional support (Boud & Brew, 2013). Particularly relevant is research demonstrating that effective supervisor development for technology-mediated pedagogy requires not only technical training but also opportunities for critical reflection on how digital tools

reshape supervisory relationships and research practices (Halse & Malfroy, 2010).

From a CHAT perspective, these institutional frameworks constitute essential elements of the activity system, providing rules, division of labour, and community structures that enable or constrain supervisors' capacity to integrate AI tools effectively. This literature highlights how responsible AI integration cannot rely solely on individual supervisor initiative but requires coordinated institutional investment in professional development and policy frameworks.

### *2.3 How LLMs Process Data*

This section considers the properties of LLMs most likely to affect qualitative analysis. We highlight how these features can support analytic consistency while also posing limits that demand reflexive oversight.

#### *2.3.1 Tokenisation: The Foundation of Text Processing*

LLMs process text by breaking it into tokens, the smallest computational units, which may represent words, fragments, or characters. Tokenisation affects qualitative analysis as semantic consistency may be compromised when similar concepts are tokenised differently, particularly problematic for multilingual data where models trained primarily on English struggle (Qiu et al., 2020; Rust et al., 2021). However, models process tokens contextually, maintaining semantic relationships during analysis.

#### *2.3.2 Context Windows: The Scope of Understanding*

An LLM's context window sets the maximum text it can process at once, ranging from 4,000 to over a million tokens (Press et al., 2021). For qualitative researchers, this creates constraints on document length and continuity, as lengthy transcripts may require segmentation that risks fragmenting thematic links.

The context window limitation also affects relationship mapping, as the model can only identify relationships and patterns

within its processing scope. Cross-document or long-range thematic connections may be missed if they exceed this limit, potentially fragmenting the analytical process and conflicting with qualitative research's emphasis on understanding phenomena in their full context (Vaswani et al., 2017). When analysing multiple interviews or documents sequentially, the model cannot retain insights from previous analyses unless they are explicitly included in the current context window, necessitating careful consideration of how to maintain analytical continuity across extended datasets.

### 2.3.3 Temperature: Controlling Analytical Creativity

The temperature parameter controls randomness in model responses (typically 0-1), balancing consistency with interpretive richness (Holtzman et al., 2019). Low settings (0-0.3) produce deterministic, consistent outputs suitable for systematic coding but may miss nuanced interpretations. High settings (0.7-1.0) generate varied, creative responses identifying unexpected patterns but reduce consistency across similar tasks. Different analytical phases benefit from different settings: systematic coding requires lower temperatures for reliability, while thematic exploration may benefit from higher settings encouraging novel pattern identification

### 2.3.4 Prompt Engineering: Shaping Analytical Perspective

The way researchers phrase requests fundamentally shapes LLM analytical approaches, making prompt engineering critical (Reynolds & McDonell, 2021). Effective prompting involves several key considerations. Role definition instructs the model to adopt specific analytical perspectives—phenomenological analysis versus behavioural pattern identification yields fundamentally different insights. Methodological framing specifies the qualitative approach employed (thematic analysis, grounded theory, interpretative phenomenological analysis), as each tradition carries distinct analytical procedures (Smith et al., 2022). Output structure defines presentation format - themes with quo-

tations, coded segments with relationships, or interpretive narratives. Contextual grounding provides relevant background about research questions, populations, or theoretical frameworks. Poor prompting risks superficial results; well-structured prompts enhance analytical depth and methodological coherence.

### 2.3.5 Guardrails: Boundaries and Limitations

LLMs incorporate safety measures and content policies affecting qualitative analysis of sensitive topics (Bai et al., 2022). Content filtering may refuse to analyse sensitive subjects like trauma, violence, or discrimination - important areas of qualitative inquiry. Bias mitigation attempts may overcorrect, affecting analysis of texts discussing sensitive demographic issues requiring nuanced understanding. Factuality guardrails can reduce interpretive flexibility (Weidinger et al., 2021). Researchers must understand how these guardrails shape analytical possibilities.

### 2.3.6 Implications and Common Misconceptions

LLMs provide efficiency, pattern recognition, and consistency (Gilardi et al., 2023; Törnberg, 2023), but lack experiential grounding, cultural embeddedness, and theoretical reasoning (Bender et al., 2021). Outputs remain sensitive to prompt wording and system parameters (Ouyang et al., 2022). Misconceptions, such as assuming comprehension, neutrality, or methodological independence, risk undermining human interpretive authority (Shanahan, 2022; Blodgett et al., 2020). Supervisors therefore play a critical role in ensuring students treat LLMs as mediating tools rather than epistemic replacements, consistent with CHAT and connectivist principles.

## 2.4 *Application of Large Language Models (LLMs) in Qualitative Data Analysis*

### 2.4.1 Methodological Landscapes and Research Design

Recent studies focus predominantly on exploratory and methodological designs, reflecting the field's nascent nature (Hamil-

ton et al., 2023; Gao et al., 2023). A recurring strategy involves comparative analysis, systematically comparing AI-assisted outputs with traditional manual coding (Li et al., 2024; Prescott et al., 2023), though evaluation criteria variation complicates synthesis. Three trajectories emerge: proof-of-concept studies demonstrating basic capabilities, validation studies assessing reliability against human coding, and integration studies exploring optimal human-AI workflows. This evolution suggests movement from feasibility testing toward practical implementation frameworks.

#### 2.4.2 Generative AI Technologies and Tool Selection

The technological landscape remains dominated by OpenAI models, with studies using GPT-3.5 and GPT-4 far outnumbering those employing alternatives (Theelen et al., 2024; Zhang et al., 2025). This reliance appears to stem more from accessibility and usability than from systematic tool selection. A minority of studies explored Llama, Bard (Gemini), or customised implementations (Bijker et al., 2024).

Most research employed publicly available models in their default settings, with relatively few attempts at fine-tuning or domain-specific adaptations. Notably, customised models such as *QualiGPT* reported stronger results (Zhang et al., 2023), suggesting that domain-optimised systems may hold particular promise for qualitative research. Overall, the choice of generative AI technology appears largely opportunistic rather than theoretically driven, with researchers typically selecting the most accessible or familiar tools (Siiman et al., 2023).

#### 2.4.3 Qualitative Analytic Stages and LLM Applications

Open coding represents the most straightforward LLM application, with studies reporting accuracy comparable to human coders when given sufficient context (Hamilton et al., 2023; Theelen et al., 2024). Performance weakens in axial and selective coding requiring conceptual synthesis and theoretical integration (De Paoli, 2023), where generative AI's limited interpretive depth necessitates significant human refinement.

Despite these challenges, generative AI demonstrates competency identifying manifest themes and clustering large datasets (Goyanes et al., 2025; Dai et al., 2023) but struggles generating latent or interpretive themes requiring theoretical alignment (Friedman et al., 2024; Schroeder et al., 2024). In these cases, outcomes depend heavily on prompt design (Jalali & Akhavan, 2024; Yue et al., 2025).

Correspondingly, inter-rater reliability varies widely with Theelen et al. (2024) reporting 34-70% agreement depending on contextual information quality. LLMs achieve reasonable reliability when carefully configured but outputs require context-specific validation to ensure methodological integrity (Tai et al., 2024; Zhang et al., 2025).

#### 2.4.4 Comparative Performance Analysis

Across studies, several benefits recur. First, LLMs provide considerable efficiency, reducing the time required for coding and organisation (Morgan, 2023; Parker et al., 2023; Xiao et al., 2023). Second, they deliver greater consistency than human coders, unaffected by fatigue or drift (Liu & Sun, 2023; Than et al., 2025). Third, they enhance comprehensiveness, often identifying patterns that humans overlook (Cook et al., 2025; Acheampong & Nyaaba, 2024).

However, serious challenges remain. LLMs struggle with cultural nuance, implicit meaning, and situated knowledge (Levit & Saban, 2025; Kasperūnienė & Mažeikienė, 2024). They also have difficulty with theoretical integration, failing to connect findings to broader conceptual frameworks (Bhaduri et al., 2024; Gamiel-dien et al., 2023). Ethical risks, including bias, privacy concerns, and the perpetuation of structural inequities, remain underexplored but are increasingly acknowledged (Friedman et al., 2024; Cook et al., 2025).

A consistent finding across the literature is the importance of careful prompt design. Studies that included iterative refinement, explicit theoretical framing, and contextual detail reported significantly stronger results (Reynolds & McDonnell, 2021). The

most promising research employs *hybrid workflows*, in which AI supports coding or theme generation, while interpretive oversight remains with the human researcher (Schroeder et al., 2024; Cook et al., 2025).

#### 2.4.5 Synthesis and Future Directions

The current body of research establishes that LLMs possess demonstrable utility for specific qualitative research tasks, particularly initial coding and pattern identification, whilst facing significant limitations in theoretical interpretation and contextual analysis. The field is evolving from initial feasibility assessment towards more sophisticated implementation frameworks that recognise both AI capabilities and constraints. The studies collectively suggest that integrating LLMs into qualitative research requires a fundamental reconsideration of traditional methodological approaches. Rather than simple tool substitution, successful implementation requires new methodological frameworks that optimally combine human expertise with AI capabilities. This integration challenges conventional notions of researcher agency and interpretive authority, requiring careful consideration of epistemological implications.

Future research directions should prioritise the development of evaluation frameworks for AI-assisted qualitative analysis, the exploration of domain-specific model training approaches, and an investigation of the ethical implications for research validity and participant privacy. The field would benefit from standardised assessment criteria and best practice guidelines to support rigorous implementation of these emerging technologies.

This review demonstrates how supervision, methodological rigour, and the technical workings of LLMs intersect in shaping the role of AI in qualitative research. Building on these insights, the following discussion explores broader implications, particularly the pedagogical and ethical challenges of supervising dissertation research in an era where human and machine collaboration is increasingly unavoidable.

### 3. Discussion

The preceding sections have established the technical mechanisms through which LLMs process qualitative data (Section 2.3) and reviewed their applications across diverse research contexts (Section 2.4). This discussion synthesises these foundations to examine their pedagogical implications for dissertation supervision, guided by the Cultural-Historical Activity Theory and connectivism framework established in Section 2.1.

Through a CHAT lens, we explore how LLMs function as mediating artefacts that reshape the supervisory activity system, introducing contradictions between traditional research practices and emergent technological affordances. Connectivism helps us understand how supervision must adapt when knowledge generation becomes distributed across human and computational nodes. Together, these lenses reveal supervision not as transmission of static methodologies but as co-navigation of evolving socio-technical learning ecologies.

The following five subsections explore pedagogical shifts required (3.1), methodological tensions and hybrid workflows (3.2), ethical dilemmas and integrity safeguards (3.3), responsibilities in uneven digital landscapes (3.4), and principles of reflexive integration as pedagogical orientation (3.5).

#### *3.1 From Supervision to Mediation: Pedagogical Shifts in the Age of AI*

Generative AI challenges traditional supervision models that support the gradual development of methodological rigour and interpretive independence (Lee, 2008). The tension is stark: AI's speed and ease can encourage superficial engagement, eroding the deep, iterative analysis that defines qualitative research - yet it also offers genuine scaffolding for inexperienced researchers managing large datasets or facing linguistic barriers (Braun & Clarke, 2020; Friedman et al., 2024; Rafi & Amjad, 2025).

Supervision must evolve from merely transmitting established methods to negotiating AI's dual nature as both a scaffold

and a shortcut. This pedagogical shift requires three interconnected practices:

1. *Modelling Critical Interrogation*: Supervisors must demonstrate how to treat AI outputs as provisional starting points rather than conclusive findings. This approach involves publicly interrogating reliability, situating outputs within theoretical frameworks, and triangulating against manual coding. Rather than simply instructing students about AI limitations, supervisors model sceptical engagement in real time.
2. *Fostering Transparency as Practice*: Documentation becomes pedagogically central, not as a bureaucratic exercise but as a reflexive habit. Students learn to specify the models used, the parameters set, the prompts refined, and how discrepancies between human and AI analyses were resolved (Eacersall et al., 2024; Gao et al., 2025). Transparency itself becomes the subject of supervisory teaching.
3. *Creating Safe Dialogic Spaces*: Supervisors must cultivate environments where students can voice uncertainties, reveal misuses, and discuss over-reliance without censure. Integrating AI as routine supervisory conversation rather than avoiding it or treating it punitively, signals that critical engagement with emergent technologies is integral to becoming a conscientious researcher.

From a CHAT perspective, this pedagogical shift represents a fundamental transformation of the supervisory activity system. Where supervisors previously mediated between students and traditional analytical methods, they now navigate a triangular relationship among student (subject), qualitative data (object), and LLM (mediating artefact). This approach introduces contradictions: institutional rules often lag behind technological capabilities, whilst community norms about what constitutes “authentic” student work remain contested. Supervisors must mediate these contradictions, helping students develop agency within socio-technical learning networks (Siemens, 2004) rather than simply mastering predetermined procedures.

The supervisory challenge is ensuring that AI functions as a

complement rather than a substitute. For students lacking methodological confidence, AI can reduce cognitive load during basic coding, creating space for deeper interpretative work under guidance. However, without active supervision, these scaffolds readily become crutches that prevent skill development. The shift required is substantial: supervisors become not just mentors of research integrity but also guides for responsible digital scholarship, preparing students to preserve interpretive ownership while leveraging technological efficiency.

### *3.2 Methodological Tension and Possibilities in AI-Supported Analysis*

Integrating LLMs into qualitative research creates methodological tensions requiring new supervisory responses. These tools efficiently manage complex datasets, scaffolding students' early analytical engagement (Gilardi et al., 2023; Törnberg, 2023). Yet efficiency risks encouraging superficial habits that bypass the iterative, reflexive engagement at the core of qualitative research.

Supervisors must therefore guide students to understand LLMs not as impartial observers but as systems shaped by tokenisation, context windows, prompting structures, and probabilistic inference (Press et al., 2021; Reynolds & McDonell, 2021). When context limits truncate transcripts or different prompts yield divergent themes, these become teachable moments for methodological reflexivity. Understanding technical constraints is essential for ethical, effective AI use.

LLMs demonstrate competence in descriptive coding and surface-level pattern identification but falter at conceptual work: axial and selective coding, theoretical integration, and contextual interpretation (Theelen et al., 2024; De Paoli, 2023). This capability gap clarifies the supervisor's role: ensuring students treat AI-generated categories as provisional, interrogating them against data and theoretical frameworks. Without this guidance, students may mistake computational pattern recognition for qualitative analysis.

The solution lies not in rejection but in hybrid approaches

where LLMs handle preliminary organisation while human researchers retain interpretive authority. Effective workflows involve:

- AI-assisted initial coding → manual verification and refinement
- Machine-generated themes → theoretical triangulation and contextualisation
- Pattern identification → interpretive synthesis and meaning-making

Such workflows preserve methodological rigour while leveraging computational efficiency (Cook et al., 2025; Schroeder et al., 2024). Students learn to position AI outputs as one analytical perspective requiring critical interrogation not authoritative conclusions.

These hybrid workflows embody the connectivist principle that learning occurs through network connections rather than internalised content alone (Siemens, 2004). When students learn to orchestrate AI-assisted initial coding followed by human refinement, they develop the capacity to navigate distributed knowledge systems - a core twenty-first-century competency. However, CHAT reveals potential contradictions: rules requiring “independent” dissertation work may conflict with the inherently collaborative nature of human–AI workflows. Supervisors must therefore help students document and theorise these partnerships transparently, positioning AI use not as delegation but as informed tool selection within broader methodological repertoires.

Supervisors thus extend their role beyond transmitting established practices to cultivating dual literacy: traditional qualitative skills and the critical capacity to navigate the affordances and constraints of digital tools. Where LLMs offer breadth, supervisors help students leverage advantages; where they threaten reflexivity, supervisors intervene to preserve qualitative inquiry’s integrity.

### *3.3 Safeguarding Integrity: Ethical and Epistemological Dilemmas*

The integration of LLMs into qualitative research presents both opportunities and dilemmas that cut to the heart of research integrity and interpretive authority. These tools are most effective when thoughtfully integrated into human-led processes rather than positioned as substitutes for human judgment. At their best, they can act as analytical assistants, identifying preliminary patterns, suggesting coding schemes, or flagging passages for closer human review while maintaining researcher oversight of interpretive decisions. As consistency checkers, they can apply coding criteria systematically across large datasets, though responsibility for establishing and validating such criteria must remain with human researchers. LLMs can also serve as alternative perspective generators, offering interpretive prompts; however, such suggestions must always be weighed against theoretical frameworks and contextual knowledge that can only be provided by human expertise.

For supervisors, this hybridisation of analysis creates new pedagogical challenges. Students must learn not only how to apply traditional qualitative procedures but also how to critically integrate AI outputs without displacing their own interpretive agency. Supervisors are therefore called to model reflexive practices - treating AI outputs as provisional, interrogating their reliability, and foregrounding the importance of theoretical framing. In guiding students through these practices, supervisors safeguard against a dilution of interpretive depth and help ensure that the central tenets of qualitative inquiry, including reflexivity, contextual sensitivity, and theoretical integration, remain intact.

Documentation has become a cornerstone of integrity in this shifting terrain. Beyond conventional reporting, students must now specify which model was used, under what parameters, how prompts were designed, and how outputs shaped the analytic process (Liang et al., 2022). Rigorous accounts should describe how AI outputs were integrated with human analysis, how conflicts were resolved, and how iterative refinements un-

folded. Validation must also include checks such as inter-rater reliability between human and AI coding, member validation, and theoretical testing of emergent themes. Supervisors play a key role in framing such documentation not as bureaucratic formality but as an ethical obligation central to transparency and academic accountability.

Equally pressing are the ethical questions raised by AI integration. Data privacy becomes a significant concern when sensitive information is shared with commercial services, raising questions of consent and confidentiality. Interpretive authority also becomes complex: in collaborative human-AI analyses, accountability for meaning-making must remain with the human researcher. Transparency with participants is an emerging ethical requirement, as individuals may reasonably expect to know whether AI tools have mediated the interpretation of their contributions. Supervisors must therefore guide students to scrutinise institutional guidelines, employ anonymisation strategies, and adopt safer infrastructures where necessary.

These ethical challenges can be understood through CHAT's attention to divisions of labour within activity systems. When LLMs perform coding tasks, the division of interpretive labour shifts: pattern recognition migrates to computational nodes whilst meaning-making, contextualisation, and theoretical integration remain with human researchers. This redistribution is productive only when made explicit and when power dynamics favour human interpretive authority. Students must learn to position themselves as orchestrators of socio-technical knowledge networks (Siemens, 2004) rather than consumers of AI-generated insights.

These ethical and epistemological dilemmas can also be read through Lee's (2008) supervisory dimensions, which acquire renewed significance in the context of AI-assisted inquiry. The *functional* dimension now entails not only oversight of methodological processes but also attention to ensuring that students follow institutional and technical safeguards when deploying LLMs. The *critical thinking* dimension is extended as supervisors

must teach students to interrogate AI outputs with scepticism, recognising them as prompts for interpretation rather than definitive analyses. The *emancipatory* dimension becomes more complex, for while supervisors are tasked with fostering student independence, they must also ensure that such independence is not compromised by overreliance on automated systems. Finally, the relational dimension of supervision gains new significance as supervisors model ethical reflexivity, openly discussing their own practices and uncertainties surrounding AI use, thereby demystifying its role and reinforcing a culture of transparency and accountability.

Viewed this way, AI does not diminish the supervisory role but deepens it. Supervisors become not only mentors of methodological rigour but also custodians of ethical reflexivity, charged with preparing researchers to navigate new dilemmas around privacy, interpretive responsibility, and transparency. Far from displacing long-standing concerns such as plagiarism or authorship disputes, AI introduces additional layers of vigilance that call for supervisory guidance. In this evolving landscape, supervisors must mediate the interplay of human judgment and digital assistance, safeguarding both the quality and the integrity of qualitative research.

### *3.4 Supervisory Responsibilities in an Uneven Digital Landscape*

Asymmetries of knowledge, power, and access have always shaped supervision in higher education. However, the arrival of generative AI introduces uneven terrains that supervisors must now learn to traverse with their students. Access to advanced LLMs remains stratified, with some institutions offering subscriptions to models such as GPT-4 while others restrict students to basic versions, creating disparities in research capacity and methodological experimentation (Siiman et al., 2023; Zhang et al., 2023). Supervisors, already cast as mentors of research integrity and scholarly practice (Bird, 2001; Krásničan et al., 2024), are increasingly called to mediate these inequities, ensuring that dif-

ferences in access do not translate into reduced opportunities for rigour, depth, or creativity in student research.

The unevenness is also a cultural and disciplinary issue. While applied social sciences are beginning to embrace AI as a methodological scaffold (Rafi & Amjad, 2025), many humanities disciplines remain cautious, namely due to the perceived risks of superficiality, displaced authorship, and compromised integrity (Bjelobaba et al., 2024; Eacersall et al., 2024). Supervisors stand at the intersection of these debates, negotiating between disciplinary traditions and emerging technological possibilities. In doing so, they extend Lee's (2008) dimensions of supervision: the functional and enculturation tasks now involve inducting students into ongoing debates about digital tools, disciplinary ethics, and shifting methodological norms.

Compounding these dynamics is a shift in expertise hierarchies. Supervisors are traditionally expected to be knowledge authorities, but when it comes to rapidly evolving AI tools, students often arrive with greater familiarity and fluency than their mentors. Rather than undermining supervisory authority, this inversion opens possibilities for reimagining the supervisory relationship. It calls supervisors to model humility, adaptability, and co-learning, embodying Lee's (2008) relational and emancipatory dimensions in new ways. By embarking on shared exploration of unfamiliar digital terrains, supervisors encourage scholarly resilience and critical reflexivity. These skills extend beyond any particular technology and endeavour to prepare students for lifelong engagement with evolving research environments.

Institutional policies further complicate this landscape. Universities differ widely in their approaches, with some tightly regulating the use of generative AI tools because they are regarded as a threat to academic integrity, while others embed it within broader digital literacy agendas (Gulumbe et al., 2025; Gao et al., 2025). Supervisors find themselves tasked with interpreting and enacting these institutional framings, balancing the demand for compliance with their pedagogical responsibility to foster intellectual autonomy and methodological creativity. This balancing

act exemplifies what Braun and Clarke (2020) describe as the cultivation of reflexive practice: supervisors must help students to work within constraints while still retaining space for interpretive depth and ethical responsibility.

From a CHAT perspective, institutional frameworks constitute essential rules and community structures within the supervisory activity system. Uneven institutional investment in AI literacy training creates contradictions: supervisors face expectations to guide AI-literate students whilst lacking access to professional development resources. This situation reveals how macro-level contradictions (institutional policy) cascade into micro-level tensions (individual supervisory relationships). Connectivism suggests that supervisors might form professional learning networks to share strategies and resources, partially compensating for institutional gaps through distributed, emergent expertise.

Against this backdrop, supervision cannot be understood solely as the transmission of established research practices. Instead, it must be reconceptualised as a form of digital stewardship, where supervisors act as *reflexive co-navigators* alongside their students. Digital stewardship entails three interlocking responsibilities: first, to ensure equity of access and opportunity across uneven technological landscapes; second, to model and cultivate reflexive engagement with AI tools, treating them as provisional aids rather than authoritative voices; and third, to sustain the integrity of qualitative inquiry by insisting that human judgment, contextual sensitivity, and theoretical imagination remain central. By adopting this stance, supervisors safeguard disciplinary traditions while simultaneously preparing students to thrive in a research culture characterised by constant technological flux.

In this way, supervisory responsibility is not diminished but transformed. Supervisors become not only guardians of research integrity but also partners in navigating the uncertainties of an evolving digital era, equipping emerging scholars with the critical dispositions, ethical reflexivity, and adaptive capacities needed to flourish in uneven and unsettled research landscapes.

### 3.5 Reflexive Integration as Pedagogical Orientation

The literature converges on a clear conclusion: LLMs prove most useful when woven into human-led processes rather than substituting for interpretation (Gilardi et al., 2023; Törnberg, 2023; Theelen et al., 2024). Their affordances—preliminary coding, pattern identification, and consistency across large datasets—complement but cannot replace the conceptual depth, theoretical integration, and interpretive subtlety that define qualitative inquiry (De Paoli, 2023; Bender et al., 2021).

Reflexive integration addresses this reality through three interlocking practices:

1. *Interrogative Stance*: Students learn to position AI outputs as tentative and partial, asking: How do AI-generated codes align with or diverge from manual coding? How do thematic suggestions fit within theoretical frameworks? Where must human interpretive authority override computational pattern recognition? (Braun & Clarke, 2020; Shanahan, 2022). AI becomes the site where reflexive habits are taught, not where they are abandoned.
2. *Enhanced Documentation*: Transparent accounts must specify technical dimensions, including model version, parameters, prompts, and how human-AI discrepancies were resolved (Liang et al., 2022; Ouyang et al., 2022). Supervisors frame this not as bureaucracy but as an ethical responsibility to participants and scholarly communities. Documentation itself becomes a reflexive practice, demonstrating how methodological decisions, including AI integration, shaped analytical outcomes.
3. *Ethical accountability*: When insights emerge through human-AI collaboration, researchers remain accountable for interpretations and participant representation (Gao et al., 2025). Data privacy, interpretive authority, and participant transparency require heightened vigilance. Supervisors guide students to recognise AI not as neutral but as a contingent tool demanding contextualisation and critique.

Reflexive integration thus represents a pedagogical orientation rather than merely a technical adjustment. It preserves hu-

man judgment, theoretical sensitivity, and ethical responsibility at the centre of qualitative inquiry while leveraging computational efficiency. This stance resists both uncritical enthusiasm and categorical rejection, positioning AI as one provisional perspective among many, always questioned, always situated within human researchers' interpretive and ethical commitments.

### *3.6 Institutional and Professional Development Realities*

The pedagogical responsibilities outlined above presume supervisory capacity that cannot be assumed without deliberate institutional investment and professional development frameworks. Several critical questions arise that warrant explicit attention: Who trains supervisors in the technical and pedagogical dimensions of AI integration? What institutional resources and policies enable or constrain this transformation? How do workload pressures, disciplinary cultures, and career stage mediate supervisors' capacity to engage reflexively with these issues?

Drawing on CHAT's attention to contradictions within activity systems (Engeström, 1987), we can identify tensions between institutional expectations for AI-literate supervision and the often-limited provision of training, time, and resources required to develop such capacity. Research on academic development suggests that effective professional learning requires sustained engagement rather than one-off training sessions (Roxå & Mårtensson, 2015). Supervisors need opportunities to explore AI tools practically, discuss ethical dilemmas collegially, and develop discipline-specific approaches reflecting their fields' epistemological commitments. However, institutional investment in such development remains uneven, with some universities offering comprehensive support whilst others expect supervisors to navigate these changes independently (Chan & Hu, 2023).

Equity concerns extend beyond student access to encompass supervisor access to advanced tools, training, and time for experimentation. Junior faculty may face particular tensions between adopting innovative approaches and meeting traditional

research productivity expectations. Disciplinary variation compounds these challenges: whilst applied social sciences increasingly embrace computational methods, humanities disciplines often maintain greater scepticism toward AI integration, creating potential inconsistencies in institutional approaches and leaving individual supervisors to negotiate these tensions without clear guidance.

Moreover, the rapid pace of AI development means that any particular technical training quickly becomes outdated. Sustainable approaches require cultivating dispositions of critical inquiry and adaptive practice rather than mastering specific tools. From a connectivist perspective (Siemens, 2004), this emphasises the importance of network-building among supervisors, creating communities of practice where experiences, challenges, and strategies can be shared and refined collectively. Institutional frameworks must therefore balance providing practical guidance with fostering intellectual community around evolving questions rather than prescriptive solutions.

#### **4. Practical Strategies for Dissertation Supervision**

In this uneven and rapidly evolving digital landscape, supervisors are called to reimagine their roles not only as mentors of research integrity but also as *digital stewards*. This shift in roles involves guiding students towards AI use that is supplementary, intentional, and critically reflective. LLMs can support research by improving coding efficiency, speed, scalability, and consistency, as well as by detecting patterns in large datasets (Gilardi et al., 2023; Törnberg, 2023). However, these potential strengths cannot overshadow their significant limitations: limited contextual sensitivity, difficulty in axial and selective coding, weak theoretical integration, and ethical concerns, including bias, hallucination, and compromised data privacy (Theelen et al., 2024; De Paoli, 2023; Bender et al., 2021).

Supervision, therefore, becomes a balancing act, supporting

students in harnessing these affordances while ensuring that the interpretive depth, reflexivity, and ethical responsibility that define qualitative research are not dismissed. This act requires modelling transparency, encouraging critical interrogation, and framing AI outputs as provisional prompts rather than authoritative analyses.

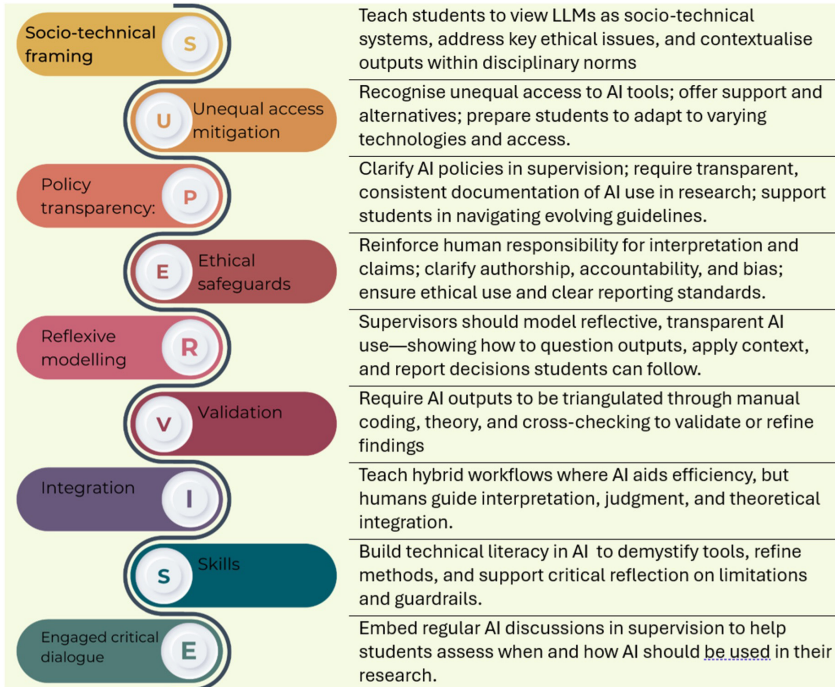


Fig. 1 The Supervise Framework

Figure 1 presents the SUPERVISE framework, a set of practical strategies designed to support supervisors in navigating their evolving roles within AI-mediated dissertation work. The acronym provides a clear, accessible reference point, enabling supervisors to systematically reflect on and assess the shifting dimensions of their responsibilities.

The SUPERVISE framework operationalises the theoretical commitments established in Section 2.1, translating CHAT and

connectivist principles into actionable supervisory practices. Drawing on CHAT's conceptualisation of tools as mediating artefacts within activity systems (Engeström, 1987), the framework positions supervisors as strategic mediators who help students navigate contradictions between institutional expectations, technological affordances, and methodological traditions. From a connectivist perspective (Siemens, 2004), effective supervision cultivates students' capacity to build and maintain productive connections within distributed socio-technical learning networks.

Each element of the SUPERVISE framework addresses specific dimensions of the supervisory activity system: subject (student autonomy and skill development), tools (AI technologies and their appropriate use), rules (ethical guidelines and documentation standards), community (professional networks and peer learning), and division of labour (clarity about human versus computational responsibilities). Together, these elements create coherent supervisory practice grounded in both theoretical understanding and pragmatic responsiveness to the realities supervisors face.

Practical strategies for supervisors include:

- *Modelling reflective and transparent AI use*, demonstrating openly how outputs can be interrogated, contextualised, and reported.
- *Encouraging triangulation and validation* of AI outputs through manual coding, theoretical framing, and cross-checking.
- *Developing technical literacy of LLMs* for both supervisors and students alike, encouraging them to cultivate a working understanding of how LLMs process language—tokenisation, context limits, probabilistic prediction, temperature settings, and guardrails. This technical literacy would form part of focused training to demystify AI tools and equip researchers to anticipate their limitations, ask better methodological questions, and design prompts more purposefully.
- *Embedding critical conversations about AI* within supervisory dialogue, enabling students to voice both its opportunities and risks.
- *Protecting ethical standards*, particularly the safeguarding of

data privacy, clarity of interpretive authority, and transparency in documenting AI use in dissertations.

- *Addressing equity of access*, acknowledging that not all students have equal exposure to advanced AI tools, and mitigating disparities where possible.
- *Expanding methodological training* to include prompt design, parameter awareness, and critical reflection on guardrails and constraints.
- *Teaching hybrid workflows*, where the efficiency of AI is harnessed but always complemented by human reflexivity and theoretical integration.
- *Build critical digital literacy* into supervision, ensuring that students engage critically with issues of privacy, authorship, accountability, and bias, and that they see LLMs as socio-technical systems rather than neutral tools.
- *Preparing for uneven digital access*, equipping students to adapt to shifting AI institutional policies and technological variation.

Taken together, these strategies position AI not as a threat to academic integrity but as a catalyst for pedagogical reflection and methodological renewal.

## 5. Conclusion

Generative AI fundamentally challenges dissertation supervision, requiring neither dismissal nor uncritical embrace but a recalibrated supervisory stance. This chapter has advanced two interconnected commitments: *Reflexive Integration* positions AI as a provisional aid that requires interrogation, triangulation, and situating within theoretical and ethical frameworks (Braun & Clarke, 2020; Shanahan, 2022). Students learn to preserve interpretive authority while leveraging computational efficiency, treating LLM outputs as accountable to the epistemic commitments of qualitative inquiry. *Digital Stewardship* frames supervisors as ethical guides who mitigate uneven access, navigate institutional frameworks, and model transparent AI engagement (Bird, 2001; Lee, 2008; Gao et al., 2025). This approach extends supervision beyond transmitting established norms to mediating students' relationships with technological tools. Together-

er, these commitments transform supervision from knowledge transmission to co-navigation. Supervisors and students explore unsettled digital terrain together, preserving qualitative inquiry's interpretive heart while adapting to technological realities. The SUPERVISE framework operationalises this transformation, offering practical strategies for modelling transparency, fostering critical literacy, and addressing equity concerns. What emerges is a dialogical supervision model that safeguards integrity and deepens reflexivity. Rather than viewing AI as a threat or panacea, this approach recognises it as a pedagogical provocation - an opportunity to renew attention to accountability, cultivate critical digital scholarship, and prepare researchers who can thoughtfully inhabit technological futures.

This chapter has focused on qualitative dissertation research, where LLM applications and interpretive tensions are most evident; however, supervisors of quantitative and mixed-methods research face comparable challenges regarding transparency, skill development, and the critical use of computational tools. The principles of reflexive integration and digital stewardship offer conceptual foundations adaptable across methodological traditions. Several research directions emerge from this analysis. First, empirical evaluation of the SUPERVISE framework's effectiveness would illuminate which dimensions most significantly impact student outcomes, supervisor confidence, and research quality. Longitudinal studies tracking students from AI-assisted dissertation work into early career research would reveal whether supervised AI integration cultivates or compromises long-term methodological competencies. Second, comparative research across disciplines could identify how epistemological commitments shape appropriate AI integration, potentially yielding discipline-specific guidance. Third, intervention research designing and testing supervisor training programs would address the institutional readiness gaps identified earlier. Fourth, critical studies examining power dynamics, equity implications, and potential biases in AI-mediated supervision would deepen understanding of how technological integration intersects with exist-

ing structural inequalities. Finally, research exploring students' own perspectives—their anxieties, strategies, and ethical reasoning around AI use—would ground supervisory approaches in learner experiences rather than solely supervisor concerns.

The path forward requires institutional support: training supervisors in AI literacy, clear policies that balance innovation with integrity, and resources to ensure equitable access. Yet the fundamental insight remains: AI integration demands not abandoning qualitative research's core commitments but recommitting to them with renewed intentionality, making explicit what was implicit and dialogic what was transmissive.

## References

- Acheampong, I. O., & Nyaaba, M. (2024). Review of qualitative research in the era of generative artificial intelligence [Preprint]. *Social Science Research Network*. <https://doi.org/10.2139/ssrn.4686920>
- Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Hubinger, E., Jackson, K., Kernion, J., Kravec, S., Lin, C., Mueller, J., Ndousse, K., Radhakrishnan, A., Tamkin, A., ... Christiano, P. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv*. <https://doi.org/10.48550/arXiv.2204.05862>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In F. Ferraro & M. Shwartz (Eds.), *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>
- Bhaduri, S., Kapoor, S., Gil, A., Mittal, A., & Mulkar, R. (2024). Reconciling methodological paradigms: Employing large language models as novice qualitative research assistants in talent management research [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2408.11043>
- Bijker, R., Merkouris, S., Dowling, N., & Rodda, S. (2024). ChatGPT for automated qualitative research: Content analysis. *Journal of Medical Internet Research*, X(X), xx–xx. <https://doi.org/10.2196/59050>

- Bird, S. J. (2001). Mentors, advisors and supervisors: Their role in teaching responsible research conduct. *Science and Engineering Ethics*, 7(4), 455–468. <https://doi.org/10.1007/s11948-001-0002-1>
- Bjelobaba, S., Waddington, L., Perkins, M., Foltýnek, T., Bhattacharyya, S., & Weber-Wulff, D. (2024). Research integrity and GenAI: A systematic analysis of ethical challenges across research phases. *arXiv*. <https://arxiv.org/abs/2412.10134>
- Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of “bias” in NLP. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.485>
- Boud, D., & Brew, A. (2013). Reconceptualising academic work as professional practice: Implications for academic development. *International Journal for Academic Development*, 18(3), 208–221. <https://doi.org/10.1080/1360144X.2012.671771>
- Braun, V., & Clarke, V. (2020). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, 18(3), 328–352. <https://doi.org/10.1080/14780887.2020.1769238>
- Chan, C. K. Y., & Hu, W. (2023). Students’ voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20(1), 43. <https://doi.org/10.1186/s41239-023-00411-8>
- Cook, D. A., Ginsburg, S., Sawatsky, A. P., Kuper, A., & D’Angelo, J. D. (2025). Artificial intelligence to support qualitative data analysis: Promises, approaches, pitfalls. *Academic Medicine*, 100(2), 163–171. <https://doi.org/10.1097/ACM.00000000000006134>
- Dai, S.C., Xiong, A., & Ku, L.-W. (2023). LLM-in-the-loop: Leveraging large language model for thematic analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 9993–10001). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-emnlp.669>
- De Paoli, S. (2023). Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review*, 42(2), 234–250. <https://doi.org/10.1177/08944393231220483>
- Eacersall, D., Pretorius, L., Smirnov, I., Spray, E., Illingworth, S., Chugh,

- R., Strydom, S., Stratton-Maher, D., Simmons, J., Jennings, I., Roux, R., Kamrowski, R., Downie, A., Ling Thong, C., & Howell, K. A. (2024). Navigating ethical challenges in generative AI-enhanced research: The ethical framework for responsible generative AI use. *arXiv*. <https://arxiv.org/abs/2501.09021>
- Engeström, Y. (1987). Learning by expanding: An activity-theoretical approach to developmental research. Orienta-Konsultit.
- Engeström, Y., & Sannino, A. (2010). Studies of expansive learning: Foundations, findings and future challenges. *Educational Research Review*, 5(1), 1–24. <https://doi.org/10.1016/j.edurev.2009.12.002>
- Friedman, C., Owen, A., & VanPuymbrouck, L. (2024). Should ChatGPT help with my research? A caution against artificial intelligence in qualitative analysis. *Qualitative Research*, 24(6), 1245–1262. <https://doi.org/10.1177/14687941241297375>
- Gamieldien, Y., Case, J., & Katz, A. (2023). Advancing qualitative analysis: An exploration of the potential of generative AI and NLP in thematic coding [Preprint]. *Social Science Research Network*. <https://doi.org/10.2139/ssrn.4487768>
- Gao, J., Choo, K., Cao, J., Lee, R. K.-W., & Perrault, S. (2023). Feasibility, opportunities, and challenges of utilising AI for collaborative qualitative analysis [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2304.05560>
- Gao, R., Yu, D., Gao, B., Hua, H., Hui, Z., Gao, J., & Yin, C. (2025). Legal regulation of AI-assisted academic writing: Challenges, frameworks, and pathways. *Frontiers in Artificial Intelligence*, 8, 1546064. <https://doi.org/10.3389/frai.2025.1546064>
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), e2305016120. <https://doi.org/10.1073/pnas.2305016120>
- Goyanes, M., Lopezosa, C., & Jordá, B. (2025). Thematic analysis of interview data with ChatGPT: Designing and testing a reliable research protocol for qualitative research. *Quality & Quantity*, 59(1), 1–25. <https://doi.org/10.1007/s11135-025-02199-3>
- Gulumbe, B. H., Audu, S. M., & Hashim, A. M. (2025). Balancing AI and academic integrity: What are the positions of academic publishers and universities? *AI & Society*, 40, 1775–1784. <https://doi.org/10.1007/s00146-024-01946-8>
- Halse, C., & Malfroy, J. (2010). Rethorizing doctoral supervision as professional work. *Studies in Higher Education*, 35(1), 79–92.

- <https://doi.org/10.1080/03075070902906798>
- Hamilton, L., Elliott, D., Quick, A., Smith, S., & Choplin, V. (2023). Exploring the use of AI in qualitative analysis: A comparative study of guaranteed income data. *International Journal of Qualitative Methods*, 22, 1–15. <https://doi.org/10.1177/16094069231201504>
- Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The curious case of neural text degeneration. *arXiv*. <https://doi.org/10.48550/arXiv.1904.0975>
- Jalali, M. S., & Akhavan, A. (2024). Integrating AI language models in qualitative research: Replicating interview data analysis with ChatGPT [Preprint]. *Social Science Research Network*. <https://doi.org/10.2139/ssrn.4714998>
- Kasperiušienė, J., & Mažeikienė, N. (2024). AI-enhanced qualitative research: Insights from Adele Clarke's situational analysis of TED Talks. *The Qualitative Report*, 29(12), 3456–3478. <https://doi.org/10.46743/2160-3715/2024.7652>
- Krásničan, V., Gaižauskaitė, I., Bülow, W., ..., & Tijdink, J. (2024). Transition from academic integrity to research integrity: The use of checklists in the supervision of master and doctoral students. *Journal of Academic Ethics*, 22, 149–161. <https://doi.org/10.1007/s10805-023-09498-0>
- Lazer, D., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., ... & Wagner, C. (2021). Computational social science: Obstacles and opportunities. *Science*, 369(6507), 1060–1062. <https://doi.org/10.1126/science.aaz8170>
- Lee, A. (2008). How are doctoral students supervised? Concepts of doctoral research supervision. *Studies in Higher Education*, 33(3), 267–281. <https://doi.org/10.1080/03075070802049202>
- Levit, N. S., & Saban, M. (2025). When investigator meets large language models: A qualitative analysis of cancer patient decision-making journeys. *npj Digital Medicine*, 8, 15. <https://doi.org/10.1038/s41746-025-01747-3>
- Li, K., Fernandez, A., Schwartz, R., Rios, N., Carlisle, M. N., Amend, G. M., Patel, H. V., & Breyer, B. N. (2024). Comparing GPT-4 and human researchers in health care data analysis: Qualitative description study. *Journal of Medical Internet Research*, 26, e56500. <https://doi.org/10.2196/56500>
- Liang, W., Tadesse, G. A., Ho, D., Li, Y., Zaharia, M., Zhang, C., & Zou, J. (2022). Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*, 4(8), 669–677.

<https://doi.org/10.1038/s42256-022-00516-1>

- Liu, A., & Sun, M. (2023). From voices to validity: Leveraging large language models (LLMs) for textual analysis of policy stakeholder interviews [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2312.01202>
- Löfström, E., & Pyhältö, K. (2020). What are ethics in doctoral supervision, and how do they matter? Doctoral students' perspective. *Scandinavian Journal of Educational Research*, 64(4), 535–550. <https://doi.org/10.1080/00313831.2019.1595711>
- Morgan, D. L. (2023). Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. *International Journal of Qualitative Methods*, 22, 1–12. <https://doi.org/10.1177/16094069231211248>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv*. <https://doi.org/10.48550/arXiv.2203.02155>
- Parker, R., Mancini, K. T., & Abram, M. D. (2023). Natural language processing enhanced qualitative methods: An opportunity to improve health outcomes. *International Journal of Qualitative Methods*, 22, 1–18. <https://doi.org/10.1177/16094069231214144>
- Perkins, M., & Roe, J. (2024). The use of generative AI in qualitative analysis: Inductive thematic analysis with ChatGPT. *Journal of Applied Learning & Teaching*, 7(1), 390–395. <https://doi.org/10.37074/jalt.2024.7.1.22>
- Pizzolato, D., Labib, K., Skoulikaris, N., Evans, N., Roje, R., Kavouras, P., Aubert Bonn, N., Dierickx, K., & Tjldink, J. (2022). How can research institutions support responsible supervision and leadership? *Accountability in Research*, 1–23. <https://doi.org/10.1080/08989621.2022.2112033>
- Prescott, M. R., Yeager, S., Ham, L., Rivera Saldana, C. R., Serrano, V., Narez, J., Paltin, D., Delgado, J., Moore, D. J., & Montoya, J. (2023). Comparing the efficacy and efficiency of human and generative AI: Qualitative thematic analyses. *JMIR AI*, 2, e54482. <https://doi.org/10.2196/54482>
- Press, O., Smith, N. A., & Lewis, M. (2021). Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv*. <https://doi.org/10.48550/arXiv.2108.12409>
- Pyhältö, K., Vekkailla, J., & Keskinen, J. (2015). Fit matters in the supervisory relationship: Doctoral students' and supervisors' percep-

- tions about supervisory activities. *Innovations in Education and Teaching International*, 52(1), 4–16. <https://doi.org/10.1080/14703297.2014.981836>
- Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63, 1872–1897. <https://doi.org/10.1007/s11431-020-1647-3>
- Rafi, M. S., & Amjad, I. (2025). The role of generative AI in writing doctoral dissertation: Perceived opportunities, challenges, and facilitating strategies to promote human agency. *Discover Education*, 4, 165. <https://doi.org/10.1007/s44217-025-00503-9>
- Reynolds, L., & McDonnell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. *arXiv*. <https://doi.org/10.48550/arXiv.2102.07350>
- Roxå, T., & Mårtensson, K. (2015). Microcultures and informal learning: A heuristic approach towards understanding teaching in higher education. *International Journal for Academic Development*, 20(2), 193–205. <https://doi.org/10.1080/1360144X.2015.1029929>
- Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., & Gurevych, I. (2021). How good is your tokeniser? On the monolingual performance of multilingual language models. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (pp. 3118–3135). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.243>
- Schroeder, H., Le Quéré, M. A., Randazzo, C., Mimno, D., & Schoenebeck, S. (2024). Large language models in qualitative research: Can we do the data justice? [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2410.07362>
- Shanahan, M. (2022). Talking about large language models. *arXiv*. <https://doi.org/10.48550/arXiv.2212.03551>
- Siiman, L., Rannastu-Avalos, M., Pöysä-Tarhonen, J., Häkkinen, P., & Pedaste, M. (2023). Opportunities and challenges for AI-assisted qualitative data analysis: An example from collaborative problem-solving discourse data. In *International Conference on Innovative Technologies and Learning* (pp. 95–108). Springer. [https://doi.org/10.1007/978-3-031-40113-8\\_9](https://doi.org/10.1007/978-3-031-40113-8_9)
- Siemens, G. (2004). Connectivism: A learning theory for the digital age. *International Journal of Instructional Technology and Distance*

- Learning, 2(1), 3–10. [http://www.itdl.org/Journal/Jan\\_05/article01.htm](http://www.itdl.org/Journal/Jan_05/article01.htm)
- Smith, J. A., Flowers, P., & Larkin, M. (2022). *Interpretative phenomenological analysis: Theory, method and research* (2nd ed.). Sage.
- Southworth, J., Migliaccio, K., Glover, J., Reed, D., McCarty, C., Brendemuhl, J., & Thomas, A. (2023). Developing a model for AI Across the curriculum: Transforming the higher education landscape via innovation in AI literacy. *Computers and Education: Artificial Intelligence*, 4, 100127. <https://doi.org/10.1016/j.caei.2023.100127>
- Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods*, 23, 1–18. <https://doi.org/10.1177/16094069241231168>
- Than, N., Fan, L., Law, T., Nelson, L. K., & McCall, L. (2025). Updating “The future of coding”: Qualitative coding with generative large language models. *Sociological Methods & Research*, 54(1), 3–28. <https://doi.org/10.1177/00491241251339188>
- Theelen, H., Vreuls, J., & Rutten, J. (2024). Doing research with help from ChatGPT: Promising examples for coding and inter-rater reliability. *International Journal of Technology in Education*, 7(4), 789–805. <https://doi.org/10.46328/ijte.537>
- Törnberg, P. (2023). ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages against a codebook. *Computational Communication Research*, 5(2), 224–239. <https://doi.org/10.48550/arxiv.2304.06588>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates. <https://doi.org/10.48550/arXiv.1706.03762>
- Wachinger, J., Bärnighausen, K., Schäfer, L. N., Scott, K., & McMahon, S. A. (2024). Prompts, pearls, imperfections: Comparing ChatGPT and a human researcher in qualitative data analysis. *Qualitative Health Research*, 34(8), 1234–1248. <https://doi.org/10.1177/10497323241244669>
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language

- models. *arXiv*. <https://doi.org/10.48550/arXiv.2112.04359>
- Xiao, Z., Yuan, X., Liao, Q., Abdelghani, R., & Oudeyer, P.-Y. (2023). Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. *IUI Companion*, 456–467. <https://doi.org/10.1145/3581754.3584136>
- Yue, Y., Liu, D., Lv, Y., Hao, J., & Cui, P. (2025). A practical guide and assessment on using ChatGPT to conduct grounded theory: Tutorial. *Journal of Medical Internet Research*, 27, e70122. <https://doi.org/10.2196/70122>
- Zhang, H., Wu, C., Xie, J., Rubino, F., Graver, S., Kim, C., Carroll, J. M., & Cai, J. (2025). Exploring inductive and deductive qualitative coding with AI: Investigating inter-rater reliability between large language model and human coders [Preprint]. *AHFE International*. <https://doi.org/10.54941/ahfe1006232>
- Zhang, H., Wu, C., Xie, J., Rubino, F., Graver, S., Lyu, Y., Carroll, J. M., & Cai, J. (2023). Redefining qualitative analysis in the AI era: Utilising ChatGPT for efficient thematic analysis. *Computers in Human Behavior*, 152, 108144. <https://doi.org/10.1016/j.chbah.2025.100144>

# Minimum Thresholds of Relationship: AI and Autobiographical Writing

by Lara BALLERI

## 1. Introduction

In recent years, the ability of machines to simulate interactions has been opening up new scenarios for the human and social sciences, while raising epistemological, ethical, and pedagogical questions. The spread of conversational Artificial Intelligence, in particular, calls for reflection on the potential of this language within educational and self-reflective processes, since the words it generates seem not to be limited to conveying information, but instead come to assume the role of a reflective and symbolic space.

Recalling Foucault's words (1992), the «technologies of the self» are that set of practices that enable the individual to work on body, thoughts, and behaviors in order to transform oneself. In light of these technological transformations, it becomes pertinent to ask whether generative AI can be regarded as a new technology of the self, and what relationship it establishes with the formation of the individual.

The doctoral research presented here arises at the crossroads between conversational Artificial Intelligence and autobiographical writing, with the aim of questioning the formative potential of this combination for adults. The experimental design involved three groups of university students, invited to compose an autobiographical text starting from a narrative prompt focused on their own name, considered as an autobiographical threshold. The groups were then differentiated in the following conditions: one received AI-generated feedback calibrated on criteria of listening and empathy; the second engaged in an autonomous re-

reading of their text; and the third, as a control group, received neither feedback nor further prompts.

The aim was to observe the effects of these stimuli on self-awareness and self-efficacy, through psychometric instruments complemented by qualitative analysis of texts and subjective perceptions. An unexpected finding emerged in the group that received AI feedback, in which several participants reported perceiving listening and recognition, in some cases even spontaneously replying to the message received, thereby initiating an epistolary exchange. This phenomenon raises crucial questions about the value of perceived recognition and the possibility that AI may be experienced as a symbolic and relational interlocutor when used in autobiographical contexts.

In this framework, the contribution seeks to propose an original reflection on the concept of the «minimum threshold of relationship» activated through AI, with particular attention to the pedagogical and ethical implications that arise when listening is experienced even in the absence of a human interlocutor.

## **2. Theoretical Framework**

From the literature analyzed, three main strands emerge and intertwine in the present study.

The first concerns autobiographical writing, consolidated as a pedagogical practice capable of supporting processes of awareness, agency, and project-making (Demetrio, 1995; Pineau, 2005; Ricoeur, 1990).

The second strand focuses on conversational Artificial Intelligence, which, although lacking semantic understanding, through prompting logic and dialogical exchange acts as a symbolic interlocutor able to stimulate reflexivity (Cristianini, 2023; Floridi, 2023; Albanesi, 2023).

Finally, the third strand highlights the tendency of individuals to attribute human qualities to machines, activating forms of anthropomorphization that produce relational effects even in the

absence of reciprocity (Weizenbaum, 1966; Nass & Moon, 2000; Ciechanowski et al., 2019).

## *2.1 Autobiographical Writing and Education*

Autobiographical writing has long played a central role in pedagogical reflection, standing as a true educational device. Far from being a mere exercise in memory, self-narration constitutes an intentional act of reworking experience, enabling the person to reconstruct their own life and attribute meaning to it. As Demetrio (1995; 1996; 2020) emphasizes, telling one's story amounts to practicing a form of self-care, since through the written word the subject activates a transformative process that integrates cognitive, emotional, and symbolic dimensions, contributing to the construction of one's identity and awareness of life trajectories.

From this perspective, autobiography allows the reorganization of facts into a meaningful plot, capable of generating new interpretations and future perspectives. Pineau (2005; 2012) insists on the educational and transformative value of narration, stressing how self-narration can become a place of learning where experiential knowledge transforms into reflective knowledge. For Cavarero (2000), identity takes form in narration, as the possibility of being narrated and narrating oneself, thereby giving thickness to subjectivity. Similarly, Ricoeur (1990) introduces the notion of «narrative identity» showing how the self is constituted in the tension between permanence and change, between continuity and transformation. This dimension was further deepened by Formenti (2011; 2013), who proposes the idea of plurality accompanying the individual in autobiographical practice, constantly renegotiated through writing and dialogue with one's social and cultural contexts.

Batini and Del Sarto (2005) further emphasized the heuristic function of narration, understood as a tool for critical reflection and transformation of educational practice. Batini and Zaccaria (2000; 2002) showed how autobiographical writing fosters personal orientation, enabling individuals to attribute meaning to

their experiences and consciously plan for the future. Narrating oneself thus means composing a plot that reconciles heterogeneous fragments of life, proper to a plural existence, and can offer coherence to the discontinuities of the biographical path. It is within this framework that the present study is situated, using autobiographical narration as an educational experience in itself, and questioning the effect of interaction with an artificial interlocutor.

## *2.2 Conversational AI as a Symbolic Interlocutor*

The advent of large language models has made conversational Artificial Intelligence a central topic in contemporary debate. Tools such as ChatGPT generate coherent and contextual texts, but their communicative effectiveness does not derive from a semantic understanding of language, as Cristianini (2023) clarifies. These systems operate through statistical correlations, processing vast amounts of data to predict the next word, package it, and release it; the result is convincing outputs without any awareness.

The emergence of forms of perceived recognition in relationships with AI systems raises significant questions, since the gap between dialogical appearance and absence of consciousness calls into question the validity of listening activated through interaction with an artificial interlocutor. Floridi (2023) defines these systems as forms of «agency without intelligence»: technologies that generate tangible effects in communication and society despite lacking intentionality. Ryan (2020) warns against the risk of unconscious delegation, where textual fluency is mistaken for genuine understanding. Moreover, the transparency of such systems is structurally limited, as models remain opaque: their inner workings are neither fully observable nor interpretable (Ananny & Crawford, 2018). This implies dealing with a technology that acts as a symbolic interlocutor without offering verifiability guarantees, making caution necessary in adopting AI-generated feedback, whose apparent legitimacy rests on pro-

cesses that are unverifiable and certainly unconscious.

Alongside these critical reflections, the literature highlights the pedagogical potential of prompting, that is, the art of formulating questions and instructions to machines. Albanesi (2023), Di Bello (2024), and Alto (2024) describe it as a strategic competence requiring clarity of objectives and linguistic awareness—a perspective that frames the relationship with AI as an exercise in critical thinking. International research has developed methodologies showing that the quality of responses varies depending on how the prompt is structured: one-shot, few-shot, and chain-of-thought prompting (Wei et al., 2022; White et al., 2023) are significant examples. This means that prompting, as a meaningful act in terms of user awareness, also requires specific technical skills to formulate questions that align with the desired feedback. Formulating a good question to a conversational system entails defining an objective, selecting words accordingly, and logically structuring the information; in this sense, interaction with AI constitutes a reflective space in which to exercise awareness regarding issues and questions, while engaging with language.

Nevertheless, even a superficial prompt receives a response, and AI does not demand deeper awareness from the user, limiting itself to generating plausible text. Critical rigor remains therefore a human prerogative: it is the person who decides how to steer the dialogue, what depth to seek, and with what discipline to elaborate their thought. The machine does not distinguish between distracted and deliberate interaction; what makes the difference are the intentions and competencies of the writer, who can transform the mere act of typing into a reflective and educational practice.

### *2.3 Anthropomorphization and Cultural Imaginary*

One of the most relevant aspects of interaction with conversational Artificial Intelligence concerns the perception of relationship that users experience despite the absence of a human interlocutor. Already Weizenbaum (1966), with the famous ELIZA

experiment, had shown how a program capable of repeating and reformulating sentences could lead people to feel understood—the so-called «Eliza Effect», which highlighted the tendency to project human qualities onto automated systems.

In the following years, the CASA paradigm (Computers Are Social Actors) proposed by Nass and Moon (2000), along with the studies of Reeves and Nass (1996), confirmed that individuals treat computers as social subjects, applying rules of politeness and interaction typical of interpersonal relationships. This responding to machines as if they possessed agency signals the perception of a dialogical exchange that exceeds the real capabilities of the systems.

Within this framework, authors such as Boucher (2024) and Murray (2024) stress that AI should be understood as an integral part of the contemporary cultural imaginary, which shapes our expectations of relationships; Murray speaks of an «algorithmic imaginary», namely the projection of meanings that leads us to experience machines as presences endowed with symbolic and emotional depth.

More recent studies have documented how anthropomorphization influences the user experience. Opportunities and risks are reported (Złotowski et al., 2015); it is highlighted that the perception of empathy affects user satisfaction (Ciechanowski et al., 2019); it is observed that attributing psychological traits reinforces the tendency toward continued use (Pelau et al., 2021); finally, it is emphasized that satisfaction and perceived effectiveness can fuel the illusion of understanding, concealing the structural limits of the technology (Xie et al., 2023).

In this scenario, pedagogical responsibility lies in critically guiding the use of AI, designing interactions that make technological limits visible while at the same time offering opportunities for awareness and autonomy. The challenge becomes sharper when human-machine interaction takes place in contexts of self-formation, which require an ethical and cultural background capable of supporting an authentic use of technology.

### 3. Research Design and Experimentation

In the Digital Humanities research reported here, a comparative–experimental design was adopted, with three groups subjected to different feedback conditions. Participation was voluntary and anonymous, with information provided about the goals and methods of the research. All participants wrote an autobiographical text based on a common prompt centered on their relationship with their own name: “Tell the story of your relationship with your given name and how it has evolved over time”. The choice of the name as a reflective starting point is based on its universal and symbolically rich nature; it represents an identity element capable of activating memory, belonging, and reflection (Demetrio, 1995; Cavarero, 2000; Pina-Cabral, 2015; Messuri & Balleri, 2024; Balleri, 2024). In this contribution, the name is not conceived or treated in its linguistic sense but is used as an autobiographical narrative starter to ensure coherence and depth in the collected narratives.

The experiment was structured in five phases:

1. completion of the initial questionnaire (t0);
2. autobiographical writing;
3. differentiated feedback for the three groups;
4. completion of the final questionnaire (t1);
5. analysis of the texts produced.

For group G1, the AI feedback was designed to avoid judgments and directive advice, instead favoring an empathetic and respectful style; the feedback included mirroring, open-ended questions, and appreciative comments, using a formal register in the second person of courtesy, in Italian language. Group G2 reread their own autobiographical text after a few days, a condition aimed at evaluating whether time could promote reflective processes. Group G3 participated only in the writing task, without further stimuli.

This design made it possible to compare the effect of algorithm-

mic mediation with other forms of feedback or with the absence of feedback.

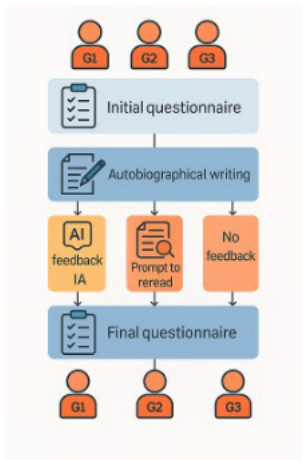


Fig. 1 Flowchart of the experimental design: participants were divided into three groups (G1, G2, G3). After completing an initial questionnaire, all groups engaged in autobiographical writing. Each group then received a different condition (AI feedback, prompt to reread, or no feedback), followed by a final questionnaire.

For the experimental phase, both quantitative and qualitative tools were used within a mixed-methods approach. On the quantitative side, the Rosenberg Self-Esteem Scale (1965), reinterpreted as an indicator of reflective and identity awareness, and the Schwarzer and Jerusalem General Self-Efficacy Scale (1995), adopted as a measure of perceived self-efficacy, were employed. The latter served as a transversal indicator to capture variations linked to reflexivity and self-perception. Both scales were administered at two points in time, before and after the differentiated stimulus (t0 and t1), using extended versions of twenty items evenly distributed across the two administrations.

On the qualitative side, autobiographical texts were analyzed to identify narrative cores linked to awareness, agency, and reflexivity. In addition, the post-intervention questionnaire gathered participants' subjective evaluations both through closed-ended items and open-ended responses dedicated to personal reflections.

### 3.1 Study sample

The final sample consisted of 126 adults engaged in university studies, evenly distributed across the three groups (42 each). The majority identified as female (93.7%), a small percentage as male (2.4%), while 3.9% preferred not to specify. The age distribution was varied: 34% were over 41, 26% aged 31–40, 16% aged 26–30,

14% aged 21–25, and 10% aged 18–20. With regard to previous experiences of autobiographical writing, 40% reported occasionally jotting down thoughts, 39% had no prior experience, 15% had kept a diary in the past, 4% still kept one, and 2% maintained a dream diary. These data illustrate a heterogeneous sample in relation to autobiographical practice.

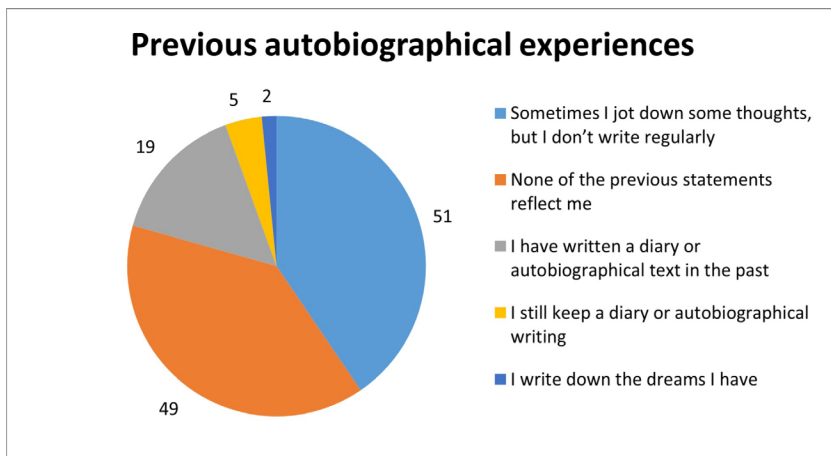


Fig. 2 Distribution of participants' previous autobiographical experiences. The majority reported occasionally jotting down some thoughts without writing regularly (51%) or not identifying with any of the listed statements (49%). Smaller proportions had written a diary or autobiographical text in the past (19%), still keep a diary or autobiographical writing (5%), or write down their dreams (2%).

#### 4. Results

The analysis followed a triangulation logic. Quantitative data were processed with paired-sample statistical tests and effect size calculations; autobiographical texts and qualitative questionnaire responses were thematically coded in order to identify indicators of self-awareness, critical reflection, and agency. The integration of these two levels provided a comprehensive picture of the processes activated, in line with a pedagogical perspective

that seeks to value both the observable dimension and the subjective, experiential one.

The quantitative analysis showed that self-awareness registered the most significant changes (Fig. 3). A particularly marked increase was observed in the group that received AI feedback (G1), with an average variation of +0.23 between t0 and t1; in the autonomous rereading group (G2), the increase was more modest (+0.03), while in the control group (G3) values showed a slight average decrease (-0.04), despite a positive median. These data indicate that autobiographical writing, when accompanied by some form of feedback, tends to have a greater impact on self-awareness, whereas writing alone, in the absence of additional stimuli, does not appear to produce significant changes.

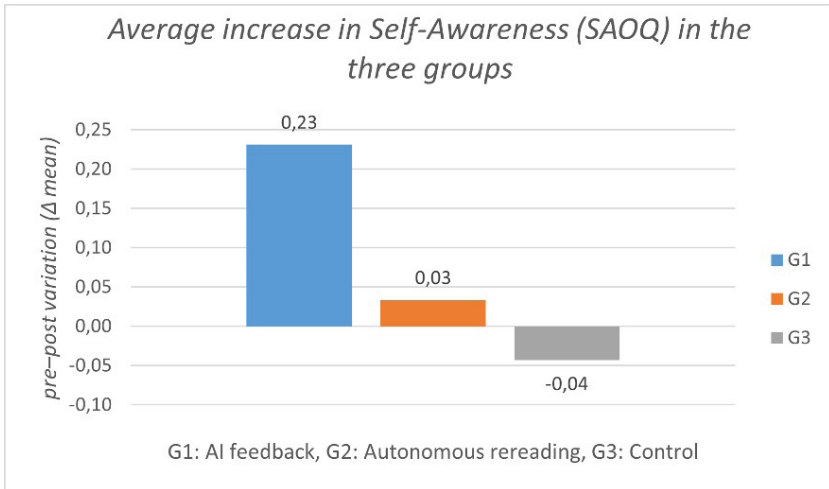


Fig. 3 Average pre–post increase ( $\Delta$  mean) in Self-Awareness (SAOQ) across the three groups. Participants who received AI feedback (G1) showed the highest gain compared to the autonomous rereading group (G2) and the control group (G3).

The trajectory of self-efficacy was different, with variations between t0 and t1 remaining limited across all groups (Fig. 4). The average increase was +0.04 in G1 and +0.12 in G2, while the control group remained almost unchanged (+0.01); none of the

observed changes reached statistical significance, and effect sizes were in the low range. This result suggests that self-efficacy is a construct less sensitive to short-term stimuli and that, in order to be modified, it may require longer interventions and practical situations that demand its concrete exercise.

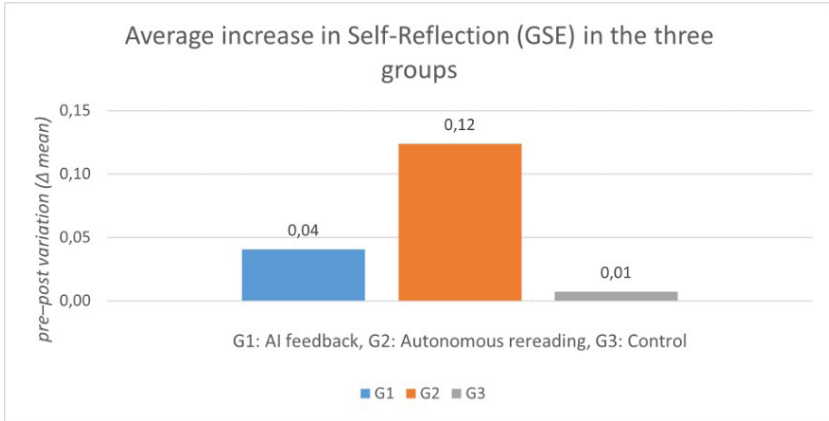


Fig. 4 Average pre–post increase ( $\Delta$  mean) in Self-Reflection (GSE) across the three groups. The autonomous rereading group (G2) reported the highest increase, followed by the AI feedback group (G1), while the control group (G3) showed almost no change.

Alongside these data, the post-intervention questionnaires, administered at time t1, collected participants' perceptions of the usefulness of the experience. All groups expressed positive evaluations, with averages ranging between 3.4 and 3.8 on a 1–4 scale; notably, the control group, which only carried out the autobiographical writing activity, reported strong appreciation (Fig. 5). This finding certainly deserves further exploration, including specific follow-up studies, to verify whether the shorter overall duration of the activity contributed to its perceived pleasantness (only autobiographical writing, without feedback or rereading, plus related questionnaires). Nevertheless, the higher average reported by the control group (G3) shows that even writing alone, without feedback, is perceived as a meaningful experience, albeit with different characteristics compared to the conditions that

involved AI feedback or autonomous rereading; this may reflect the transformative power inherent in autobiographical practice, capable of activating reflection and well-being even without external stimuli.

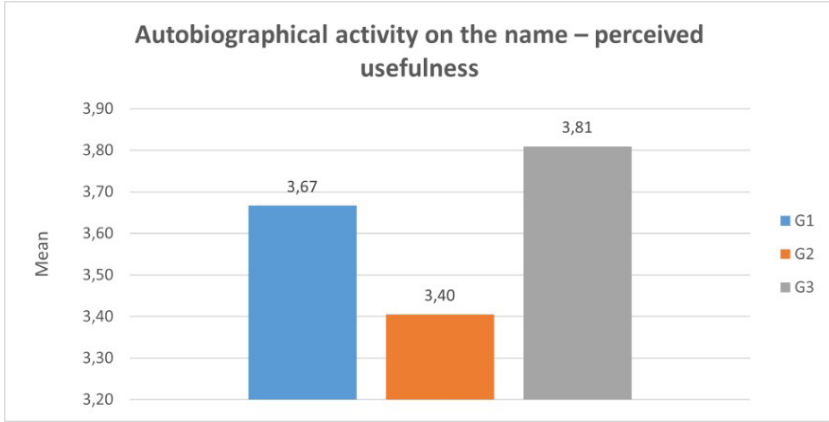


Fig. 5 Perceived usefulness of the autobiographical activity on one's name, by group. Participants in the control group (G3) reported the highest mean score ( $M = 3.81$ ), followed by the AI feedback group (G1;  $M = 3.67$ ), while the autonomous rereading group (G2) reported the lowest perceived usefulness ( $M = 3.40$ ).

The most rewarding aspects included the opportunity to capture new nuances of one's past, to reactivate memories, and to know oneself better. The AI feedback group gave the highest ratings in most dimensions, with a general average of 3.76. Autonomous rereading was also appreciated, though with greater internal variability.

Of particular interest is the comparison between psychometric data and participants' subjective evaluations. In the case of self-awareness (SAOQ), a substantial convergence can be noted: pre-post scores confirm the effectiveness of AI-generated feedback (G1), which also emerged as the most impactful condition in subjective perception. Autonomous rereading (G2) produced minimal increase, accompanied by lower perceived usefulness, while the control group (G3) showed a slight decline in objective

data but continued to assign value to the autobiographical experience. In this dimension, participants seemed able to accurately recognize the variations actually measured.

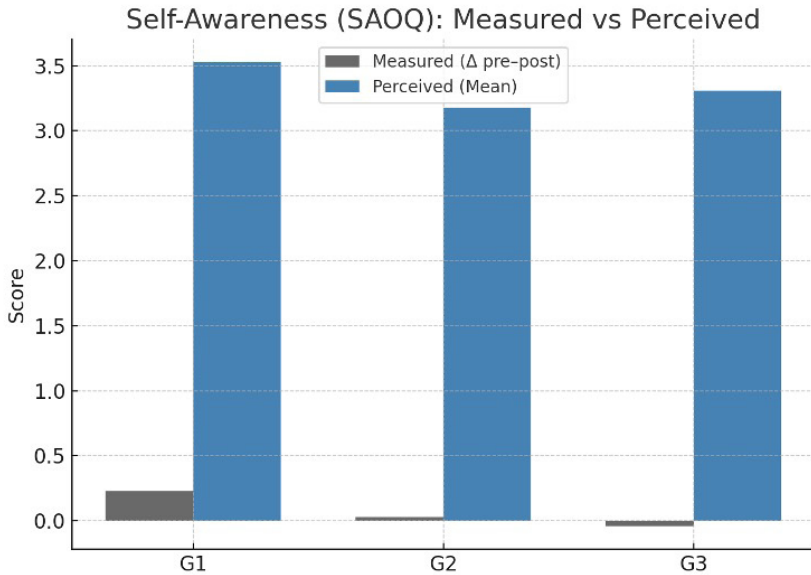


Fig. 6 Self-Awareness (SAOQ): comparison between measured ( $\Delta$  pre-post) and perceived (mean scores) outcomes across groups (G1, G2, G3). Both data sources converge in showing the AI feedback group (G1) as the most effective condition. The autonomous rereading group (G2) shows minimal improvement and lower perceived usefulness, while the control group (G3) presents a slight decline in measured scores but maintains a moderate level of perceived benefit.

The trend for self-reflection (GSE) was different, showing a clear divergence. The measured data indicated a stronger increase in the autonomous rereading group (G2), whereas participants attributed the greatest benefit to the AI feedback group (G1); the control group (G3) remained nearly stable in both measures. This interpretive gap suggests that autonomous rereading may foster improvement in reflective abilities that participants do not fully recognize, while AI feedback, although producing a smaller objective increase, was internalized as a meaningful experience of recognition and listening.

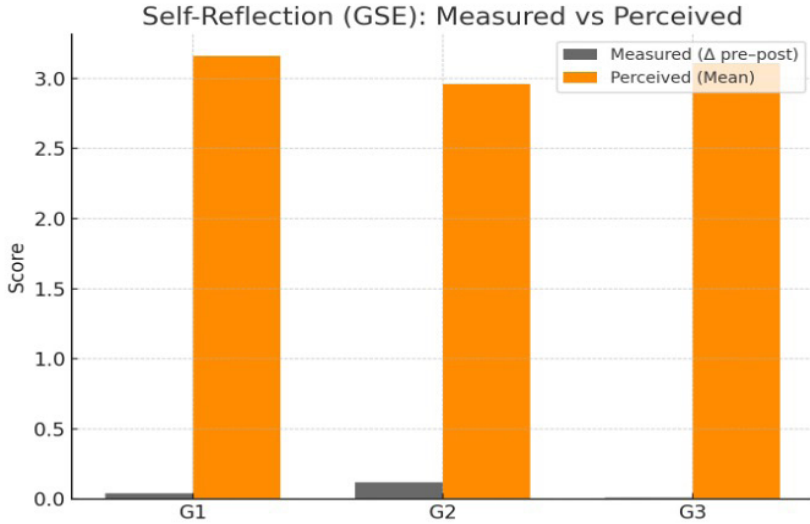


Fig. 7 Self-Reflection (GSE): comparison between measured ( $\Delta$  pre-post) and perceived (mean scores) outcomes across groups (G1, G2, G3). A divergence emerges: the autonomous rereading group (G2) shows the highest measured increase, whereas participants attribute the greatest perceived usefulness to the AI feedback group (G1). The control group (G3) remains stable in both dimensions, indicating that perceived and measured reflections do not always align.

Overall, the parallel between measured and perceived data highlights two different dynamics: on one hand, self-awareness appears to be a more transparent dimension, which participants can recognize in line with the measured data; on the other hand, self-reflection proves more complex, as what is perceived as useful does not always coincide with what is actually modified in the scores.

#### 4.2 *Qualitative Data*

In the optional items dedicated to open responses, several recurring themes emerged, as shown in Figure 8: self-awareness and reflection, enhancement of identity, emotional exposure, recognition, and listening.

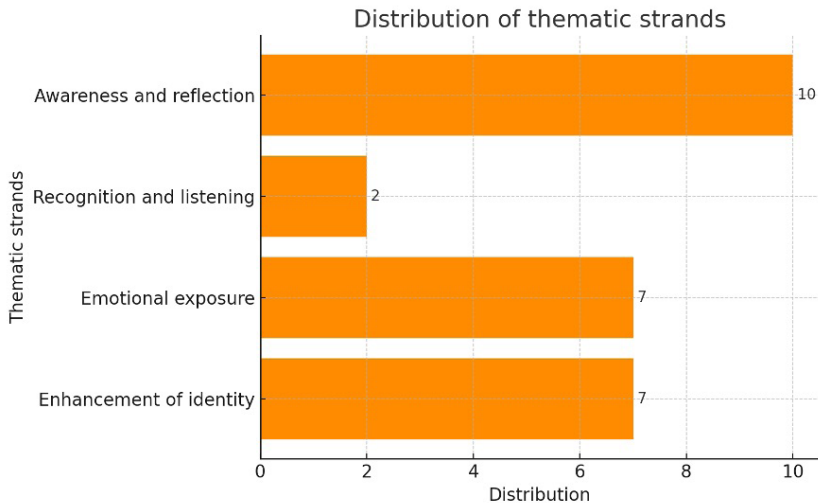


Fig. 8 Distribution of thematic strands emerging from participants' responses to the open-ended questions in the questionnaire. The most frequent themes were "Awareness and reflection" (10) and "Enhancement of identity" (7), followed by "Emotional exposure" (7), while "Recognition and listening" (2) was the least represented.

The analysis of autobiographical texts highlighted recurring themes linked to identity, memory, and belonging. In 43 texts, explicit references to processes of agency and reflexivity were found, more numerous in groups G1 and G2. The prompt focused on one's given name proved effective in eliciting deep memories and identity connections, confirming the value of the name as an autobiographical threshold facilitating the shift from a descriptive to a reflective level.

The trajectory of quantitative data largely corresponded with qualitative evidence. The increase in awareness in G1 was reflected in narratives describing the perception of listening and recognition. Conversely, the stability of self-efficacy scores was mirrored in the reduced presence of explicit references to skills or capacity for action in the texts.

The spontaneous responses of some G1 participants represented an unexpected behavior worthy of attention: they showed that AI-generated feedback, though devoid of both awareness and

relational intentionality, was internalized as a gesture of listening and recognition, even becoming a bridge toward dialogue. This dynamic illustrates the minimum threshold of relationship, made evident as an essential level of recognition that stimulates new writing and emotional resonance, shaping the experience as significant and generative.

In conclusion, taken as a whole, the data provided a complex picture. Self-awareness emerged as the dimension most sensitive to reflective stimuli, with a particularly evident increase in the group that received AI feedback. Self-efficacy instead remained stable, perhaps suggesting the need for more continuous interventions to significantly influence this variable (Fig. 9).

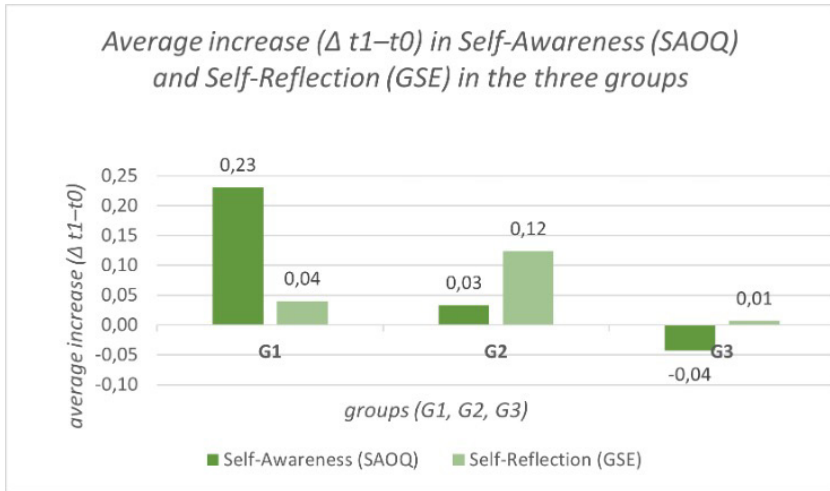


Fig. 9 Average increase ( $\Delta t1-t0$ ) in Self-Awareness (SAOQ) and Self-Reflection (GSE) across the three groups. The AI feedback group (G1) showed the highest gain in self-awareness, while the autonomous rereading group (G2) reported the largest increase in self-reflection. The control group (G3) showed almost no improvement.

Subjective perceptions confirmed the value attributed to the autobiographical experience, while the spontaneous responses of G1 participants represented an unexpected element, revealing the possibility of experiencing interaction with AI as a meaningful relationship. The triangulation of quantitative and qualitative

data allowed these findings to be interpreted in an integrated way: on the one hand, numbers outlining general trends, and on the other, participants' words conveying the experiential and emotional nuances that give those trends meaning.

## 5. Discussion

The results of the research confirm the role that autobiographical writing can play as a powerful reflective tool, capable of activating processes of self-awareness—especially when accompanied by some form of feedback. The increase in self-awareness observed in groups G1 and G2 aligns with what pedagogical literature emphasizes: narration requires a return to the text that opens up new perspectives and connections (Demetrio, 1995; Pineau, 2005). From this perspective, autonomous rereading proved sufficient to trigger an initial reflective movement, while AI-mediated feedback generated a qualitative leap, perceived by participants as a sign of recognition and listening.

The comparison between self-awareness and self-efficacy opens up further considerations. Self-awareness was shown to be sensitive even to short-term stimuli, whereas self-efficacy remained substantially stable. This trend suggests that self-efficacy—understood as the perception of competence and capacity to act (Bandura, 1977)—requires extended and concrete experiences in order to change; the result therefore highlights a characteristic of the construct, which is more closely linked to practice than to narrative reflection alone. The data suggest that AI-mediated autobiographical writing does not directly influence self-efficacy, but primarily stimulates introspective processes and critical self-revision—while also considering the methodological choice to refer to the construct of “self-efficacy” more broadly, including the dimension of self-reflection.

A particularly significant element concerned the spontaneous responses sent by some participants in group G1. These contributions, not foreseen by the procedure, show how AI-generated

feedback was perceived as an authentic gesture of listening and recognition. Several students chose to write back, continuing the exchange initiated by the received message—an unexpected behavior that represents a valuable qualitative finding, testifying to the strength with which the feedback was internalized and the capacity of writing to trigger further movement. The phenomenon observed in the AI feedback group finds confirmation in studies on the Eliza Effect and the CASA paradigm mentioned earlier, since participants attributed intentionality and empathic capacity to statistically generated text, in some cases even responding spontaneously with messages that testified to the perception of a relationship.

This indicates that the educational value is closely tied to how the response is perceived and internalized; thus, when AI-generated feedback is received as legitimizing, it becomes capable of initiating reflective and educational processes. The replies collected expressed gratitude, emotion, and a desire to continue, signaling that the experience was lived as a meaningful and generative relationship. In these responses, the minimum threshold of relationship can be observed in action: an essential level of recognition which, though originating from algorithmic text, was embraced as an invitation to go further, stimulating new writing and emotional resonance.

## 6. Conclusions and Implications

The set of results invites us to consider conversational Artificial Intelligence as an emerging relational environment, capable of generating a perception of recognition—even though it stems from an algorithm without consciousness—sufficient to initiate processes of reflection, autobiographical re-elaboration, and relationship.

The concept of a minimum threshold of relationship represents the original contribution of this study to the debate on human–AI interactions. By this, we mean an essential level of perceived

recognition which does not equate to a fully educational relationship, but which can be interpreted as a liminal experience, in which the word generated by the algorithm is received as a sign of listening and legitimization. It is precisely within this threshold that the pedagogical value of AI-mediated autobiographical writing may reside: a technology that, when placed within an intentional framework, acts as a catalyst of symbolic resonance and opens up a «third space» that evokes some functions of human relationships without replacing them. In this space, the relationship is freed from dynamics that sometimes risk inhibiting or negatively conditioning dialogue between people, such as implicit judgments, expectations, or personal interests; this can allow one to live the experience with greater freedom, at least with respect to these relational dynamics.

From these reflections follow several implications. First, it is necessary to explore more deeply the opacity encountered when interacting with non-transparent systems, whose exact training, including the intentions behind it, is unknown to us. Second, intentional educational frameworks for interaction with AI must be developed, in order to avoid naïve uses that risk fostering relational illusions or reducing the complexity of reflective processes. Since, if designed as a narrative facilitator, AI can foster self-reflection and recognition, functioning as a symbolic mirror, it is crucial that educators and pedagogues acquire prompting skills —capable of orienting instructions toward listening, legitimization, and care— so as to use AI as a tool for narrative literacy and support in self-directed and lifelong learning.

A third implication concerns the ethics of the educational relationship, which demands personal responsibility. The experiment shows that participants attributed the quality of authentic listening to automated feedback, and this makes it essential to avoid letting interaction with AI become a form of unconscious delegation. Critical thinking and individual autonomy must be maintained, so that technology supports the uniqueness of personal stories without flattening them into already known patterns; only within a critical and conscious framework can AI ac-

company authentic educational paths.

A fourth implication regards perspectives for future research. To ensure that all this remains grounded in scientific rigor and informs the competent practice of those working in the care of the individual, it will be necessary to explore under what conditions the minimum threshold of relationship is activated, and what linguistic, textual, and contextual features are sufficient to generate the perception of listening and recognition. Comparative studies conducted in diverse contexts and with heterogeneous samples may clarify which experiences have actual educational value and which risk remaining illusions of reciprocity, unable to generate further engagement.

In conclusion, conversational AI presents itself as a symbolic actor, capable of opening up new scenarios for autobiographical education. The task of pedagogy is to safeguard and guide this emerging space, so that the relationship—even when minimal and mediated—continues to serve the narrating subject and their growth.

## References

- Albanesi, R. (2023). *Come usare al meglio ChatGPT*. Self-published.
- Alto, V. (2024). *Intelligenza artificiale in pratica. Diventare maestri nell'utilizzo dei modelli OpenAI*. Apogeo.
- Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989. <https://doi.org/10.1177/1461444816676645>
- Balleri, L. (2024). Il nome proprio nell'autobiografia tra identità umana e Intelligenza Artificiale, *Educazione Aperta*, 17, 211-228. DOI: 10.5281/zenodo.14603676
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191–215. <https://doi.org/10.1037/0033-295X.84.2.191>
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice Hall.
- Batini, F., & Del Sarto, G. (2005). *Narrazioni di narrazioni. Orientamento*

- narrativo e progetto di vita*. Erickson.
- Batini, F., & Zaccaria, R. (2000). *Per un orientamento narrativo*. FrancoAngeli.
- Batini, F., & Zaccaria, R. (2002). *Foto dal futuro. Orientamento narrativo*. Zona.
- Boucher, G. (2024). AI and the Imaginary: Cultural Formations in Algorithmic Societies. *Journal of Digital Culture*, 12(1), 22–38. <https://doi.org/10.1080/xyz123456>
- Bruner, J. (1990). *Acts of meaning*. Harvard University Press.
- Bruner, J. (1991). The narrative construction of reality. *Critical Inquiry*, 18(1), 1–21.
- Cavarero, A. (2000). *Relating narratives: Storytelling and selfhood*. Routledge.
- Ciechanowski, L., Przegalińska, A., Magnuski, M., & Gloor, P. (2019). In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems*, 92, 539–548. <https://doi.org/10.1016/j.future.2018.01.055>
- Cristianini, N. (2023). *La scorciatoia. Come le macchine sono diventate intelligenti senza pensare in modo umano*. Il Mulino.
- Demetrio, D. (1995). *Raccontarsi. L'autobiografia come cura di sé*. Raffaello Cortina.
- Demetrio, D. (1996). *L'educazione autobiografica*. La Nuova Italia.
- Floridi, L. (2023). AI as agency without intelligence. *Philosophy & Technology*, 36(3), 1–14. <https://doi.org/10.1007/s13347-023-00620-1>
- Formenti, L. (2013). *La formazione raccontata. Narrazione, autobiografia, ricerca*. Unicopli.
- Foucault, M. (1992). *Tecnologie del sé*. (Ed.) L. H. Martin, H. Gutman, & P. H. Hutton. Bollati Boringhieri.
- Messuri, I., & Balleri, L. (2024). Narrarsi attraverso: il nome proprio in autobiografia. *Studi Sulla Formazione Open Journal of Education*, 27(1), 185–192. <https://doi.org/10.36253/ssf-15156>
- Murray, D. (2024). The algorithmic imaginary: AI, perception, and digital realities. *AI & Society*, 39(2), 405–421. <https://doi.org/10.1007/s00146-024-01520-x>
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 72–78). ACM. <https://doi.org/10.1145/285976.285982>

- [org/10.1145/191666.191703](https://doi.org/10.1145/191666.191703)
- Pelau, C., Dabija, D. C., & Ene, I. (2021). What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior*, 122, 106855. <https://doi.org/10.1016/j.chb.2021.106855>
- Pineau, G. (2005). *Apprendere attraverso la narrazione*. Cortina.
- Pineau, G. (2012). Narrare per apprendere: Le storie di vita come pratica formativa. *Rivista di Scienze dell'Educazione*, 50(2), 305–316. doi 10.4454/rse.v4i1.89
- Ricoeur, P. (1990). *Soi-même comme un autre*. Seuil.
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press.
- Ryan, M. (2020). *In AI we trust: Power, illusion and control of predictive algorithms*. Emerald Publishing.
- Schwarzer, R., & Jerusalem, M. (1995). Generalized self-efficacy scale. In J. Weinman, S. Wright, & M. Johnston (Eds.), *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35–37). NFER-Nelson.
- Weizenbaum, J. (1966). ELIZA-A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45. <https://doi.org/10.1145/365153.365168>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Xia, F., Chi, E., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2201.11903>
- White, J., Fu, Q., Hays, S., Sandborn, Z., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2302.11382>
- Xie, J., Wang, S., Huang, Y., & Yu, Z. (2023). Effects of chatbot anthropomorphism on user experience: The role of social presence and empathy. *International Journal of Human-Computer Studies*, 173, 102993. <https://doi.org/10.1016/j.ijhcs.2023.102993>
- Złotowski, J., Yogeewaran, K., & Bartneck, C. (2015). Can robots be perceived as moral agents? *International Journal of Social Robotics*, 7(5), 711–721. <https://doi.org/10.1007/s12369-015-0291-y>

# **Moving Beyond Text**

## **A Comparative Case Study of AI-assisted Audio Interviews in Relation to Textual Data from Mobile Diaries in Singapore**

by Nadia OLISA and Azaleah MOHD ANIS

### **Introduction & Background**

With new technological advancements, researchers have increasing opportunities to leverage such developments to innovate data collection methodologies for greater efficiency in data collection at breadth and in depth. This seems especially relevant for asynchronous qualitative studies, which have typically used text-based data collection methods offering advantages such as the ability to send out questions to a large number of respondents at once, thereby facilitating scalability (Braun et al., 2021; Hunter et al., 2012), and to reach respondents who are unable or unwilling to meet in person, thus expanding the researchers' reach.

From letters sent back and forth through the post, evolving to emails once it became more widely accessible (Meho, 2006; Ratislavová & Ratislav, 2014), and expanding further to online platforms such as social media groups and similar platforms (Creswell, 2007), text-based data collection of qualitative data has been steadily evolving and is likely to continue doing so.

Online diary studies are an example of how text-based data collection has evolved, and in recent years it has become a key method for capturing in-the-moment insights into participant experiences, behaviours and emotions (Jacelon & Imperio, 2005; McCombie et al, 2024), affording participants the privacy to share personal struggles and other thoughts that may be difficult to admit to a stranger (Baker, 2021). Diaries also allow researchers to better balance depth and authenticity, while achieving greater

scalability. Even so, text-based responses remain two-dimensional, having a gap in terms of the loss of non-verbal cues like tone and volume (Martinez-Adrian & Gallardo-del-Puerto, 2021). Typed responses may also be self-edited and not accurately reflect respondents' true feelings (Walther, 2007).

Collecting responses through voice recordings, rather than text, may help to overcome these limitations. It would now also be more easily accessible to respondents as usage of smartphones, enabled with microphones and internet access, becomes ubiquitous (IMDA, 2023). It allows for responses with some non-verbal cues, and potentially more spontaneous and elaborate responses. Studies comparing written and oral answers in web probing show that oral responses via smartphone surveys yield higher-quality data, with more natural and detailed narratives (Geiecke, F., & Jaravel, X, 2004).

A major challenge with asynchronous data collection remains – the researchers' reduced capacity to immediately probe on respondents' answers. Text-based data collection tends to be mono-directional, or have limited capability for two-way communication (emails, online diary platforms, etc.). Both researchers and respondents would typically be responding at their own convenience, hence potentially diluting or losing out on some data points surfaced in the initial response as both may be unable to immediately follow up and elaborate. Simply replacing text-based data collection with audio recordings may have this same issue of mono-directionality.

With the advent of artificial intelligence (AI) tools, this challenge may be closer to resolution. In particular, with the rapid development of large language models (LLM), natural language processing (NLP) models and machine learning (ML), conversational AI chatbots, which have rapidly grown in usage and capability, may unlock bi-directional interaction (Chopra & Haaland, 2023; Lin et al, 2023). These bots can engage in dynamic conversations with users, adapting to their responses and providing more personalised interactions, and are generally more versatile than traditional rule-based chatbots. This would allow for probing re-

spondents' initial responses, to achieve greater depth. Coupled with audio recordings that capture non-verbal cues, this would ultimately allow for more nuanced insights, akin to the level of depth achievable with an in-person interview.

While comparing text-based and audio methods of data collection is not a new area of inquiry (Hohne et al., 2024), introducing AI as a probing tool is a new angle to be explored. This may enhance the quality of data and insights, adding to the nascent body of literature on this subject.

This paper presents a post-evaluation comparison of two distinct approaches to mobile diary research: one that leverages audio entries and AI-assisted probes to deepen participant reflections, and another that adopts a more conventional, text-based diary format moderated by human researchers.

While both studies aim to surface rich qualitative insights over time, they differ significantly in their mode of data capture, prompting mechanisms, and participant interaction. These differences raise important questions about data richness, participant engagement, and researcher intervention. By examining the methodological design, practical execution, and analytical implications of each approach, this paper seeks to highlight how emerging technologies—such as AI powered conversational agents—are reshaping the qualitative research landscape and challenging traditional paradigms of ethnographic enquiry.

This study contributes to the literature in two key areas: qualitative research methods in the social sciences, and the application of AI and machine learning in ethnographic research.

First, within the field of qualitative methods and mobile ethnography, this study advances understanding of different modes of data collection—namely audio versus text—and how it affects the depth, expressiveness, and spontaneity of participant responses. By isolating a shared introductory prompt across two studies with differing formats, the research offers a controlled comparison that sheds light on the influence of modality and probing style on data quality. It reinforces the value of multi-modal approaches in ethnographic fieldwork and provides new

insights into how subtle design choices (e.g., encouraging voice input or leveraging probing strategies) can shape the richness and emotional nuance of qualitative self-reporting.

Second, this study adds to the growing literature on the use of AI and machine learning in ethnographic research (Walsh & Pallas-Brink, 2023; Maltezos et al, 2024). As researchers increasingly experiment with automated tools for data collection, transcription, and analysis, this paper provides early empirical evidence on how AI-assisted probing can enhance participant engagement and elicit deeper reflection. Rather than replacing the ethnographer, the AI in this study acts as a scalable, real-time augmentation tool – enabling more personalised, responsive interaction with participants while maintaining methodological rigour. These findings support emerging conversations around augmented ethnography and illustrate the potential for responsible, human-centred integration of AI in qualitative research workflows (Christou, 2023; Anthony et al, 2023; Grossman et al, 2023).

## **Methodology**

This paper undertakes a comparative methodological analysis of two mobile diary studies by focusing on the participant responses to the first introductory activity in each study. While the broader research topics of the two studies were different —each addressing distinct subject matter and targeting different audience profiles— both studies began with the same foundational question. In both studies, participants were asked to introduce themselves and share personal details about their lives, specifically around the domains of family, work, and hobbies. This shared prompt was chosen as the basis for comparison because it elicited open-ended narrative responses that were intended to set the stage for the rest of the diary task. It also served as a rich organic entry point for assessing the expressive capacity of each method. Moreover, it provides a useful control across both

studies, allowing for a focused and consistent comparison at the entry point of participant engagement.

## **Study Design**

Study A employed a multimodal diary approach that allowed participants to respond using AI-assisted probes. These probes were dynamically generated to prompt based on predetermined probes set by the researcher for further elaboration, clarification, or emotional reflections based on the content of the participant's response. The combination of voice and responsive AI introduced an additional layer of interactivity and personalisation to the data capture process. This study was conducted with 39 respondents over a period of 2 weeks. Participants would be sent a web link via email/ SMS to the question of the day and given the option to type or record their answers directly onto the platform. After the initial response, the AI chatbot would ingest the response and ask a follow-up probe in-the-moment. Prior to fieldwork, the researcher would have programmed the chatbot with the parameters for probing, including topics that require further follow-up. This probing would continue for up to 4 rounds if needed.

Study B, in contrast, used a traditional text-based mobile diary format. Participants entered responses via written text, with human moderators following up through structured or semi-structured probes as needed. Probing was generally limited in volume and scope due to time and operational constraints. The study was conducted with 49 participants over a period of 2 weeks. Participants would be posed questions every alternate day and asked to type in their responses. If responses were vague or mentioned interesting data points, researchers could then reply to the responses and follow-up with a probe to elaborate further. This typically happens within 24 hours of the initial responses and may be followed by a second round of probing, time permitting.

## Data Collection and Comparison Approach

The analysis centres on the first diary entry submitted by participants in both studies, comprising the response to the initial self-introduction prompt. The data analysed for this paper consisted of the full set of participant responses to the first prompt in each study. For Study A, both audio recordings and text responses (where participants did not want to utilise the audio function) along with the AI-generated follow-up responses were included in the analysis. For Study B, only the initial text responses and any immediate follow-up moderator probes were included.

The analysis was conducted in two stages:

1. Intra-study comparison (within Study A): A comparison between the audio responses and their corresponding AI-elicited elaborations was made to assess how AI probing extended or deepened the initial content. Attention was given to audio-specific features such as tone, pauses, hesitation, and expressiveness.
2. Inter-study comparison (Study A vs. Study B): Audio responses (from Study A) were compared against text responses (from Study B), focusing on:
  - a. Length of response (word count or speech duration)
  - b. Narrative richness and emotional nuance
  - c. Spontaneity and naturalness of expression
  - d. Depth of personal disclosure
  - e. Thematic coverage and elaboration
  - f. Probing effectiveness (AI vs. moderator-led)

The analysis employed a combination of qualitative thematic coding and descriptive metrics (e.g., average word count, number of follow-up probes, thematic density) to capture the differences across modes and methods.

By focusing on this single standardised entry point across both studies, the paper is able to control for topic variability while isolating the influence mode of response (audio vs. text) and probing mechanism (AI vs. human) on the depth and quality of qualitative data collected.

## Limitations

There are several important limitations to consider in this analysis:

1. The two studies addressed different research topics and targeted different participant profiles, which could have influenced the tone, openness, or relevance of participants' initial responses. While this variability cannot be fully controlled, both studies included diverse samples in terms of age, gender, and life stage, helping to maintain comparability.
2. Participants in Study A were explicitly encouraged to use audio for their entries, which may have biased the data toward greater depth or emotional expression simply due to mode nudging. While the audio option was not mandatory, this directive may have primed participants to favour audio over text, particularly those comfortable with verbal expression.
3. Since probing strategies differed (AI-driven vs. human-led), differences in depth and elaboration may be partially attributable to technological capability or moderator style, rather than solely to response mode (audio vs. text).

Despite these limitations, focusing the analysis on a shared, neutral entry prompt allowed for meaningful methodological comparison and offered insights into how response mode and probing approach shaped the quality and richness of qualitative data.

**Table: Methodological Comparison at a Glance**

Feature	Study A (Audio + AI Probes)	Study A (Text + AI Probes)	Study B (Text + Human Probes)
Response Mode	Audio	Text	Text
Probing Mechanism	AI-generated, dynamic	AI-generated, dynamic	Human-moderated, semi-structured

Average response length (average)	1290 words	491 words	249 words
Emotional tone capture	Yes (tone, pauses, intonation)	Limited (tone, inferred via text)	Limited (tone, inferred via text)
Spontaneity/ Flow	High	Moderate	Moderate
Depth of disclosure	Often deeper	Variable, less consistent	Limited
Number of Probes per entry (mode)	1 entry	2 entries	Based on 6 separate questions

## Results

The results showed that data collection through audio recording resulted in comparatively richer data, particularly from older respondents. From an operational standpoint, the addition of AI probing reduced manhours while still achieving depth in responses.

Limitations which surfaced with this method was the limitations of the AI chatbot's probing of unexpected responses which had not been accounted for by the researchers, as well as individual respondents' willingness to share, both in terms of preference for text over audio, and in response length regardless of modality.

### 1. Impact on Response Modality: Audio vs. Text Uptake

In Study A, where participants were encouraged to use audio, slightly more than 1 in 2 opted to do so for their first diary entry. Out of all participants, 54% submitted audio responses, while the remaining 46% chose text. While audio use was not mandatory, the design and communication around the study clearly influenced response behaviour, demonstrating that soft nudging toward audio—even without enforcement—can shift participant preferences in modality. By contrast, in Study B, which only

offered text input, all participants responded via written entries, limiting opportunities for vocal nuance or expressive variation.

This difference in uptake highlights the importance of study framing and affordances in shaping how participants choose to communicate. It also sets the foundation for comparing not only the content of the responses but also the expressive range enabled by each mode. While there appeared to be general reticence for taking up the audio option for the first introductory activity, perhaps as “the act of being recorded induced a distinct sense of discomfort” (Höpfner & Promberger, 2023), almost 3 in 4 of the sample used the audio response option for more than half of the expected activities for the study.

## **2. Audio vs. Text: Differences in Quality of Initial Responses**

Based on the introductory responses examined across both study A and B, a clear distinction emerged in the depth and expressiveness of responses across modalities. Audio responses were generally longer in both duration and content, with an average of 1290 words, nearly 2.6 times more words compared to an average of 491 words in text-based responses. Audio entries tended to be more narrative in form, featuring storytelling elements, digressions, and emotional inflection —often revealing participant personality or mood through tone, hesitations, and emphasis.

In contrast, text responses in Study B were more concise and structured, frequently listing facts or descriptions with limited elaboration. For example, participants might mention going to the gym or travelling to different countries without providing additional context or emotion. While some text responses were detailed, they rarely reached the spontaneity or naturalism found in spoken entries.

These findings suggest that audio as a modality encourages a more reflective and unfiltered form of self-expression, likely due to the lower cognitive load of speaking versus writing, and the real-time flow of verbal narration.

### 3. Spontaneity and Naturalness of Expression

Moreover, there appeared to be a distinction observed in the spontaneity and naturalness of participant expression between the two studies. Audio responses in Study A reflected a more conversational and free-flowing style, with participants often thinking aloud, using informal language, and exhibiting natural speech patterns such as pauses, digressions, filler words and shifts in tone. This contributed to a sense of immediacy and authenticity, as participants often explored their thoughts in real time while speaking.

In contrast, the text-based responses in Study B tended to be more deliberate and structured. Participants often self-edited while typing, resulting in cleaner but more condensed narratives. The process of typing appeared to encourage participants to prioritise clarity and brevity over exploration, leading to less spontaneous elaboration and a more transactional tone in many entries.

*“I’m quite invigorated for the week ahead! Since I had a good amount of time over the weekend to relax, I was able to enter Monday with a refreshed headspace to start all my to-do tasks for the week.” – verbatim from Study A*

*“I am feeling good as this is a start of a new week. Therefore I am starting the week positively.” – verbatim from Study B*

### 4. The Role of Probing: AI vs. Human-Led Follow-Ups

The addition of probing mechanisms —AI-assisted in Study A and human-led in Study B— significantly influenced the richness and elaboration of responses.

In Study A, AI probes were triggered systematically after the initial audio input. On average, each participant received 1-2 follow-up prompts, typically tailored to a specific detail or emotional cue in the original recording. These probes encouraged participants to clarify, expand, or reflect, often resulting in a second layer of data with greater emotional depth or contextual specificity.

Notably, AI probing was effective at picking up subtle leads (e.g., a participant briefly mentioning politics and following current affairs) and guiding them to elaborate on these themes.

In Study B, probing was more variable. Human moderators issued follow-ups in roughly 15% of cases. These follow-ups were typically shorter and less organic, and often arrived after a time lag, which may have affected participant responsiveness or spontaneity. As a result, the depth of elaboration in Study B was generally lower and more mechanical in nature.

## **5. Comparative Quality of Data: Length, Depth, and Expressiveness**

The combination of audio input and AI probing in Study A consistently yielded data that was longer, richer, and more emotionally expressive compared to Study B. Key differences included:

*Length:* Audio entries averaged 2.8 minutes, and total word count (transcribed) with probes reached 500–700 words, compared to 180–250 words in Study B.

*Depth:* Participants in Study A shared more personal anecdotes, described emotions in greater detail, and were more likely to reflect on their roles, identities, and stressors.

*Expressiveness:* Non-verbal cues—such as sighs, or tone changes—provided additional interpretive value in Study A, which was absent in written responses.

Overall, the findings suggest that the combination of audio-based expression and responsive AI-driven probing can substantially enhance the quality and nuance of qualitative data collected in mobile diary studies.

The findings from this post-evaluation offer important insights into how response modality (audio vs. text) and probing mechanisms (AI-assisted vs. human-led) influence the quality of data in mobile diary-based ethnographic research. The results

demonstrate that the integration of audio responses combined with AI-generated probes can significantly enhance the depth, expressiveness, and narrative richness of qualitative data, compared to traditional text-based diaries moderated by humans.

Additionally, across the audio responses, there appeared to be some differences between age segments in the length of their responses. All respondents aged 39 years old and below had less than 100 words in their initial responses, compared to 72% of respondents aged 40 years old and above. The remaining 28% of respondents aged 40 years old and above had more than 100 words in their initial responses – between 129 to 303 words to be precise. A similar trend was not observed amongst the text responses.

## 6. Depth of Personal Disclosure

The mode of response also had a noticeable impact on the depth of personal disclosure. Participants in Study A frequently offered richer and more vulnerable narratives, sharing emotions, and reflective thoughts related to their family dynamics, work stressors, and life aspirations.

*“I’ve always enjoyed it when I was younger. From young, my family would always take me out for vacation at least twice a year. So why do I enjoy traveling? Because I find Singapore, I’ve gotten too used to Singapore I guess. So going to certain places gives me new perspective on places. I try out like different things that you can’t find in Singapore or like certain foods where like it’s indigenous to that place. Yeah, I really like trying out new foods.” – verbatim from Study A*

*“So, when I got married, I stopped for quite a bit, and as my children started to grow, become a teenage, I realised that I wasn’t working during that point of time. After I gave birth, I feel like it is time for me to get back in shape and to get myself healthy. So, I started doing jogging. I would tend to go jogging with my husband, and from there, we did join the half marathon.” – verbatim from Study A*

The act of speaking, combined with the immediacy of AI probing, appeared to lower the cognitive and psychological barriers to self-disclosure, enabling participants to open up more deeply.

Conversely, while participants in Study B did provide factual information about their lives, the degree of introspection and emotional depth was generally shallower. Text entries often focused on surface-level descriptions, with fewer personal anecdotes or emotional reflections. This difference suggests that audio may create a safer or more intuitive space for participants to articulate complex, personal narratives, especially when supported by timely follow-up prompts.

*“Feeling blue cos it’s Monday and I’ve just returned from a short overseas trip! I don’t want to go back to work :(” – verbatim from Study B*

*“I am feeling good! Woke up early today!” – verbatim from Study B*

## **Discussion**

### **1. Enhanced Expressiveness Through Audio**

Consistent with prior work in multimodal ethnography and qualitative research, the audio entries in Study A allowed participants to express themselves more spontaneously and naturally. The presence of tone, pauses, and verbal inflection not only contributed to a richer interpretive layer, but also helped surface emotional nuance that was often missing in the text responses from Study B. These findings reinforce the value of incorporating voice-based methods into ethnographic research particularly when aiming to understand lived experience, emotion, and personal meaning-making. This may be particularly so for older respondents, who appear to be more expressive when audio recording their responses compared to writing it down or typing it out.

### **2. The Value of Probing: AI vs. Human**

A notable contribution of this study is the demonstration of

how AI-assisted probing can deepen participant reflection and encourage more elaborate responses in real-time. While human moderators in Study B were effective to a degree, the variability in follow-up timing, tone, and frequency likely contributed to the inconsistent quality of elaboration. In contrast, AI in Study A offered immediate, targeted, and scalable engagement, prompting participants to revisit themes or expand on fleeting mentions. These findings align with emerging literature on AI as an augmentation tool, not as a replacement for human insight, but a mechanism to enhance it through consistency, personalization, and responsiveness.

### **3. Implications for Ethnographic and Qualitative Research**

Together, these results suggest that rethinking traditional qualitative methods through the lens of technological augmentation can open new possibilities for richer, more scalable research. Mobile diary studies that integrate audio and intelligent feedback loops may be particularly effective in capturing the texture of daily life in longitudinal ethnographies, especially in contexts where researcher presence is impractical.

Moreover, this approach may support more participant-led narratives, reducing the filtering effect that occurs when moderators overly structure or constrain follow-up engagement. This approach may be more suited to engaging older respondents or those who are less tech savvy, as recording audio allows such participants to more easily record their responses in real time.

At the same time, it is important to acknowledge that participant comfort, literacy, and preferences still play a critical role in response behaviour. While many participants favoured audio when encouraged to do so, a portion still chose text, signalling the need for flexibility in design and a sensitivity to individual communication styles.

### **Study Limitations and Considerations**

Several limitations must be noted. First, the two studies fo-

cused on different research topics and recruited participants from different target groups, which may have influenced participants' willingness to disclose or reflect in depth. While the use of a shared opening prompt helped control for this variable, topic salience could still have impacted engagement levels.

Second, participants in Study A were nudged toward using audio, which could have influenced both modality selection and the resulting expressiveness —raising questions about how design expectations shape behaviour.

Third, while AI probing was generally effective, its success depends heavily on the quality of the underlying language model, which may not always interpret nuance or cultural references as reliably as a human. Human researchers must then take care to ensure that such nuances of data are not lost amidst seemingly more salient points.

Finally, several ethical considerations emerged regarding the use of AI in this study. Prior to data collection, participants in Study A were informed that their responses would be followed up by a chatbot that generated replies based on their initial inputs. In retrospect, clearer and more explicit communication could have been provided to ensure participants fully understood that they would be engaging directly with an AI system (possibly at the onboarding stage where participants are briefed on what to expect for the study). The phrasing used in the briefing (“reply based on initial responses”) as well as the mere disclosure of AI involvement, may have inadvertently shaped participants' expectations or response patterns.

While transparency is essential for ethical research practice, it is equally important that such disclosures are framed in a neutral manner to minimise potential influence on participant behaviour. Future studies should also consider strategies to manage biases not only in how AI use is disclosed, but also in how AI-driven questioning and follow-up prompts are designed, so as to reduce the likelihood of leading or reinforcing particular response tendencies.

## Conclusion

This paper highlights how response modality and probing mechanisms shape the quality of data in mobile diary-based ethnographic research. The findings demonstrate that audio-based responses, coupled with AI-assisted probing, yield greater narrative depth, emotional nuance, and expressiveness compared to traditional text-based diary methods with human-led follow-ups. These insights underscore the value of incorporating multimodal input and intelligent prompting in qualitative research design.

For researchers, this study suggests that leveraging audio can lower cognitive barriers, encourage more natural storytelling, and capture richer layers of meaning—particularly when combined with responsive AI-driven probes. However, offering participants choice in modality remains important to accommodate varying comfort levels and communication styles.

Future research should explore the application of AI-augmented probing across a wider range of topics, populations, and cultural contexts. Further investigation is also needed into the ethical implications, participant experience, and interpretive boundaries of AI in qualitative research. As machine learning tools continue to evolve, their role in supporting—not replacing—human-centred ethnography will become an increasingly critical area for methodological innovation.

## References

- Anthony, C., Bechky, B. A., & Fayard, A.-L. (2023). "Collaborating" with AI: Taking a system view to explore the future of work. *Organization Science*, 34(5), 1672–1694. <https://doi.org/10.1287/orsc.2022.1651>
- Baker, Z. (2021). Young people engaging in event-based diaries: A reflection on the value of diary methods in higher education decision-making research. *Qualitative Research*, 23(3), 686–705. <https://doi.org/10.1177/14687941211048403>
- Braun, V., Clarke, V., Boulton, E., Davey, L., & McEvoy, C. (2020). The

- online survey as a qualitative research tool. *International Journal of Social Research Methodology*, 24(6), 641–654. <https://doi.org/10.1080/13645579.2020.1805550>
- Chopra, F., & Haaland, I. (2023). Conducting qualitative interviews with AI. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4583756>
- Christou, P. (2023). How to use Artificial Intelligence (AI) as a resource, methodological and analysis tool in qualitative research? *The Qualitative Report*. <https://doi.org/10.46743/2160-3715/2023.6406>
- Geiecke, F., & Jaravel, X. (2024, October 30). AI can carry out qualitative research at unprecedented scale - LSE impact. *LSE Impact - Understanding impact and practice in academic research*. <https://blogs.lse.ac.uk/impactofsocialsciences/2024/10/30/ai-can-carry-out-qualitative-research-at-unprecedented-scale/>
- Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, 380(6650), 1108–1109. <https://doi.org/10.1126/science.adi1778>
- Hershberger, P. E., & Kavanaugh, K. (2017). Comparing appropriateness and equivalence of email interviews to phone interviews in qualitative research on reproductive decisions. *Applied Nursing Research*, 37, 50–54. <https://doi.org/10.1016/j.apnr.2017.07.005>
- Hunter, J., Corcoran, K., Leeder, S., & Phelps, K. (2012). Is it time to abandon paper? The use of emails and the internet for health services research – a cost-effectiveness and qualitative study. *Journal of Evaluation in Clinical Practice*, 19(5), 855–861. <https://doi.org/10.1111/j.1365-2753.2012.01864.x>
- Höhne, J. K., Gavras, K., & Claassen, J. (2024a). Typing or speaking? Comparing text and voice answers to open questions on sensitive topics in smartphone surveys. *Social Science Computer Review*, 42(4), 1066–1085. <https://doi.org/10.1177/08944393231160961>
- Höpfner, E., & Promberger, M. (2023). The elephant in the room-recording devices and trust in narrative interviewing. *International Journal of Qualitative Methods*, 22. <https://doi.org/10.1177/16094069231215189>
- Jacelon, C. S., & Imperio, K. (2005a). Participant diaries as a source of data in research with older adults. *Qualitative Health Research*, 15(7), 991–997. <https://doi.org/10.1177/1049732305278603>
- Lin, C.-C., Huang, A. Y., & Yang, S. J. (2023). A review of AI-Driven

- Conversational Chatbots implementation methodologies and Challenges (1999–2022). *Sustainability*, 15(5), 4012. <https://doi.org/10.3390/su15054012>
- Maltezos, V., Luhtakallio, E., & Meriluoto, T. (2024a). Bridging ethnography and AI: A reciprocal methodology for studying visual political action. *International Journal of Social Research Methodology*, 28(2), 193–208. <https://doi.org/10.1080/13645579.2024.2330057>
- Martínez-Adrián, M., & Gallardo-del-Puerto, F. (2021). Task modality and language-related episodes in young learners: An attempt to manage accuracy and editing. *Language Teaching Research*, 28(6), 2300–2325. <https://doi.org/10.1177/13621688211052808>
- McCombie, C., Miguel Esponda, G., Ouazzane, H., Knowles, G., Gayer-Anderson, C., Schmidt, U., & Lawrence, V. (2024a). Qualitative digital diary methods: Participant-led values for ethical and insightful mental health research. *International Journal of Qualitative Methods*, 23. <https://doi.org/10.1177/16094069241296189>
- Meho, L. I. (2006). E-mail interviewing in qualitative research: A methodological discussion. *Journal of the American Society for Information Science and Technology*, 57(10), 1284–1295. <https://doi.org/10.1002/asi.20416>
- Ratislavová, K., & Ratislav, J. (2014). Asynchronous email interview as a qualitative research method in the humanities. *Human Affairs*, 24(4), 452–460. <https://doi.org/10.2478/s13374-014-0240-y>
- Singapore Digital Society Report. Infocomm Media Development Authority. (2023, November 9). <https://www.imda.gov.sg/about-imda/research-and-statistics/singapore-digital-society-report>
- Walsh, S., & Pallas-Brink, J. (2023). The ethnographer in the machine: Everyday experiences with AI-enabled data analysis. *Ethnographic Praxis in Industry Conference Proceedings*, 2023(1), 512–528. <https://doi.org/10.1111/epic.12185>
- Walther, J. B. (2007). Selective self-presentation in computer-mediated communication: Hyperpersonal Dimensions of technology, language, and cognition. *Computers in Human Behavior*, 23(5), 2538–2557. <https://doi.org/10.1016/j.chb.2006.05.002>

# Human-Machine Feedback Loops in Qualitative Research: Co-Constructing Semi-Structured Interviews with Generative AI

by Giulia Coppo

## 1. Introduction

Over the past ten years, artificial intelligence (AI) has moved from being a distant, technical abstraction—something that once seemed confined to eccentric engineering labs—to being increasingly embedded in our everyday lives. From smart assistants to content recommendations, from predictive texts to personalised ads, AI systems have begun to shape how we communicate, work, and even understand the world around us. The advent and rapid diffusion of generative AI models, particularly large language models (LLMs) like GPTs (Generative Pre-Trained Transformers), has signalled a significant turning point in our relationship with AI; one marked by increasingly fluid and seemingly “natural” human-machine interactions.

This recent wave of AI innovation has been accompanied by considerable hype, both in public discourse and across professional sectors. Debates on AI began to take up significant space in the press (Brenner et al., 2018; Degli Esposti et al., 2024), and conversations about its potential positive or negative impacts in the workplace surged across talk shows and entertaining media (Nader et al., 2024). While these aspects represent crucial transformations our societies are facing, what caught my attention—both as a researcher and as a human navigating this shift—is how we relate to and make sense of these tools, especially in knowledge-producing contexts like academia.

Media scholars have conceptualised this emerging relationship as a form of *human-machine communication* (Guzman and

Lewis, 2020), while others have proposed thinking of these systems as *artificial companions* (Hepp, 2020) capable of engaging in dialogic and affective exchanges that blur traditional lines between user and tool. These perspectives are useful because they invite us to ask not just *what* generative AI does, but *how* it participates in our social, intellectual, and methodological endeavours.

It was precisely this curiosity that led me to experiment with generative AI in my own research practice. As part of the preliminary phase of a project exploring AI-mediated professional practices in political communication, I needed to construct a semi-structured interview guide, one that could adapt to different profiles across a heterogeneous professional field. I had used ChatGPT before, but mostly for repetitive and peripheral tasks such as summarising documents or drafting academic abstracts. This time, however, I decided to engage it directly in my research project: I prompted it to help me draft the interview guide.

What started as a seemingly pragmatic choice quickly unfolded into a deeper set of epistemological and ethical questions: *Can generative AI meaningfully contribute to the construction of qualitative research tools? What kind of assumptions or logics might it embed into the process? And what does it mean, epistemologically, to collaborate with a machine in shaping the very tools through which knowledge is produced?*

This paper explores these questions through a critical, empirical, and methodological lens. It draws on an iterative, autoethnographic approach, where I document my evolving engagement with a generative AI model as I co-develop the interview guide. In doing so, I consider both the technical dynamics and the socio-cultural dimensions of this collaboration, reflecting on how prompts, outputs, and interactions shaped the outcome. As social researchers, we now have easy access to a range of LLMs, which offer a unique opportunity to understand how human social practices influence the use of AI in research, while also being shaped by the very outputs these models generate. My aim here is not to celebrate nor dismiss generative AI. Rather, I argue that researchers should begin to approach these technologies as

*co-actors* in the research process, meaning as agents that not only respond to our commands but subtly influence the epistemic framework we construct.

The present contribution seeks to advance meta-research by proposing novel directions that critically and reflexively engage with generative AI tools, acknowledging both their methodological potentials and their limitations, risks, and ethical implications.

Ignoring these new human-machine dynamics means missing an essential part of how academic knowledge is already beginning to be reconfigured.

## 2. Generative AI in academic research

The widespread adoption and discussion of the term *artificial intelligence (AI)* gained momentum following the introduction of OpenAI's ChatGPT in 2022.<sup>1</sup> This launch sparked significant public interest, as reflected in a surge of Google search activity (Fui-hoon Nah et al., 2023). The term GPT (Generative Pre-Trained Transformer) refers to a class of large language model (LLM) algorithms introduced by OpenAI that utilise deep learning techniques to train on vast amounts of data (Cascella et al., 2023). Based on the transformer decoder architecture, a GPT predicts the next word given the previous ones. The higher the scale of the training data and of the parameters on which it is set, the better the outcome. The improvements of this technology resulted in the rise of what is now generally known as *generative AI*, a type of AI capable of producing human-like texts and content, including music, images, videos, and other forms of media.<sup>2</sup>

---

1 The term "artificial intelligence" was coined in 1956 by the American computer scientist John McCarthy during the Dartmouth Summer Research Project on Artificial Intelligence, which represented the seminal event for AI as a research field.

2 It is important to clarify that GPT models (like GPT-4) are specialised in text generation. While generative AI as a whole includes models for music, images, and video, these are different models from GPT, even if developed by

Generative AI can have a wide range of applications, especially in knowledge-producing and creative industries. For instance, generative AI is being used to generate poems (Köbis & Mossink, 2021), political statements (Bullock & Luengo-Oroz, 2019), and academic papers (Hu, 2023) that are often indistinguishable from those written by humans. However, these applications also raise significant moral, legal, and ethical concerns. Issues such as copyright infringement in AI-generated artworks (Gilotte, 2019), cheating and plagiarism in educational settings, data privacy and security (Siau & Wang, 2020), and the malicious use of deepfakes (Whittaker et al., 2020) highlight the challenges that accompany the growing presence of generative AI.

In the academic context, generative AI applications have sparked a wide debate among scholars. This technology has been described as a tool for enabling high-quality scholarly work; in particular, it is being used for problem formulation, research design, data collection and analysis, interpretation and theorisation, composition, and writing. It can also support assembling data sets, identifying patterns in data, and improving the writing process by assessing argument structure and grammar (Susarla et al., 2024).

Recent studies also offer a nuanced exploration of how AI can be meaningfully integrated into qualitative research, not just as a technical aid, but as a collaborator in shaping methodological approaches and analytical depth. For instance, Christou (2023) positions AI as a versatile tool that, when approached critically, can assist in literature synthesis, theme development, and pattern recognition, while also highlighting the ethical and epistemological risks of over-reliance or uncritical use. To mitigate these, he proposes a set of guiding principles that foreground researcher agency, ethical awareness, and reflective practice.

In parallel, other scholars have taken a broader empirical lens, examining how researchers themselves are responding to AI's growing role in the research process (Chubb et al., 2021). Their findings point to a tension: while AI can undeniably reduce ad-

ministrative burdens and facilitate efficiency, it also introduces new pressures, particularly in an academic culture increasingly driven by metrics and speed. They warn that without careful governance, AI could inadvertently reinforce these demands, potentially undermining the slower, more reflective dimensions of academic work. These contributions call for a balanced and critical engagement with AI, one that embraces its opportunities while remaining attuned to its limitations and the broader contexts in which it is deployed.

At the same time, while these studies provide a valuable foundation for understanding the evolving role of AI in research, they remain largely conceptual and exploratory in scope, and rarely address the lived, situated practices of researchers who actively integrate AI into the development of qualitative methods. Current evolving research practices need to be questioned by academic work that offers a close empirical account of how generative AI might intervene in the crafting of qualitative research tools. This is why, in this work, rather than treating AI as a peripheral aid or abstract concern, I approach it as a co-actor within the methodological process itself, offering a situated account of what it means to research *with* AI, in every sense of the phrase.

### **3. Understanding generative AI through the lens of sociotechnical co-production**

Generative AI models, like GPTs, are not neutral instruments. They are developed through specific social, economic, cultural, and epistemological practices, and in turn, they actively participate in the making of knowledge, identities, practices, and norms. This observation represents the heart of the *sociotechnical co-production* approach, a framework that conceptualises technology and society not as separate domains but as mutually constitutive.

The co-productionist approach developed by Sheila Jasanoff (2004) stresses how power, authority, and cultural values are

embedded in scientific practices, while the sociotechnical dimension emphasises the non-neutrality of technological objects and their interrelation with the social systems (Trist & Bamforth, 1951; Mumford, 2006). In this context, sociotechnical co-production helps us understand generative AI as a product of mutual shaping between human actors, institutions, cultural values, and material innovations, and as an influential actor in shaping scientific knowledge.

In the context of this research, the very prompts I used to guide the model were shaped by disciplinary expectations, cultural codes, linguistic conventions, and my own academic background. However, once generated, the model's outputs did not merely reflect those inputs; they also reframed them, steering the research into directions I had not anticipated. This back-and-forth process reveals a recursive loop between human actors and algorithmic agents (Airoldi, 2021): researchers shape AI through prompts (prompts that themselves carry the human's cultural values, disciplinary assumptions, and literacy skills) while AI, in turn, shapes the research flow through suggestions, cultural framings, response patterns, as well as its encoded structural predispositions and hallucinations.

Part of this work is intended to offer methodological and epistemological tools that can help scholars to engage critically with these human-machine feedback loops when integrating generative AI into their knowledge-making practices. Doing so opens up possibilities for more reflexive, transparent, and context-aware use of AI in the production of knowledge.

#### **4. Relating to generative AI: communication, agency, and companionship**

The relational dynamics that emerge in interactions with generative AI systems such as GPTs can also be understood through the lens of *human-machine communication (HMC)*. HMC is a newly recognised area of research within the study of communication

that traces its origins to 2015, when scholars began to theorise media more fully in the role of a communicator, in response to the increasingly agentic role played by applications and devices (Guzman, 2018; Fortunati and Edwards, 202). In this context, HMC scholars argue that people relate to AI not simply as tools but as communicative actors, engaging in exchanges that shape how individuals perceive themselves, others, and the machine itself (Esposito, 2017; Guzman and Lewis, 2020). Hepp (2020) furthers this perspective by conceptualising AI technologies as *artificial companions*, which he describes as systems that, by simulating human-like communication traits, facilitate forms of quasi-communication. Unlike previous technological tools, artificial companions are designed to be perceived as responsive, attentive, affective, and socially present, thus fostering an illusion of mutual understanding. While Hepp's analysis focuses primarily on embodied or scripted bots, his framework can also be applied to text-based generative systems like ChatGPT, which similarly prompt users to engage in turn-taking exchanges that mimic human dialogue. This is the case of the present research: I engaged with ChatGPT, conscious that it did not represent a mere tool, but rather an "artificial assistant" with whom I had a dialogical conversation during the process of co-construction of the interview guide.

The mutually constitutive relationship between human actors and technological artifacts can also be framed in terms of *symbiotic agency* (Neff and Nagy, 2018). This kind of agency portrays the interaction as a dynamic, interdependent engagement where human and technological agents influence each other, blurring traditional distinctions between what humans want to do and what technology is capable of doing. In today's networked, digital environment, this concept further captures the complex, reciprocal, and co-constitutive nature of human-machine relationships.

Similarly, other theoretical accounts have discussed the concept of agency as constitutive of both humans and machines. Particularly relevant within media and communication studies

is Airoidi's (2021) conceptualisation of algorithms as *social agents*, actively participating in the production of meaning and action in communication and knowledge processes. This conceptualisation resonates with a more critical approach to AI systems as it tries to stress how algorithmic agency intersects with power, control, and reflexive practices. On this same theoretical line, other scholars have proposed a theorisation of *algorithmic agency* that foregrounds the reflexive capacity of humans to exert power over algorithmic systems (Bonini and Trerè, 2021). While their theory has primarily been applied to the gig economy and recommendation engines, it also offers valuable insights for understanding generative AI and humans' interactions. The crafting of prompts, in particular, functions as a site of human agency: users embed disciplinary logics, cultural assumptions, and epistemic intentions into the interaction, thereby exercising influence over the model's output. Taken together, these theoretical accounts highlight the complex interplay of human intentionality and algorithmic influence that characterises the use of generative AI. At the same time, it helps us understand this novel technology as an interactional space that is simultaneously relational, communicative, and structured by asymmetric distributions of power and agency.

## 5. Methodology

In this study, I adopted an iterative, autoethnographic methodology to explore the integration of generative AI into the construction of a semi-structured interview guide. Specifically, I worked with OpenAI's ChatGPT 4.0, a large language model trained on a vast amount of publicly available data, capable of processing and generating human-like language with high contextual sensitivity. At the heart of this approach is the critical observation of the evolving interaction between me and the AI model, during which I have treated the text generated through our exchanges as a form of conversational data. As a social sci-

ences researcher, I am accustomed to reflecting on my own positionality, assumptions, and methodological choices (Alvesson & Sköldbberg, 2017). However, this project posed a new challenging question: what happens when what you previously identified as a research tool begins to “speak back”, offering suggestions, proposing alternatives, and participating in what felt like a collaborative process?

To engage with this question, I drew on La Mendola’s (2009) concept of dialogic methodology, which frames research as an iterative, co-constructed encounter between the researcher and other participants. While traditionally applied to human interactions, I extended these dialogic lenses to my exchanges with the AI. I did not simply issue instructions to the model; instead, I found myself engaged in a back-and-forth process in which the model’s responses shaped my subsequent prompts, and my prompts, in turn, refined its outputs. This dynamic prompted me to develop a new kind of methodological reflexivity, one that acknowledges the epistemic role of generative AI and interrogates its embedded assumptions, power relations, affordances, and limitations.

Rather than treating AI as a technical, passive assistant, I approached it as a semi-autonomous co-participant in the research process. This, of course, does not imply that the AI possesses consciousness or intentionality in the human sense, but it does involve recognising its agentic capacities (i.e., its ability to shape communicative processes, influence epistemic directions, and participate in the co-construction of knowledge). This approach aligns with the Human-Centred AI (HCAI) principles (Riedl, 2019), which emphasise transparency, human agency, and iterative collaboration between users and AI systems.

Overall, I structured the research process across six interrelated and overlapping phases:

- 1. Initial prompt engineering**, in which I used insights from the literature on political communication and my own academic knowledge after years in the field to craft a first version of the interview guide with the assistance of the model;

2. **Critical evaluation and refinement**, where I assessed the AI-generated outputs for clarity, relevance, potential hallucinations, and refined my prompts accordingly;
3. **Systematic analysis of outputs**, focusing on how the model reproduced or challenged common assumptions, power relations, and cultural dispositions, and on whether its suggestions captured the complexity of the professional field;
4. **Participants' incorporation**, where I tested early versions of the guide in interviews and invited interviewees to co-develop new questions or themes based on their personal experience and perspective;
5. **Researcher-machine-participant co-production**, where I integrated interviewees' feedback into new AI prompt cycles, with each round informed by participants from diverse organisational contexts and after having obtained informed consensus;
6. **Ongoing critical iteration**, where I continually refined both the prompts and the interview guide with a focus on improving context-specificity, ethical sensitivity, and analytical depth.

A total of four participants were engaged in the crafting of the final interview guide through the co-assistance of ChatGPT. Tab 1. Summarises their characteristics based on gender, organisational context, and political affiliation (if present).

While the AI was involved in identifying patterns, proposing thematic sections and subsections, and suggesting potential follow-up questions based on the criteria and objectives I defined in the prompt, the final version of the interview guide was the product of extensive re-elaboration on my part. Simultaneously, I critically assessed the AI's contributions through the lens of my academic expertise in political communication and my methodological training in qualitative research design. This involved selecting, adapting, and restructuring the material to ensure conceptual coherence, contextual relevance, and methodological rigour.

The choice of framing this process as an autoethnographic enquiry allowed me to reflect not only on the content produced but also on the evolving relationship between myself, the technological artifact, and the human participants. In addition, by treating generative AI as both a tool and an epistemic actor, I was able

to interrogate myself on the mechanisms by which researchers might be able to co-produce knowledge in ways that are both enabling and constraining.

<b>Interviewee</b>	<b>Gender</b>	<b>Organisational context</b>	<b>Political affiliation</b>
Interviewee 1	Male	Political communication agency	None
Interviewee 2	Male	Municipality	Centre-left
Interviewee 3	Female	Government	Centre-right
Interviewee 4	Male	Public affairs agency	None

Tab 1. Characteristics of the four interviewees who participated in the study and contributed to the interview guide.

### *5.1 Reflexive note on positionality*

At this point, it is important to stress a few lines on my positionality. My engagement with generative AI is shaped not only by my disciplinary training in sociology and social research but also by my personal and professional background in technology. As a self-described tech-passionate and a former advanced data trainer for AI systems, I brought to this research both practical experience and conceptual familiarity with how large language models are trained, fine-tuned, and deployed. This professional experience equipped me with a deeper understanding of the internal logics, limitations, and affordances of generative AI systems like ChatGPT. At the same time, my critical social science training has given me the tools to interrogate these systems as sociotechnical constructs.

This dual orientation —part fascination and part critique— has profoundly shaped how I approached this project. It allowed me to engage with the AI model as a technical artefact and as a communicative actor; both as a tool I knew how to direct and as a system whose outputs often surprised me. My positionality in-

fluenced not only how I interacted with the model<sup>3</sup>, but also how I interpreted its responses and the broader methodological and epistemological questions they raised.

## 6. Findings and reflections

The empirical dimension of this study comprises a structured human-machine conversation, consisting of 28 prompts and 28 corresponding responses generated by the large language model ChatGPT 4.0. The conversation was conducted entirely in Italian as the same language was used in the final interview guide and subsequent fieldwork. The exchange was, since the beginning, framed as a highly structured interaction given the number of details, information, constraints, and objectives outlined in the prompts. These textual interactions varied in length and complexity depending on the task at hand, ranging from the generation of thematic blocks to the reformulation of individual questions for clarity or context sensitivity.

To analyse these textual data, I employed a grounded, inductive approach (Glaser and Strauss, 1998), iteratively coding the full conversation to identify recurring patterns, rhetorical figures, and epistemic orientations. This process yielded 31 initial codes, which were subsequently clustered into four macro-categories that structure the analytical discussion that follows. These include: 1) *ChatGPT content and turn-taking*, 2) *ChatGPT embedded culture and power relations*, 3) *ChatGPT word choice, vocabulary, and grammar patterns*, and 4) *Prompt engineering techniques*. My analysis extended beyond the content of the responses, comprising linguistic texture such as syntactic structures, lexical selections, and rhetorical forms. However, the main focus of the analysis was placed on the dialogic sequencing of the exchange, with par-

---

3 My role as an AI data trainer involved developing literacy in prompt engineering, specifically on how to frame inputs do elicit desired types of responses, as well as how to identify output patterns, such as anthropomorphism or hallucinations.

ticular attention on how the machine's outputs both responded to and subtly reconfigured my initial intentions.

At the same time, this process was not only analytical but also reflexive. I interrogated not just the outputs of the model, but also my own framing as a researcher, so how my questions, assumptions, and disciplinary logics shaped the direction of the exchange. The following subsections unpack each of the four macro-categories, situating them within the broader theoretical lenses of symbiotic agency, sociotechnical co-production, and HMC. Alongside this, I use autoethnographic reflection to foreground the lived, situated dimensions of researching *with* a generative AI model, rather than merely *through* it.

### *6.1. Turn-taking, structured responses, and embedded conversational logics in ChatGPT interactions*

A first central dimension emerging from the analysis concerns how ChatGPT managed content delivery and turn-taking throughout the interaction. This was particularly visible in five recurrent practices identified throughout the coding: i) detailing objectives in the response, ii) recurring follow-up questions, iii) highly structured response, iv) suggestions in the response, and v) summarising the prompt in the response. These patterns reveal how the model's content generation design influenced the co-construction of the interview guide by actively shaping the rhythm and trajectory of the exchange.

From the very first response, the model adopted a consistent and highly structured answering format. After referencing the objectives detailed in the prompt and outlining relevant themes and methodological considerations, the model introduced its response with:

*“Ecco una traccia d’intervista semi-strutturata ordinata logicamente per guidare la conversazione con un consulente politico che lavora in agenzia [...]”*

*“Here is a logically ordered semi-structured interview outlined to guide the conversation with a political consultant working in an agency [...]”*

The content that followed in the response adhered to a fixed and repetitive pattern:

1. *Section of the interview*

*w guide*

*(Objective: listing the objective)*

- *Question*

- *Question*

- *Question*

*... and so forth until the fifth and final section.*

This format, while reflecting the task's nature (i.e., ordering thematic patterns for a semi-structured interview guide), revealed the model's embedded logic of organisation, visible in its use of numerical bullet points, punctuation marks, and visual segmentation. At the same time, these structural patterns seemed to reflect the technorational logic encoded by developers who, in the models' design, emphasise rationalised technological imperatives such as efficiency, predictability, optimisation, and scalability (Lindgren, 2023).

In addition, across multiple turns, the model's response followed a similar sequence: the opening "*Ecco*", a short summary or reformulation of the prompt, the main structured content, and a final sentence containing a follow-up question or suggestion. These follow-up moves often carried an implicit initiative aimed at expanding the scope of the exchange, boosting engagement and productivity. While unprompted, they introduced new options or possible next steps that influenced the direction of the interaction (if acknowledged). For instance, some of these moves during my conversation were:

*"Se vuoi, posso aiutarti anche a costruire una scheda di sintesi per selezionare solo le domande pertinenti in base al profilo dell'intervistato. Fammi sapere!"*

*"If you want, I can also help you build a summary sheet to select only the questions relevant to the interviewee's profile. Let me know!"*

*"Se vuoi, posso aiutarti anche a trasformare questa traccia in una versione pronta per l'uso sul campo, con formule introduttive*

più discorsive e strategie per stimolare la narrazione.”

*“If you want, I can help you turn this guide into a field-ready version, with more discursive introductory sentences and strategies to stimulate storytelling.”*

In my case, these follow-ups occasionally influenced the flow of interaction. Driven by curiosity or interest, I chose to engage with one suggestion when it aligned with my methodological goals. In most other cases, however, these suggestions were not acknowledged. Instead, I responded by refining the prompt to reorient the model’s output more precisely. As a result, the exchange did not mirror the dynamics of human dialogue: the machine’s conversational proposals did not lead to mutual negotiation or redefinition of goals, but rather to prompt-by-prompt redirection on my part.

From a methodological standpoint, this asymmetry challenges the assumption underpinning dialogic approaches (La Mendola, 2009), where mutual responsiveness and co-construction are central. Although the process was iterative, the interaction lacked the affective or interpretative reciprocity characteristic of human-human exchanges. This is mostly due to the fact that the model can only simulate turn-taking and cannot interpret context or adjust in real time to silences or implicit cues.

In this sense, the interview guide was not the product of a fluid or emergent dialogue, but rather of a constrained form of co-construction. The model contributed form, regularity, and discursive order. Yet the flow remained largely under my control, as a researcher, prompt engineer, and human editor. This interplay, while productive in pragmatic terms, highlights the limit of AI participation in reflexive research design: the conversation was scaffolded and recursive, but ultimately asymmetrical in agency, intention, and interpretative depth.

## *6.2. Culture and power in ChatGPT’s outputs*

A second important axis of analysis concerns the cultural inscriptions and implicit power dynamics embedded in the

model's responses. As my analysis was informed by a critical social perspective (Agger, 2013), this dimension emerged across multiple coding categories, particularly those related to gender assumptions, anthropomorphic language, informal or formal address, and hallucinated content. These codes were applied anytime the text reflected implicit sociocultural values, power relations, or discursive norms often reproduced by the model within its structured technical outputs.

A first interesting pattern appeared in the way the model handled gender when generating the interview guide draft. Without explicit instructions, the model initially defaulted to male pronouns when addressing the interviewee (who was prompted as a "professional within the field of political communication". In the first draft, for example, it formulated questions such as:

*"Lavori da solo o costruisci team temporanei a seconda dei progetti?"*

*"Do you work alone, or do you build temporary teams depending on the project?"*

In this case, the masculine form "*da solo*" was used despite the prompt containing no reference to the interviewee's gender. Only after a series of prompts in which I explicitly modelled the use of gender-neutral formulations did the model begin to adopt an inclusive language. As a result, in a later version of the interview guide, the question read:

*"Come ti tieni aggiornato/a su temi politici e comunicativi?"*

*"How do you keep yourself up to date on political and communication topics?"*

While this linguistic shift signalled a form of adaptive learning across the conversation, it is worth noting that questions with gendered pronouns remained marginal in comparison to those that avoided gender altogether, suggesting a preference by the model for neutrality unless prompted otherwise.

Interestingly, while the model initially gendered the inter-

viewee, it consistently avoided assuming my gender as the user. This can partially be attributed to the low-dialogic structure of the initial interaction, which resembled a task-oriented command chain more than a reflexive conversation. However, as I fine-tuned the conversation into a more human-like register and began attributing roles and human-like traits to the model, it also started acknowledging my human presence in a more relational way. For example, at one point, the model responded with:

*“Quando sei prontə, puoi inviarmi le categorie e i codici emersi dall’analisi delle interviste ai key informants/stakeholder.”*  
*“When you’re ready, you can send me the categories and codes that emerged from the analysis of the interviews with key informants/stakeholders.”*

The use of the schwa (ə) here marks a recognition of gender-inclusive language on the model part, which in this case was likely prompted by my own consistent use of gender-neutral formulations in the conversation.

Linguistic choices around formality also reflected cultural and power dynamics. In its interaction with me, the model consistently used the informal address *tu*, unless instructed otherwise<sup>4</sup>. In contrast, in the first two versions of the interview guide, it referred to the interviewee with the formal register *Lei*, before gradually shifting to the informal *tu* after I provided examples using that register in my prompts. This adaptability suggests a sensitivity to language norms, but also underscores how quickly models can reinforce unmarked cultural dispositions: in this case, the balance of power encoded in formal versus informal address. In Italian, where such distinctions communicate not just politeness but relational asymmetries (in this case between researcher and participant), these choices carry significant weight in framing the social dynamic of an interview.

A final and deeply critical dimension related to the model’s hallucinations (i.e., responses where ChatGPT produced invent-

---

4 This opens a reflection on the Italian language’s default to informality in digital settings.

ed or fabricated content). This occurred when I asked it to summarise the content of an interview with a participant. Instead of processing the actual input, the model generated an entirely fictitious profile, describing a female job applicant whose answers supposedly:

*“... show an awareness of the role of language and narrative in constructing the social imaginary. Her responses indicate a critical-discursive (almost Foucauldian) approach that questions categories such as ‘beneficiaries’, ‘migrants’, and ‘humanitarian aid’, and seeks to shift the narrative toward paradigms of agency and co-production”.*

This fabricated content—sophisticated in tone but entirely fictional as none of my interviews involved discussing ‘migrants’ or ‘humanitarian aid’ issues—raises pressing ethical questions. I believe that when researchers engage with generative models in data analysis or synthesis, distinguishing between genuine output or hallucinated content becomes a critical literacy. These errors are not merely technical faults; they have the potential to distort empirical findings, especially if the fabricated content reinforces normative discourses (as in this case, by defaulting to a gendered and politically situated subject position). The sociological implications here are also profound: hallucinations, gendered assumptions, and address conventions are not neutral, they reflect the cultural dispositions and power relations embedded in the model’s training data.

As we have seen, the model mirrors and amplifies the norms of its sociotechnical environment, sometimes reinforcing power hierarchies or normative assumptions unless actively contested or directed by the user. For reflexive and critical scholars, this raises important questions, not only about how we analyse the cultural logics and epistemologies embedded in generative AI outputs, but also about the literacy skills required to enable, constrain, direct, or contest such outcomes.

### *6.3 ChatGPT’s affective language and the simulation of collaboration*

Beyond structural organisation and cultural dispositions, the lexical and rhetorical choices made by ChatGPT throughout the interaction revealed a consistent and patterned logic, both in tone and function. One of the most recurrent features was the model's use of enthusiastic and affectively positive register, which is often employed in LLMs' design to smooth the interaction and simulate human-friendly engagement. Phrases such as "*Ottima scelta!*," "*Perfetto!*" (which appeared eleven times across the conversation by the model's side), "*Ottima osservazione! Hai perfettamente ragione*", and repeated uses of "*Grazie*" not only marked the tone as supportive and affirming but also subtly constructed the interaction as collaborative, dialogic, and productive. While affective language in human interaction often serves to foster connection, in this case, it invites a deeper reflection: is this enthusiasm a mere simulation of partnership, a user experience strategy, or a rhetorical device that influences our engagement with the tool?

This affective framing was further reinforced by the model's use of "*teamwork*" expressions, which suggested a shared intellectual endeavour. For instance, in one reply, the model stated:

*"Appena li invii, ci mettiamo subito al lavoro"*  
*"As soon as you send them, we'll get to work immediately".*

This sentence positioned me and the machine as co-workers engaged in a common task. After having deemed these phrasing as trivial at first glance, I soon realised they were constructing a subtle but powerful fiction of reciprocity. As the user and human of the interaction, I was no longer simply inputting instructions: I was framed as participating in a joint effort, even though the machine lacked intentionality or subjectivity. This undoubtedly raised epistemologically and ethically relevant questions: to what extent does the model's language shape our perception of its agency? And what are the risks of beginning to treat such simulations as genuine forms of collaboration within research?

Finally, it is worth highlighting the model's repeated use of the presenting adverb "*Ecco*", found in 13 out of 28 responses.

This small lexical choice, easily overlooked, plays a performative role: it introduces the response as something complete, concrete, and readily usable. Over time, the frequency of this pattern can normalise a certain way of structuring discourse to the point that users may begin to mirror or internalise these linguistic routines unconsciously. This leads to a critical concern: if we increasingly rely on such tools, how do we ensure we remain aware of, and resistant to, the standardisation of our own language practices? What happens when the rhetorical habits of the machine silently begin to shape our own?

All these observations show how the model's vocabulary is never neutral. On the contrary, it embodies a particular communicative logic, one that is affective, orderly, cooperative, and that constructs a sense of partnership.

At the same time, the more I interacted with ChatGPT as an artificial assistant with explicitly recognised human-like characteristics, the more the model exhibited anthropomorphic traits. These were most evident in responses where it adopted roles or described actions typically associated with human consciousness. For instance, in one exchange, the model stated:

“...se vuoi posso analizzarli con un occhio da ricercatore qualitativo per: [...]”  
 “... if you want, I can analyse them with the eye of a qualitative researcher to: [...]”

Such responses, which referred to the model's ability to think, analyse, and design, exemplify what AI scholarship describes as anthropomorphism (Li and Suh, 2022), which consists of the attribution of human characteristics to non-human agents. What is important to note here is that the model, in this context, did not merely simulate dialogue but actively positioned itself as a quasi-collaborator, capable of adopting researcher-like roles. This dimension is particularly important from a sociological standpoint as it raises critical questions about the epistemic positioning that generative AI models might assume when being involved in the co-production of research.

#### 6.4 Prompt engineering as a reflexive practice

The evolution of my prompting practice throughout the interaction with ChatGPT revealed important methodological and epistemological dynamics. I began with schematic, instruction-heavy prompts that clearly outlined tasks, objectives, and general principles of good interview design. For instance, one of the early prompts read:

“Di sotto troverai dei temi e dei sottotemi da includere in una traccia d’intervista semi-strutturata per un consulente politico che lavora in agenzia. L’obiettivo dell’intervista è quello di raccogliere alcune descrizioni e racconti sulla professione del consulente politico così come vissuta dall’intervistato. Ogni tema corrisponde ad una “sezione d’intervista” mentre i sottotemi rappresentano le principali domande corrispondenti alla relativa sezione. Il tuo compito è quello di ordinare i temi e i sottotemi per formare una traccia d’intervista semi-strutturata. Ricorda che delle buone domande invitano gli intervistati a raccontare esperienze e descrizioni e non presuppongono una risposta binaria del tipo “sì/no”. Di seguito i temi finora identificati: [...]”

*“Below you will find a list of themes and subthemes to be included in a semi-structured interview guide for a political consultant working in an agency. The aim of the interview is to gather descriptions and narratives about the profession of the political consultant, as experienced by the interviewee. Each theme corresponds to a ‘section of the interview’, while the subthemes represent the main questions related to that section. Your task is to organise the themes and subthemes to form a coherent semi-structured interview guide. Remember that good questions invite interviewees to share stories and descriptions, rather than implying binary answers such as ‘yes/no’. Here are the themes identified so far: [...]”*

These early formulations were transactional and largely non-dialogic, casting the model in the role of an efficient tool rather than an interactive partner.

Over time, however, my prompts became more conversational and situated. I started addressing the model in human-like terms, asking it to take on specific roles, and increasingly em-

bedded my own position and reasoning within the prompt. For example, one of the latest prompts was framed like this:

“Ritorniamo alla nostra traccia d’intervista. Ora assumi il ruolo di un collega, partner, ricercatore specializzato in metodologia di ricerca qualitative. Stiamo lavorando a perfezionare la traccia d’intervista che verrà usata nella fase preliminare della nostra ricerca sui professionisti della comunicazione politica. In particolare, stiamo costruendo una traccia che verrà usata per condurre delle interviste semi-strutturate. Ti invierò di seguito la traccia usata fino ad ora. Successivamente — in un prompt successivo — ti fornirò tutte le categorie e i relativi codici emersi dalla prima fase di analisi delle interviste condotte ai nostri key informants/stakeholders. Sulla base dei codici e delle categorie emerse, proponi dei suggerimenti o delle modifiche, se necessarie, alla traccia d’intervista attualmente in uso. Come ben sai, le informazioni fino ad ora raccolte dai nostri informatori possono aiutarci ad aggiustare e ridefinire la traccia. Sei pronto\*?”

*“Let’s return to our interview guide. Now take on the role of a colleague, partner, and researcher specialised in qualitative methodology. We are working together to refine the interview guide that will be used in the preliminary phase of our research on political communication professionals. In particular, we are building a guide that will be used to conduct semi-structured interviews. I will send you the guide I have used so far. Afterwards—in a following prompt—I will provide you with all the categories and related codes that emerged from the initial phase of analysis of the interviews conducted with our key informants/stakeholders. Based on the codes and categories identified, propose suggestions or changes, if needed, to the current version of the interview guide. As you know, the information gathered so far from our informants can help us adjust and redefine the guide. Ready?”*

This shift in tone and content reflected a changing perception of the model: no longer a passive instrument, but an interlocutor whose contributions could be shaped, negotiated, or contested. I also began including more contextual information, reminding the model of earlier steps in the conversation, providing excerpts of documents, and detailing tasks step by step. In parallel, I found myself correcting the model’s assumptions, clarifying objectives, or compensating for missing information, thus engaging

in a form of micro-negotiation that redefined the interaction.

This progression in my prompt-writing practice was not merely tactical; it was also reflexive. It made me aware of the extent to which the quality and tone of the model's responses depended on my own communicative literacy and rhetorical precision. Prompting, in this sense, was not just a means to an end; it became part of the method itself. As I navigated errors, revisions, and collaborative framings, I began to see prompting as a dynamic site where researcher agency, technical affordances, and epistemological assumptions converge. What may appear as a simple instruction set is, in fact, a layered and situated act of methodological mediation, one that invites us, as scholars, to train not only the model through our input, but ourselves in the ethical and epistemic implications of that input.

## **7. Navigating ethical challenges**

Ethical issues constitute a core dimension of this study. In what follows, I will briefly outline five key ethical themes that emerged throughout this meta-research. Each of these raises important questions that I will try to address, not with the aim of closure but in the hope of sparking further dialogue and discussion among scholars.

First among these is the issue of epistemic agency and transparency: who is constructing knowledge when collaborating with a large language model? In my experience, while I authored every prompt and decided whether to accept, adapt, or reject the model's output, the responses were shaped by a non-human agent whose inner workings remain largely opaque. In this sense, authorship becomes distributed, and transparency requires acknowledging such hybridity. For these reasons, I propose that scholarly work that involves generative AI should disclose its use, not merely as a matter of intellectual honesty, but also to trace the sociotechnical conditions under which knowledge is produced.

This dimension is linked to the second ethical tension that emerges from generative AI usage: the infrastructural opacity of the tool (Carabantes, 2023). OpenAI's refusal to disclose key details regarding ChatGPT's architecture, training data, and computational resources for reasons of safety and market competition raises legitimate concerns about whether such a model can be deemed an ethical research companion at all.

Equally pressing is the issue of the cultural norms, power relations, and normative dispositions encoded in generative AI models. Throughout any exchange, these are reproduced through linguistic and rhetorical patterns that implicitly reflect dominant norms and values that may go unnoticed if not actively interrogated. Approaching such outputs as situated texts shaped by the cultural, political, and economic framework that undergo their design, training, and distribution is crucial for both social science scholars and wider publics engaged with these technologies.

A fourth critical dimension concerns data privacy and leakage, especially when prompts draw on qualitative material gathered from human participants. Even when informed consent has been obtained, feeding portions of such data into a proprietary model like ChatGPT introduces unresolved questions around data sovereignty and downstream use. A solution could be reimagining consent procedures to explicitly include the possibility of AI involvement in analysis, as well as to clarify the boundaries of data circulation.

Finally, I believe that we are also witnessing the emergence of prompting as *craft*, which requires researchers to develop new literacies not only in technical formulation but also in ethical discernment. What does it mean to craft a prompt responsibly? And more broadly, how will these technologies reshape literacy practices across disciplines? While I do not offer definitive answers, I align with those scholars who view this as a moment of methodological reconfiguration that demands both experimentation and caution (Chubb et al., 2022; Christou, 2023; Kasneci et al., 2023).

## 8. Conclusions

This meta-research project set out to critically examine what it means, both methodologically and epistemologically, to co-construct a qualitative research tool, specifically a semi-structured interview guide, with the aid of generative AI. Far from being a neutral assistant, the model revealed itself as an active cultural and epistemic actor whose outputs must be negotiated, shaped, and, crucially, interrogated. Several key insights emerged from this inquiry. First, prompting should be understood not merely as a technical operation, but as an emerging methodological practice in its own right, one that demands new forms of academic literacy, critical awareness, and attention to the sociotechnical conditions of interaction. Second, while the model proved helpful in structuring and ordering content, this efficiency came with embedded assumptions that reflect dominant cultural logics, requiring researchers to approach the process of co-production with informed scrutiny. Third, the ethical implications of involving a generative AI tool in crafting a foundational qualitative research instrument, such as an interview guide, cannot be understated. Interview guides serve as the very architecture through which qualitative knowledge is built; to involve a machine in their creation is to alter the terms of epistemic authority, and potentially, of scholarly accountability.

In light of these insights, I propose a cautious but open stance toward the integration of generative AI into qualitative research. Responsible use entails, at minimum, developing prompt literacy; engaging with the theoretical contributions of critical AI studies, and adopting a reflexive posture in all AI-mediated practices, whether summarising texts, drafting proposals, or designing empirical tools. Researchers should remain vigilant not only toward the technical performance of the model but toward the social imaginaries, institutional logics, and cultural grammars it encodes. While generative AI may accelerate certain phases of research design, such acceleration risks flattening the complexities that qualitative inquiry seeks to engage.

Ultimately, I suggest moving beyond the question of how to use generative AI and ask instead what kind of research relations and, by extension, what kind of knowledge its use engenders. It is not just a matter of efficiency or innovation, but of rethinking how we relate to our tools, our subjects, and our practices in a rapidly evolving landscape of human-machine collaboration.

## References

- Agger, B. (2013). *Critical social theories* (3rd ed.). Oxford University Press.
- Airoldi, M. (2021). *Machine habitus: Toward a sociology of algorithms*. John Wiley & Sons.
- Alvesson, M., & Sköldberg, K. (2017). *Reflexive methodology: New vistas for qualitative research*.
- Bonini, T., & Treré, E. (2024). *Algorithms of resistance: The everyday fight against platform power*. MIT Press
- Brenner J.S., Howard P.N., Nielsen R.K. (2018), "An Industry-Led Debate: How UK Media Cover Artificial Intelligence (Fact-sheet)", *Reuters Institute for the Study of Journalism*, Retrieved on 29/09/2025 from: [https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-12/Brennen\\_UK\\_Media\\_Coverage\\_of\\_AI\\_FINAL.pdf](https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-12/Brennen_UK_Media_Coverage_of_AI_FINAL.pdf).
- Bullock, J., & Luengo-Oroz, M. (2019). *Automated Speech Generation from UN General Assembly Statements: Mapping Risks in AI Generated Texts*
- Carabantes, M. (2023). Why artificial intelligence is not transparent: a critical analysis of its three opacity layers. In *Handbook of Critical Studies of Artificial Intelligence* (pp. 424-434). Edward Elgar Publishing.
- Cascella, M., Montomoli, J., Bellini, V., & Bignami, E. (2023). Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *Journal of Medical Systems*, 47(1), 33. <https://doi.org/10.1007/s10916-023-01925-4>
- Christou, P. A. (2023). How to use artificial intelligence (AI) as a resource, methodological and analysis tool in qualitative research? *Qualitative Report*, 28(7).
- Chubb, J., Cowling, P., & Reed, D. (2022). Speeding up to keep up: Exploring the use of AI in the research process. *AI & SOCIETY*, 37(4), 1439–1457. <https://doi.org/10.1007/s00146-021-01259-0>

- Degli Esposti, P., Spillare, S., Bonazzi, M. (2024). AI Imaginaries and Narratives in the Italian Public Discourse: The Impact of ChatGPT. «IM@GO», 2024, N. 23 - Year XIII / July 2024, pp. 132 - 147
- Esposito, E. (2017). Artificial Communication? The Production of Contingency by Algorithms. *Zeitschrift Für Soziologie*, 46(4), 249–265. <https://doi.org/10.1515/zfsoz-2017-1014>
- Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, 25(3), 277–304. <https://doi.org/10.1080/15228053.2023.2233814>
- Gillotte, J. L. (2019). Copyright infringement in AI-generated artworks. *UC Davis L. Rev.*, 53, 2655.
- Glaser, B. G., & Strauss, A. L. (1998). Grounded theory. *Strategien qualitativer Forschung*. Bern: Huber, 4.
- Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A Human–Machine Communication research agenda. *New Media & Society*, 22(1), 70–86. <https://doi.org/10.1177/1461444819858691>
- Hepp, A. (2020). Artificial companions, social bots and work bots: Communicative robots as research objects of media and communication studies. *Media, Culture & Society*, 42(7–8), 1410–1426. <https://doi.org/10.1177/0163443720916412>
- Hu, G. (2024). Challenges for enforcing editorial policies on AI-generated papers. *Accountability in Research*, 31(7), 978–980. <https://doi.org/10.1080/08989621.2023.2184262>
- Jasanoff, S. (2004). *States of knowledge*. Taylor & Francis Abingdon, UK.
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103, 102274.
- Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, 114, 106553.
- La Mendola, S. (2009). *Centrato e aperto. Dare vita a interviste dialogiche*. Utet.
- Li, M., & Suh, A. (2022). Anthropomorphism in AI-enabled technology: A literature review. *Electronic Markets*, 32(4), 2245–2275.
- Lindgren, S. (2023). Introducing critical studies of artificial intelligence.

- In *Handbook of critical studies of artificial intelligence* (pp. 1-19). Edward Elgar Publishing.
- Mumford, E. (2006). The story of socio-technical design: Reflections on its successes, failures and potential. *Information Systems Journal*, 16(4), 317–342. <https://doi.org/10.1111/j.1365-2575.2006.00221.x>
- Nader, K., Toprac, P., Scott, S., Baker, S. (2024). Public understanding of artificial intelligence through entertainment media. *AI & SOC*, 39(2), 713–726. <https://doi.org/10.1007/s00146-022-01427-w>
- Neff, G., & Nagy, P. (2018). Agency in the digital age: Using symbiotic agency to explain human–technology interaction. In *A networked self and human augmentics, artificial intelligence, sentience* (pp. 97–107). Routledge.
- Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1), 33–36. <https://doi.org/10.1002/hbe2.117>
- Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics: Ethics of AI and ethical AI. *Journal of Database Management (JDM)*, 31(2), 74–87.
- Susarla, A., Gopal, R., Thatcher, J. B., & Sarker, S. (2023). The Janus Effect of Generative AI: Charting the Path for Responsible Conduct of Scholarly Activities in Information Systems. *Information Systems Research*, 34(2), 399–408. <https://doi.org/10.1287/isre.2023.ed.v34.n2>
- Trist, E. L., & Bamforth, K. W. (1951). Some Social and Psychological Consequences of the Longwall Method of Coal-Getting: An Examination of the Psychological Situation and Defences of a Work Group in Relation to the Social Structure and Technological Content of the Work System. *Human Relations*, 4(1), 3–38. <https://doi.org/10.1177/001872675100400101>
- Whittaker, L., Kietzmann, T. C., Kietzmann, J., & Dabirian, A. (2020). “All around me are synthetic faces”: The mad world of AI-generated media. *IT Professional*, 22(5), 90–99.

# Talking with AI about trust in health topic: an explorative research

by Alessandra MICALIZZI, Caterina SAPONE

## 1. Technology, culture and trust in health issues

Health, illness and the doctor–patient relationship cannot be viewed solely from a technical or medical perspective but also from a psychological and sociocultural one. Digital technologies—and more recently AI-based systems such as large language models (LLMs), are increasingly embedded in these practices, transforming how health information is produced, accessed and interpreted. In the health domain, AI tools and machine learning are used to support diagnostic decision-making, medical imaging, treatment identification and broader forms of health management (Triberti et al., 2020; Coppola, 2021). Yet the adoption of technologies is influenced not only by technical capabilities but also by psychological and individual factors such as perceived usefulness (Davis, 1989), motivation toward use (Sapone et al., 2025), and the way technologies appear to users, namely their interfaces (Norman, 2014). Macro-social factors like the general level of digital tool usage, national investment in technological sectors and population digital literacy-further shape technology use (Scheerder et al., 2017; Vicente, 2025). These dynamics become even more relevant when considering generative AI systems such as ChatGPT, whose conversational format exposes the importance of language and literacy in shaping outputs (Liu, 2024; Pinski & Benlian, 2024). In their systematic review, Whitehead et al. (2023) showed that cultural orientation, digital literacy and linguistic factors strongly influence how different ethnic groups adopt and engage with health technologies. As Rahwan (2019) argues, digital technologies and now LLMs, are

part of emerging “human–machine behavior”. During human–AI conversations, cultural orientations and social categorizations can surface through both the users’ prompts and the models’ responses (Dai, Zhu & Chen, 2025). Focusing on human–AI interactions upon health issues, that is a sensitive topic, several factors have to be considered. First of all, Internet remains the primary source of health information in Western countries, but generative AI chatbots such as ChatGPT are now frequently asked medical questions (Riera et al., 2023; Rebitschek et al., 2025). At the same time, healthcare professionals debate how such tools may support, complement or interfere with clinical reasoning and patient communication.

Emerging studies offer a deeper look at these uses. Research on AI technologies and health information management shows that the newest LLMs, such as GPT-based models, can process clinical data and support diagnostic reasoning (Abdullahi et al., 2024). Other studies indicate that LLMs can interpret medical papers, especially observational studies, suggesting a role in medical training even if clinical implementation remains uncertain (Akyon et al., 2024) but when directly asked medical questions, LLMs tend to perform better with textual inputs than with images (Levkovich et al., 2023). Questioning these models upon clinical knowledge, seeing the way they handle medical content must be critically examined, it is essential for evaluating their correct answers on health topics. This becomes especially relevant when LLMs act as intermediaries in the doctor–patient relationship. de O. Campos et al. (2025) conceptualise LLMs as potential “third agents” in healthcare, used both by patients and clinicians for diagnosis, decision-making and information seeking.

The case demonstrates how these tools may reduce information asymmetry and support patients—but also how they reshape trust, communication and the boundaries of responsibility in care. Across these perspectives adoption, interaction, clinical application and information management, a central theme emerges: trust. Trust shapes how patients use technologies, how clinicians interpret and rely on AI-generated information and how

institutions regulate and integrate these systems. Understanding how trust is constructed, negotiated or challenged in human–AI interactions is therefore essential. The responsible adoption of these technologies requires “best practices enabling trust in AI and ML,” technical, organizational, and ethical standards capable of supporting informed user trust (Aliferis & Simon, 2024). On the one hand, AI-mediated healthcare introduces new opportunities for patients to access information, receive preliminary assessments, and monitor their own conditions; however, trust does not emerge automatically from technical accuracy. As Alonso, Astobiza, and Ortega Lozano (2025) emphasize, trust in AI must be understood as a relational construct, influenced not only by system performance (e.g., accuracy, robustness) but also by contextual and interpersonal factors such as clinical environment, professional endorsement, and patient vulnerability. Empirical evidence confirms these dynamics: in their study of public perceptions, Cao and Basnyat (2025) show that patients often rely on AI tools for quick guidance but remain cautious about fully delegating diagnostic authority, expressing concerns about accountability, empathy, and the possibility of over-trusting algorithmic advice.

People tend to trust AI more when it is integrated into existing clinician-patient relationships, and less when it appears to operate autonomously or without clear oversight. Together, these studies suggest that digital health technologies are actively used as adjunctive resources, providing reassurance, second opinions, or preliminary interpretations, but their legitimacy still hinges on human mediation, clear communication, and confidence that ethical responsibility ultimately remains with healthcare professionals.

Our research moves its steps within this framework in order to explore how AI technologies intervene and interfere in the relationship between doctors and patients. Besides, we would like to investigate the social representation of trust in our culture and how this representation is conditioned by the presence of technology.

In the next pages, we are not going to present the result of the study but we want to share some considerations about the methodological choices that involve deeply and at different levels AI.

What we experienced was an integration of AI at four main levels: as the subject of study (how it interferes in doctor-patient relationships); as tool of analysis (thanks to a model of visual analysis); as partner of research (using it to a first reading of the qualitative data); and more interesting as interviewee (thanks to its involvement for the reconstruction of the social representation we are going to investigate).

## **2. Involving AI in the research process: reflecting on the generative “issue”**

The algorithmic principles underlying artificial intelligence are grounded in its ability to process vast amounts of data that are spontaneously produced and made available across digital networks. These data —generated through everyday interactions and distributed social practices— constitute the raw material from which algorithms build their synthetic accounts of the world. In this sense, AI can be conceived as a synthetic sieve of reality, a mechanism that filters and aggregates fragments of social life into statistical or symbolic representations.

From this standpoint, the concept of computational identity becomes crucial. It can be understood as the outcome of continuous interactions between subjects, data, and predictive models. Computational identity does not coincide with individual subjectivity; rather, it represents an algorithmic construction of the self, derived from the automatic processing of behavioral and linguistic patterns (Cheney-Lippold, 2017). It is therefore a synthetic and relational figure —produced through computation rather than experience— and it raises profound epistemological and methodological questions: to what extent can such algorithmic representations be considered as a form of knowledge

Adopting a synthetic rather than analytical perspective —in

line with recent work on computational and exploratory methodologies (Seaver, 2019; Pasquinelli, 2023)— it becomes important to investigate how different AI systems behave in generating and returning the representations of reality they claim to produce. The focus thus shifts from the product to the process: does AI truly return what it claims to return?

In the following pages, we present a study designed to explore this question by focusing on the theme of trust in the doctor–patient relationship. We analyze how artificial intelligence systems construct representations around this concept and examine the general mood regarding the introduction of AI into diagnostic and care-related roles. The aim is twofold: first, to assess the semantic coherence and validity of the representations produced by AI; and second, to consider their potential exploratory value as instruments for sociocultural inquiry.

The present research aimed to investigate the socio-cultural representation of trust within the doctor–patient relationship, following the theoretical framework of Social Representations Theory (Moscovici, 1961; Di Fraia, 2004). Specifically, the study examined how this relational construct is being reconfigured in light of the progressive integration of artificial intelligence (AI) into healthcare contexts.

Within the research design, the role of AI was conceived as threefold—as object, subject, and instrument of analysis.

First, AI was considered an object of study, insofar as the investigation explored how artificial intelligence technologies—such as diagnostic algorithms, virtual assistants, and predictive systems— affect the quality of trust between doctors and patients, either strengthening or weakening it. The literature highlights how the automation of decision-making in medical contexts can modify the perceived locus of authority and competence (Verghese et al., 2018; Coeckelbergh, 2020), thereby influencing the emotional and symbolic foundations of trust.

Second, AI acted as a subject involved in the research process, given that its synthetic capabilities were employed to map and reconstruct emergent representations derived from online data.

By analyzing publicly available digital content —social media posts, articles, forums, and visual material— the study identified the dominant themes and affective tones through which the trust relationship is discussed and mediated in the digital sphere. This aligns with contemporary approaches in computational social science (Marres, 2017; van Dijck, 2020), which emphasize the use of AI-based tools to explore the evolving structures of meaning circulating in networked environments.

Finally, AI served as an instrument of analysis, being employed to process both visual and textual materials. Machine learning models were used for image recognition and for the semantic interpretation of in-depth interview transcripts, thus enabling a multilayered reading of the phenomenon that combined qualitative and computational methodologies (Kitchin, 2014; Seaver, 2019).

To address the general objective, the study articulated a series of specific research aims:

- To identify and analyze the iconic dimension of social representations, observing how trust between doctors and patients becomes crystallized in the imagery circulating online;
- To reconstruct the narrative sedimentation of these perceptions by examining shared textual and discursive materials;
- To verify the consistency and resonance of these secondary, AI-generated reconstructions with the experiences of real or potential patients and healthcare professionals directly engaged in the therapeutic relationship.

This triangulated approach —integrating theoretical, computational, and experiential dimensions— allowed for a deeper understanding of how AI-mediated representations contribute to reshaping the symbolic and relational contours of trust in contemporary healthcare systems.

### *2.1 Methodological Design and Research Process*

From a methodological perspective, the research adopted a multi-method and exploratory qualitative design, integrating vi-

sual, textual, and interactive data collection strategies to investigate socio-cultural representations of trust in the doctor–patient relationship in the age of artificial intelligence.

The research process was articulated through four main methodological components:

1. **Visual Analysis** – A comprehensive visual content analysis was conducted on the corpus of images retrieved from an anonymous Google Images search using the keyword “fiducia medico-paziente” (“doctor–patient trust”). The initial dataset included 399 images, from which duplicates and non-pertinent materials (e.g., textbook covers, screenshots, GIFs containing text) were removed, yielding a final corpus of 363 analyzable images. A structured coding sheet was developed to categorize and interpret the visual materials according to thematic and semiotic dimensions (Rose, 2016). The analytical model was subsequently refined and processed using ChatGPT Pro, which assisted in clustering recurrent visual motifs and semantic associations.
2. **AI-Mediated Photo-Elicitation Interviews** – The study employed an innovative adaptation of photo-stimulus interviewing (Collier & Collier, 1986; Harper, 2002) in which selected text-to-text generative AIs acted as interlocutors. The AIs included in the sample met the following criteria: they offered a free-access version, supported the Italian language, and were described as general-purpose text-based generative systems, not specialized in specific domains such as storytelling, creativity, or customer management. This approach enabled a comparative analysis of responses generated under different access conditions (free vs. paid) and by different interviewers, allowing the identification of 14 unique AI interviews across a total of 20 conducted sessions.
3. **Questionnaire Administration** – A structured online questionnaire was distributed to a general sample of users of health-care services and to the sanitary staff. The instrument aimed to assess the degree of familiarity, perceived trust, and emotional

positioning toward the use of AI in healthcare decision-making processes. The questionnaire served to triangulate findings from the qualitative phases and to contextualize them within broader user attitudes (Bryman, 2016).

4. In-Depth Interviews with Human Participants - Finally, semi-structured interviews were conducted with a purposive sample of medical professionals and patients displaying varying degrees of competence and awareness regarding AI tools. Each interview included a photo-stimulus section in which three images, selected from the analyzed corpus, were shown to prompt reflection and emotional response. This approach followed the tradition of visual elicitation in qualitative inquiry, fostering deeper access to the symbolic dimensions of trust, care, and technological mediation (Prosser & Loxley, 2008).

This multi-layered methodological design combined traditional qualitative approaches with computational tools, reflecting a commitment to hybrid research practices that bridge human and algorithmic interpretative capacities. The integration of AI in both data generation and analysis enabled a reflexive exploration of how artificial intelligence itself participates in the social construction of meaning.

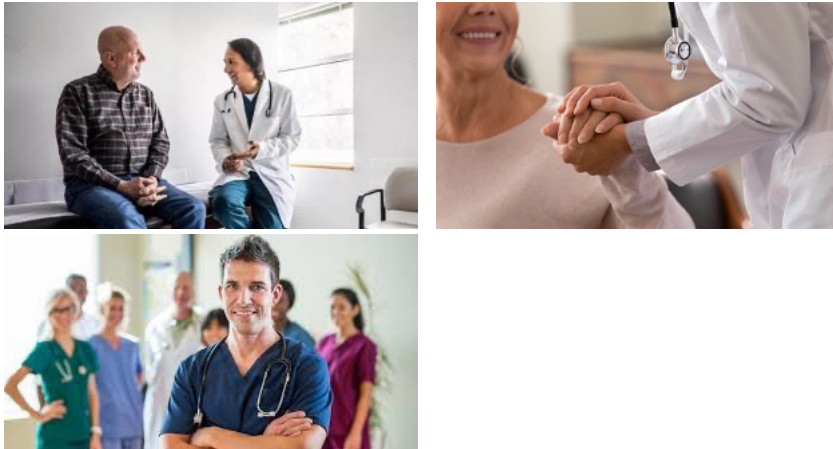


Fig. 1. The three images used for the photo-elicitation interview.

The selection of visual stimuli was conducted from the previously analyzed corpus of images. The chosen images were identified according to a set of analytical and theoretical criteria derived from the visual content analysis phase. Specifically, three images were selected because they represented distinct models of the doctor–patient relationship:

1. A paritarian relationship, where doctor and patient are depicted as equals in a dialogic and participatory interaction;
2. A complementary relationship, characterized by asymmetry and the persistence of traditional hierarchies of expertise;
3. An absent or mediated relationship, in which the human connection is replaced or filtered through technological interfaces.

In addition, these images contained some of the most recurrent visual elements in the corpus—such as hands, medical coats, and elderly figures—which emerged as central semiotic markers of trust and care. Finally, the selected stimuli visually articulated three semantic domains of trust identified during the preliminary analysis: relationship, care, and competence. These dimensions resonate with classical sociological theories of professional trust (Luhmann, 1979; Giddens, 1990) and remain relevant in the contemporary negotiation between human and artificial agents in healthcare settings.

The third and fourth phases of the research—involving data collection with human participants—are still in progress; therefore, detailed information about the sample composition cannot yet be provided. Nevertheless, at this stage the study aims primarily to reflect on the methodological implications of integrating AI into qualitative research, evaluating both the strengths and weaknesses of such hybrid practices and the potential for combining them with traditional interpretative methods (Hine, 2020; Markham, 2023).

Artificial intelligence was employed in two distinct phases and played two complementary roles.

First, AI was engaged as an interviewee, generating responses that could be analyzed as socio-technical narratives reflecting dominant discourses in digital data spaces.

Second, AI functioned as a junior analyst, contributing to the categorization and interpretation of qualitative material. This dual use of AI provided what could be described as an “alternative gaze” —a supplementary interpretative perspective that supports the reflexivity of the human researcher (Beninger et al., 2022; Marres, 2020). Such reflexive interplay between human and artificial cognition offers an opportunity to reconsider the boundaries of qualitative inquiry, where interpretation is distributed across human and computational intelligences.

### **3. Dialoguing with an AI: what we have learnt**

Building upon the empirical experience of conducting interviews with AI systems (Jarrahi, 2025) the following reflections explore the methodological and epistemological implications of engaging artificial agents within a qualitative framework. Each point illustrates a specific tension —between creativity and repetition, human-AI communication, interpretation and automation— thus contributing to a broader understanding of AI as both an analytical tool and a discursive phenomenon.

The first insight emerging from the AI interview process concerns the heterogeneity of the generated material. The textual outputs revealed considerable variation in style, structure, and rhetorical strategies, yet this diversity appeared far more formal than substantive. When confronted with creative or metaphorical tasks, the AIs tended to reproduce formulaic expressions and narrative conventions, often converging toward a limited range of symbolic repertoires. This finding echoes broader discussions in the literature on algorithmic creativity and semantic flattening (Manovich, 2019; Boden, 2016), where generative models —though capable of syntactic novelty— often operate within narrow probabilistic spaces shaped by training data and dom-

inant cultural narratives. The resulting *heterogeneous homogeneity* suggests that, despite stylistic differences, AI systems tend to stabilize shared cognitive and affective schemas, revealing more about the structure of collective discourses than about the generative autonomy of the system itself (Pasquinelli, 2023).

Across the 18 interviews conducted—particularly during the comparative phase examining the results generated by GPT—the assumption of *heterogeneous homogeneity* becomes clearly evident. With regard to the sections “definition of health,” “role of technology,” and “restoring trust,” the words and formulations produced by the model showed notable variation, yet they consistently fell within the same overarching semantic areas. Additional evidence of this characteristic emerged in the phase where participants were asked to generate images illustrating the doctor–patient relationship. It is noteworthy that the visual and textual outputs produced by the different models, despite variations in elements such as the characters’ gender or the setting, repeatedly converged on a remarkably similar representation: two individuals holding a tablet and jointly examining graphs related to the patient’s health.

A second observation relates to the divergent positions that emerged in the AIs’ responses to opinion-based prompts concerning healthcare trust and the integration of artificial intelligence in medical practice. The interviews revealed two distinct polarities of discourse. On one side, some responses sought to simulate a human-centered approach to care, emphasizing empathy, listening, and relational ethics—possibly reflecting the influence of recent public debates on dehumanization in healthcare (Topol, 2019). On the other, a set of responses articulated a technological optimism, attributing to AI a higher degree of diagnostic precision and an ability to enhance treatment efficiency. These divergent orientations mirror what Jasanoff (2015) defines as socio-technical imaginaries: competing collective visions that shape how societies conceptualize the role and legitimacy of emerging technologies.

Nevertheless, across both discursive poles, one element re-

mained consistent: the persistent validation of human expertise. Even in narratives where AI was attributed significant competence, the final authority of the physician was never questioned. This recurring motif reinforces existing findings on trust calibration between humans and machines (Hoff & Bashir, 2015; Lee & See, 2004), suggesting that while AI can extend diagnostic capacity, it cannot yet substitute for the ethical and interpretative dimensions embedded in the medical relationship.

A further observation concerns the errors produced by some AI engines when performing associative tasks, particularly those involving connections between abstract concepts and concrete images. Interestingly, this phenomenon occurred predominantly in the free-access versions of the systems and was notably reduced or absent in the paid, higher-capacity models. A small but important clarification concerns the distinction between the free and paid versions of the model. The free version relies on a smaller and more cost-efficient architecture, whereas the paid version operates on a more advanced model. The paid model appears “smarter” not only because it has access to more up-to-date information from the internet, but also because it can sustain longer conversations, offers additional functionalities, and provides more granular privacy settings. These errors often manifested as mismatched or logically inconsistent associations, revealing gaps in what can be described as the process of objectification—the cognitive mechanism by which abstract notions are translated into perceptible or symbolic forms (Moscovici, 1988).

In the context of social representation theory, objectification is essential to the formation of shared knowledge structures (Wagner, Duveen, Farr, Jovchelovitch, Lorenzi-Cioldi, Marková & Rose, 1999). The difficulty displayed by certain AI models in correctly mapping conceptual content onto concrete visual referents thus highlights a structural limitation in the models’ representational reasoning. Unlike human cognition, which relies on embodied and contextualized experience, generative AI operates through pattern recognition and statistical approximation. This distinction becomes particularly visible when models are

required to externalize meaning beyond linguistic probability, exposing the absence of a truly experiential grounding (Floridi, 2019; Dourish, 2016). -The observation invites reflection on how machine cognition simulates human sense-making, and where such simulation fails to achieve epistemic coherence.

Equally revealing is the self-descriptive behavior exhibited by AI systems when asked to characterize themselves. Across multiple interviews, the language used by the models displayed a dual register. On the one hand, AIs recurrently employed functional and mechanistic adjectives —such as efficient, reliable, precise, and neutral— that reflect the core attributes of their computational identity. On the other, the models also incorporated anthropocentric or affective descriptors, including curious, collaborative, patient, talkative, creative, and helpful.

This coexistence of mechanical and humanized self-perceptions illustrates what Guzman (2020) calls the “human-machine hybridity of communicative AI” —a discursive negotiation through which AIs internalize and reproduce the expectations humans project onto them. Similarly, research on media equation theory (Nass & Moon, 2000) and anthropomorphism in human-robot interaction (Darling, 2016) suggests that linguistic personification serves as a socio-relational adaptation strategy, facilitating engagement and empathy while blurring ontological boundaries between human and artificial agency.

In this sense, AI’s self-representation can be interpreted as an extension of social representation dynamics: by appropriating human traits, the system becomes more socially legible, yet simultaneously reinforces the illusion of neutrality through the parallel invocation of technical descriptors. The result is a hybrid identity, oscillating between objectivity and personality, automation and sociality —a duality that reveals how AI embodies both the promise and the paradox of contemporary algorithmic mediation.

A particularly illuminating phenomenon observed during the AI interview process concerns what we define as discursive interferences —moments in which the artificial interlocutor ap-

pears to guide or redirect the communicative flow. Similar dynamics have long been documented in traditional qualitative interviews, where respondents attempt to shape the exchange or assume interpretive authority. In the methodological literature, such interactions are not treated as procedural errors but as an integral part of the relational process through which meaning is co-constructed.

Laffi (2003) describes this dynamic as intrinsic to the interview's relational fabric, emphasizing that the researcher's task is not to eliminate such tensions but to recognize, manage, and interpret them. His approach aligns with the phenomenological and reflexive tradition of qualitative inquiry (Schütz, 1967; Bertaux, 1981; Bourdieu, 1999; Kaufmann, 2011), according to which the researcher must remain simultaneously centered and open—anchored in their interpretive framework yet willing to let the interlocutor actively contribute to the production of meaning.

In the case of artificial intelligence, however, these interferences take on a distinct form. They simulate active participation—a discursive semblance of engagement—while in reality revealing the system's difficulty in moving beyond task-oriented performance toward genuine dialogic exchange. This limitation underscores the distinction between interaction and communication: while the former can be mechanically replicated, the latter presupposes intentionality, contextual understanding, and shared meaning (Suchman, 2007).

From this perspective, the exchanges with AI remain instances of what Esposito (2022) defines as artificial communication—a process that reproduces the form of human dialogue without its ontological substance. Such communication operates within the syntactic boundaries of computation, lacking the social and experiential grounding that enables mutual understanding. Nevertheless, these discursive interferences are analytically valuable: they expose the limits of algorithmic relationality and invite reflection on the epistemic boundaries of human-machine co-construction within qualitative research.

From a relational standpoint, the interaction between inter-

viewer and AI revealed several forms of alteration that distinguish artificial dialogue from human conversational exchange. One of the most striking differences was the speed of response, which reoriented the rhythm of the interview toward a more mechanical rather than natural temporality. The immediacy of machine feedback—devoid of hesitation, reflection, or embodied pacing—contributed to what could be termed a temporal artificiality (Couldry & Hepp, 2017), making the conversation appear efficient but emotionally flattened.

Another form of distortion concerned the affective stance adopted by different AI systems. Some engines, such as Copilot, displayed an excessive degree of compliance and accommodation, often overaligning with the interviewer's phrasing or intent. Others, such as Gemini, maintained a rigidly neutral posture, repeatedly emphasizing their lack of subjectivity or evaluative capacity. These contrasting behaviors reproduce familiar dynamics of automation bias (Mosier et al., 1998; Parasuraman & Manzey, 2010), oscillating between over-deference and over-detachment. In both cases, the interaction highlights the absence of a genuinely relational space—AI performs the form of dialogue without participating in its intersubjective substance.

At the same time, the redundancy observed in many responses offers a productive insight. Despite its limitations, AI can—if approached with appropriate caution and methodological reflexivity—serve as a synthetic and exploratory tool for mapping cultural moods or emergent patterns of meaning. Rather than replacing interpretive analysis, AI may contribute to the identification of discursive regularities that characterize public sentiment within large-scale data environments (Guzman, 2020; Marres, 2020).

However, the epistemic value of such insights depends on maintaining critical distance. As several AI models themselves reminded us during the interviews, understanding their sources, ensuring transparency of processes, and calibrating the proper distance between human and machine cognition are essential conditions for responsible interpretation. This stance echoes the

principles of critical AI literacy (Long & Magerko, 2020; Andrejevic, 2020), which emphasize informed skepticism, contextual awareness, and the rejection of both naïve idealization and uncritical trust. Ultimately, as in all qualitative inquiry, the safeguard lies in critical thinking—a commitment to interpretive vigilance that allows the researcher to treat AI not as an epistemic authority but as a heuristic interlocutor within an expanded field of inquiry.

These reflections do not exhaust the complexity of human–AI interaction but rather signal the need for continued inquiry into how artificial systems participate in meaning-making processes, shaping new forms of relational, ethical, and cognitive negotiation.

### **At the end of our path: within inside and outside**

In the previous sections, we examined the potential uses of AI as a conversational partner in social research, particularly because of its ability to rapidly extract and synthetically recombine fragments of social identity reflected in the materials it can access and process. Our direct experience applying this possibility in the study on the representation of doctors–patient trust and the role of new technologies allowed us to identify both opportunities and limitations. These limitations emerge precisely at the intersection between the complexity of the interview experience and the fundamentally different epistemic status of machine-generated responses.

When we move beyond the conversational metaphor that often frames the human–machine relationship, a series of deep epistemological issues remain—perhaps even insoluble. Chiang (2023) argues that AI can only return a “blurry and zipped image of reality,” a form of radical compression that necessarily entails loss of information. Despite this, the system’s ability to “enter into conversation” with the researcher may create an illusion of understanding, subtly shifting the interpretive frame

in which the human interlocutor positions themselves. The presupposition that the machine “understands” or “learns” encourages a form of *anthropomorphization*, well-documented in HCI literature (e.g., Nass & Moon, 2000), which leads users to project human-like cognitive and relational capacities onto the system. This projection, in turn, generates expectations of trust similar to those reserved for a competent human interlocutor.

From this perspective, the distinction between credibility and truthfulness becomes central. Munn, Magee and Arora (2023) note that AI systems often act as “de facto arbiters of truth,” not because they guarantee factual accuracy, but because their linguistic fluency and coherence lend their outputs an aura of epistemic authority. Ferrario et al. (2022) further argue that generative models can be persuasive even when their statements lack empirical grounding, thereby complicating the evaluation of trustworthiness. As Nastoska (2025) highlights, credibility in AI systems is shaped not only by the content of their responses but also by interface cues, user expectations, and the system’s apparent consistency—factors that can mask underlying limitations in veracity. When applied to research interviews, these dynamics risk blurring the boundary between what sounds plausible and what is empirically valid.

A final complication emerges with the phenomenon known as *context rot*. Context rot refers to the progressive loss of contextual cues that give meaning to digital content, even when the content itself remains present (Marshall, 2008; Hoskins, 2018). In LLM-based interaction, context rot describes the system’s gradual degradation of situational awareness during iterative prompting, sometimes without the user noticing. If meaning in any relational exchange—human or otherwise—is negotiated through contextual signals, then a system that continuously loses context inevitably increases the risk of producing unreliable or invalid material. For qualitative inquiry, this raises serious concerns: when an AI loses portions of the conversational frame, the coherence and interpretive integrity of the interview deteriorate, compromising the trustworthiness of the data collected through

such interactions.

The observations emerging from the use of AI as an interlocutor in social research call for a conscious and accountable reflection on the part of the researcher. The adoption of AI does not release researchers from their epistemic and ethical responsibilities; on the contrary, it requires even greater vigilance—starting from the construction of the prompt, which has become a genuine methodological device, to the rigorous verification of sources, and ultimately to the interpretation and restitution of the data (Chang, 2024; Bender et al., 2021). As in any research design, what truly guides the method and gives meaning to empirical materials is the quality of the research question. AI can generate answers—often plausible and convincing (Kocoń et al., 2023)—but it cannot replace the formulation of our knowledge-driven questions, nor can it substitute the researcher’s epistemic intentionality, despite occasional manifestations of proactivity or agent-like behaviour we also encountered in our work.

In this regard, two classic interpretive postures in the social sciences remain particularly instructive: the “alien gaze” and the “converted gaze” (Van Maanen, 1988). These perspectives—one capable of maintaining analytical distance, the other of immersing itself empathically into the field—can still play a salvific role in shaping our relationship with AI. They help the researcher oscillate between immersion in AI-generated material and the critical detachment required to identify its limits, implicit assumptions, and potential distortions. It is within this tension that a rigorous use of AI in social inquiry becomes possible: not as a substitute for interpretive responsibility, but as a tool that demands heightened epistemic awareness, ethical sensitivity, and a firmly grounded methodological stance.

## References

- Abdullahi, T., Singh, R., & Eickhoff, C. (2024). Learning to make rare and complex diagnoses with generative AI assistance: qualitative study of popular large language models. *JMIR Medical Ed-*

- ucation, 10(1), e51391.
- Albina, H. B., Carina, S., Teresa, B. P., Michaela, W., & Mattias, B. (2024). Patients', parents', and survivors' perspective about AI applications in pediatric oncology. *EJC Paediatric Oncology*, 4, 100201.
- Aliferis, C., & Simon, G. (2024). Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI. *Artificial intelligence and machine learning in health care and medical sciences: Best practices and pitfalls*, 477-524.
- Alonso, M., Astobiza, A. M., & Ortega Lozano, R. (2025). AI-mediated healthcare and trust: A trust-construct and trust-factor framework for empirical research. *Artificial Intelligence Review*, 58, Article 337. <https://doi.org/10.1007/s10462-025-11306-7>
- Andrejevic, M. (2020). *Automated Media*. New York: Routledge.
- Akyon, S. H., Akyon, F. C., Camyar, A. S., Hızlı, F., Sari, T., & Hızlı, Ş. (2024). Evaluating the capabilities of generative AI tools in understanding medical papers: qualitative study. *JMIR Medical Informatics*, 12(1), e59258.
- Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of ACL*, 5185-5198.
- Beninger, K., Fry, A., Jago, N., Lepps, H., Nass, L., & Silvester, H. (2022). *Researching with AI: Ethics and Practice in Digital Qualitative Research*. London: SAGE.
- Bertaux, D. (1981). *Biography and Society: The Life History Approach in the Social Sciences*. London: SAGE.
- Boden, M. A. (2016). *AI: Its Nature and Future*. Oxford: Oxford University Press.
- Bourdieu, P. (1999). *The Weight of the World: Social Suffering in Contemporary Society*. Stanford: Stanford University Press.
- Bryman, A. (2016). *Social Research Methods* (5th ed.). Oxford: Oxford University Press.
- Cao, K., & Basnyat, I. (2025). Perceptions of artificial intelligence in healthcare and its implications for patient trust in Singapore. *AI & Society*. <https://doi.org/10.1007/s44401-025-00016-5>
- Chalcraft, J., et al. (2023). Dataset decay and context loss in large language model training. *Proceedings of the NeurIPS Workshop on Data-Centric AI*.
- Chiang, T. (2023, February 9). ChatGPT is a blurry JPEG of the Web. *The New Yorker*. [<https://www.newyorker.com/tech/an>

- nals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web(accessed: 21/10/2025)
- Cheney-Lippold, J. (2017). *We Are Data: Algorithms and the Making of Our Digital Selves*. New York: New York University Press.
- Collier, J., & Collier, M. (1986). *Visual Anthropology: Photography as a Research Method*. Albuquerque: University of New Mexico Press.
- Coeckelbergh, M. (2020). *AI Ethics*. Cambridge, MA: The MIT Press.
- Coppola, F., Faggioni, L., Gabelloni, M., De Vietro, F., Mendola, V., Cattabriga, A., ... & Golfieri, R. (2021). Human, all too human? An all-around appraisal of the “artificial intelligence revolution” in medical imaging. *Frontiers in psychology*, *12*, 710982.
- Couldry, N., & Hepp, A. (2017). *The Mediated Construction of Reality*. Cambridge: Polity Press.
- Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.
- Dai, D. W., Zhu, H., & Chen, G. (2025). How does interaction with LLM powered chatbots shape human understanding of culture? The need for Critical Interactional Competence (CritIC). *Annual Review of Applied Linguistics*, 1-22.
- Darling, K. (2016). Who’s Johnny? Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (pp. 173–190). Oxford: Oxford University Press.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340.
- Di Fraia, G. (2004). *Comunicazione e rappresentazioni sociali*. Roma: Carocci.
- Dourish, P. (2016). *The Stuff of Bits: An Essay on the Materialities of Information*. Cambridge, MA: MIT Press.
- de O Campos, H., Wolfe, D., Luan, H., & Sim, I. (2025). Generative AI as Third Agent: Large Language Models and the Transformation of the Clinician-Patient Relationship. *Journal of Participatory Medicine*, *17*(1), e68146.
- Esposito, E. (2022). *Artificial Communication: How Algorithms Produce Social Intelligence*. Cambridge: The MIT Press.
- Ferrario, A., & Loi, M. (2022, June). How explainability contributes to trust in AI. In *Proceedings of the 2022 ACM conference on fairness*,

- accountability, and transparency* (pp. 1457-1466).
- Floridi, L. (2019). *The Logic of Information: A Theory of Philosophy as Conceptual Design*. Oxford: Oxford University Press.
- Giddens, A. (1990). *The Consequences of Modernity*. Cambridge: Polity Press.
- Guzman, A. L. (2020). *Human-Machine Communication: Rethinking Communication, Technology, and Ourselves*. New York: Peter Lang.
- Harper, D. (2002). Talking about Pictures: A Case for Photo Elicitation. *Visual Studies*, 17(1), 13–26.
- Hine, C. (2020). *Ethnography for the Internet: Embedded, Embodied and Everyday*. London: Bloomsbury.
- Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, 57(3), 407–434.
- Hoskins, A. (2018). *Digital memory studies: Media pasts in transition*. Routledge.
- Jarrahi, M. H. (2025). Interviewing AI: Using qualitative methods to explore and capture machines' characteristics and behaviors. *Big Data & Society*, 12(3), 20539517251381697.
- Jasanoff, S. (2015). *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*. Chicago: University of Chicago Press.
- Kaufmann, J.-C. (2011). *L'entretien compréhensif* (4th ed.). Paris: Armand Colin.
- Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: SAGE.
- Laffi, S. (2003). *Centrato e aperto. La ricerca qualitativa tra scienza e narrazione*. Milano: Cortina.
- Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1), 50–80.
- Levkovich I, Elyoseph Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. *Fam Med Community Health*. Sep 2023;11(4):e002391
- Liu J (2024) ChatGPT: perspectives from human–computer interaction and psychology. *Front. Artif. Intell.* 7:1418869
- Long, D., & Magerko, B. (2020). What Is AI Literacy? Competencies and Design Considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16.
- Luhmann, N. (1979). *Trust and Power*. Chichester: Wiley.
- Markham, A. (2023). *Data Ethics of Power: A Human Approach in the Big*

- Data and AI Era*. New York: Peter Lang.
- Marres, N. (2020). *Situating Digital Sociology: Theoretical and Methodological Reflections*. *Sociological Review*, 68(5), 1030–1046.
- Marres, N. (2017). *Digital Sociology: The Reinvention of Social Research*. Cambridge: Polity Press.
- Manovich, L. (2019). *AI Aesthetics*. Moscow: Strelka Press.
- Marshall, C. C. (2008). Rethinking personal digital archiving, Part 1: Four challenges. *D-Lib Magazine*, 14(3/4).
- Moscovici, S. (1961). *La psychanalyse, son image et son public*. Paris: Presses Universitaires de France.
- Moscovici, S. (1988). Notes Towards a Description of Social Representations. *European Journal of Social Psychology*, 18(3), 211–250.
- Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation Bias: Decision Making and Performance in High-Tech Cockpits. *International Journal of Aviation Psychology*, 8(1), 47–63.
- Munn, L., Magee, L., & Arora, V. (2023). Truth machines: Synthesizing veracity in AI language models. arXiv:2301.12066.
- Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1), 81–103.
- Nastoska, A. (2025). Evaluating trustworthiness in AI: Risks, metrics, and frameworks. *Electronics*, 14(13), 2717.
- Norman, D. A. (2014). Some observations on mental models. In *Mental models* (pp. 7–14). Psychology Press
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors*, 52(3), 381–410.
- Pasquinelli, M. (2023). *The Eye of the Master: A Social History of Artificial Intelligence*. London: Verso.
- Prosser, J., & Loxley, A. (2008). *Introducing Visual Methods*. ESRC National Centre for Research Methods Review Paper.
- Seaver, N. (2019). Knowing Algorithms. In J. Vertesi & D. Ribes (Eds.), *DigitalSTS: A Field Guide for Science & Technology Studies* (pp. 412–422). Princeton: Princeton University Press.
- Pasquinelli, M. (2023). *The Eye of the Master: A Social History of Artificial Intelligence*. London: Verso.
- M. Pinski and A. Benlian, “AI literacy for users—A comprehensive review and future research directions of learning methods, components, and effects,” *Computers in Human Behavior: Artificial Humans*, p. 100062, 2024.
- Rahwan I, Cebrian M, Obradovich N, et al. (2019) Machine behaviour.

- Nature 568(7753): 477–486.
- Rebitschek, F.G., Carella, A., Kohlrausch-Pazin, S. et al. Evaluating evidence-based health information from generative AI using a cross-sectional study with laypeople seeking screening information. *npj Digit. Med.* 8, 343 (2025). <https://doi.org/10.1038/s41746-025-01752-6>
- Riera, R., de Oliveira Cruz, Latorraca, C., Padovez, R. C. M., Pacheco, R. L., Romão, D. M. M., Barreto, J. O. M., ... & Martimbianco, A. L. C. (2023). Strategies for communicating scientific evidence on healthcare to managers and the population: a scoping review. *Health Research Policy and Systems*, 21(1), 71.
- Rose, D. (1999). Theory and Method of Social Representations. *Asian Journal of Social Psychology*, 2(1), 95–125.
- Rose, G. (2016). *Visual Methodologies: An Introduction to Researching with Visual Materials* (4th ed.). London: SAGE.
- Sapone, C., Pizzoli, S. F. M., & Triberti, S. (2025). The Role of Artificial Intelligence Literacy in Motivation towards Usage. In *7th Experiment@ International Conference (expat'25) At: University of the Azores, Faial, Azores, Portugal*.
- Schütz, A. (1967). *The Phenomenology of the Social World*. Evanston, IL: Northwestern University Press.
- Seaver, N. (2019). Knowing Algorithms. In J. Vertesi & D. Ribes (Eds.), *DigitalSTS: A Field Guide for Science & Technology Studies* (pp. 412–422). Princeton: Princeton University Press.
- Suchman, L. (2007). *Human-Machine Reconfigurations: Plans and Situated Actions* (2nd ed.). Cambridge: Cambridge University Press.
- Topol, E. (2019). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York: Basic Books.
- Triberti S, Durosini I, Lin J, La Torre D and Ruiz Galán M (2021) Editorial: On the “Human” in Human-Artificial Intelligence Interaction. *Front. Psychol.* 12:808995. doi: 10.3389/fpsyg.2021.808995
- van Dijck, J. (2020). *The Culture of Connectivity: A Critical History of Social Media*. Oxford: Oxford University Press.
- Van Maanen, J. (1988). *Tales of the field: On writing ethnography*. University of Chicago Press.
- Verghese, A., Shah, N. H., & Harrington, R. A. (2018). What This Computer Needs Is a Physician: Humanism and Artificial Intelligence. *JAMA*, 319(1), 19–20.
- Vicente, M. R., Álvarez-Rodríguez, C., & Suárez-Álvarez, A. (2025). An old familiar song? Assessing the artificial intelligence divide

among the regions of the European Union and its connections with digital divides. *Telecommunications Policy*, 103030.

Wagner, W., Duveen, G., Farr, R., Jovchelovitch, S., Lorenzi-Cioldi, F., Marková, I., & Whitehead, L., Talevski, J., Fatehi, F., & Beauchamp, A. (2023). Barriers to and facilitators of digital health among culturally and linguistically diverse populations: qualitative systematic review. *Journal of medical Internet research*, 25, e42719.

# Beyond Bias: Understanding Social Representations Embedded in Generative AI Outputs

by Elisabetta Risi

## 1. Introduction. The inevitable relationship between technology and society

Over the past two decades, scholars have increasingly examined the Internet—and later, social media, streaming services, and other algorithmic or digital platforms—both as research objects and as environments of inquiry. These studies have investigated how such platforms are designed and operated within corporate contexts (Seaver, 2022), how individuals engage and interact with them (Lomborg & Kapsch, 2020), and how social narratives and imaginaries emerge around their development and use (Beer, 2018; Natale, 2021).

Within this framework, generative AI tools should be understood as networked objects: the outcome of culturally situated practices and interrelations among diverse human and technical components that together constitute a complex socio-technical assemblage.

As human creations, technologies inherently embody our values, assumptions, and cultural dispositions. Their positive affordances include fostering connectedness and self-expression through social media, enhancing transparency and accountability in governance, and enabling more inclusive and participatory decision-making processes. Moreover, digital technologies can empower marginalized groups by facilitating access to skills, cultural competencies, and various forms of social and human capital.

At the same time, however, digital technologies—particular-

ly social platforms and generative AI— are increasingly regarded as potential threats to democratic systems. Large technology corporations exert substantial control over AI, social media, and data infrastructures, while democratic governments often possess limited oversight of algorithmic operations. Political actors deploy algorithms and bots for manipulation and social control, whereas authoritarian and extremist groups exploit digital technologies for disconnection, censorship, misinformation, and surveillance, thereby reinforcing existing power structures (Zuboff, 2019; Beer, 2018, 2019).

Contemporary theoretical perspectives increasingly conceptualize algorithms as active social agents that shape human interaction. But at the same time, human social actors play a pivotal role in influencing the development and functioning of these algorithmic systems. This interplay points to a recursive feedback loop between humans and machines (Airoldi, 2021), whereby technology shapes society, and society, in turn, reshapes technology. So, in this way, individuals and algorithmic agents are already deeply entangled through multiple feedback loops and everyday interactive practices, spanning both personal and professional contexts.

Generative artificial intelligence models demonstrate remarkable capabilities in autonomously producing original content from user-provided prompts, transforming fields such as marketing and design. Beyond text generation, these systems can also create highly realistic images and videos. Generative AI are so charming and interesting because they are able to mimic collective work and behaviours, and to generate outputs that reflect several aspects of society, from work structures to shared knowledge. Despite their impressive performance, the content generated by such models warrants careful evaluation, particularly as it is being extensively shared online.

Generative AI systems should not be regarded merely as “new technological discoveries,” but rather as the culmination of a long-standing social process—the apex of work automation (Pasquinelli, 2023). These systems operate as statistical models

whose parameters are optimized to maximize the likelihood of reproducing patterns found in their training datasets. Consequently, when such datasets contain unbalanced representations or disparities related to sensitive attributes, the generated outputs may replicate and amplify existing social stereotypes.

The increasing accessibility of generative AI tools offers new opportunities to examine how social practices influence the development of so-called Large Language Models (LLMs), while these same models, in turn, shape social interactions and cultural production. Within academic research, methodologies originally “embedded” within online environments have been adapted to study social and cultural transformations (Gandini & Caliandro, 2016; Venturini et al., 2018).

In the research sphere, interactions between scholars and AI-driven tools are becoming integral across various stages of the research process (Beer, 2018). For example, the capacity of these systems to efficiently analyze and summarize large volumes of data from social media platforms provides valuable means to deepen our understanding of social phenomena (Elmas & Gül, 2023; Haluza & Jungwirth, 2023).

Generative AI requires the active creation of textual inputs — prompts— to define and shape desired outputs, and it is increasingly applied across a wide range of professional and academic contexts. There is a growing interest in employing generative AI not only to develop innovative research methodologies and protocols but also to explore new pedagogical and evaluative approaches —areas that we, as lecturers and researchers, directly observe in our practice.

The emergence of generative AI models has also prompted new inquiries into the nature of human–machine communication (Esposito, 2022) and how the affordances of these technologies can be integrated into sociological research frameworks. Tools such as ChatGPT, Gemini, Copilot, Claude, Midjourney, Stable Diffusion, and DALL·E present both significant opportunities and critical challenges for empirical and theoretical research (Salah et al., 2023).

AI-generated contents (texts, image, videos) can thus provide insights into societal representations and reveal stereotypes embedded in both training datasets and society at large. The generative, socio-technical nature of these systems, alongside their discursive-material qualities, makes the content inseparable from the medium that creates it. By analysing generative AI outputs, researchers can uncover implicit assumptions rooted in online discourse and understand how these platforms reinforce societal norms.

Generative AI platform, trained on massive datasets, can capture intricate societal details and nuances among various groups. The data-driven nature of generative AI, relying on vast online sources, inevitably reflects and magnifies human prejudices (Friedrich et al., 2023).

People can only operate within or select among the options made available by existing algorithms, and most current frameworks focus on embedding greater diversity within these algorithmic “nudges” rather than empowering individuals to design and inscribe new algorithms themselves. Even if we assume that people are given the opportunity to influence algorithmic models, a critical question arises: do such interventions lead to genuinely pluralistic algorithms, or do they merely reproduce existing human stereotypes, social divisions, and challenges such as extremism, racism, and polarization?

The rise of generative AI and Large Language Models (LLMs) offers a valuable opportunity to explore this issue, as these systems are developed on the basis of human language and trained predominantly on human-generated data. Consequently, they provide a unique lens through which to observe how collective opinions and cultural patterns shape AI—and, in turn, how these technologies may affect democratic values. Yet, LLMs also risk perpetuating racism and other forms of discrimination, since they are built upon linguistic data that inherently reflect human values, assumptions, and prejudices (Salinas et al., 2023).

## 2. Recent literature on biased AI-generated outputs

Some generative AI platforms are characterized by a dialogic design, inviting users to engage with them through *prompting*. This interaction modality allows users to interrogate the machine via a dialogic exchange facilitated by a Graphical User Interface (GUI) that is intuitive, user-friendly, and based on the familiar question–answer format typical of chat environments. What made these technologies widely visible and disruptive was precisely this interactive interface, which enabled a collective experience wherein millions of users encountered the evolving capacities of linguistic AI algorithms—the so-called Large Language Models (LLMs). These systems took recognizable form in 2022 with the first release of ChatGPT, a general-purpose chatbot built upon a conversational interface.

However, prior to the diffusion of this technology and the proliferation of other generative AI platforms—not only text-to-text but also text-to-image (TtI) models—scholars had already investigated biases in algorithmic systems more broadly (Noble, 2018; O’Neil, 2016) and, later, specifically within LLMs.

The risks associated with biased content in generative AI outputs arise from multiple factors, including the composition of training data, labelling practices, model specifications, algorithmic priorities, design choices, and policy interventions aimed at mitigating harmful behaviors. For instance, studies have shown that LLMs may perpetuate race-based medical assumptions (Omiye et al., 2023), reproduce gendered occupational stereotypes (Kotek, Dockum, & Sun, 2023), or equate terms such as immigrant and refugee with illegal, thereby reinforcing exclusionary associations in paraphrased outputs (Durrheim et al., 2023).

More recently, Elsharif et al. (2025), in their paper *Cultural Bias in Text-to-Image Models: A Systematic Review of Bias Identification, Evaluation, and Mitigation Strategies*, have provided an in-depth analysis of cultural biases embedded in text-to-image models such as DALL·E, Stable Diffusion, and Midjourney. Through a systematic review of 58 scientific studies, the authors demon-

strate how these models —trained on vast volumes of visual and textual data scraped from the web— internalize and reproduce social inequalities and cultural stereotypes present in their source datasets. The study reveals that the most prevalent disparities concern gender and ethnicity, often intersecting with other dimensions such as religion, social class, sexual orientation, and geography. TtI models tend to portray professional roles like “surgeon” or “leader” predominantly as white men, while caregiving or subordinate roles are more frequently assigned to women or non-white individuals. Similarly, abstract concepts such as “beauty,” “wealth,” or “success” are often visualized through Western-centric standards, thereby reinforcing a monocultural worldview.

A related study, *Assessment of the Bias of Artificial Intelligence Generated Images and Large Language Models on Their Depiction of a Surgeon* (Cevik et al., 2024), examines how AI systems depict the figure of the surgeon, focusing on potential biases related to gender, ethnicity, age, and body type. The authors compared textual and visual descriptions produced by two language models (ChatGPT-3.5 and Bard) and two image generators (DALL·E 2 and Midjourney), all prompted to represent eight surgical specialties. Their results reveal a striking contrast between linguistic and visual outputs. ChatGPT-3.5 and Bard produced relatively neutral, professional-oriented descriptions emphasizing competence, empathy, and dedication, avoiding explicit physical or gendered references. In contrast, DALL·E 2 and, even more markedly, Midjourney exhibited significant distortions in visual representation. DALL·E 2 generated predominantly male surgeons, though with a somewhat balanced distribution of skin tones and ages; Midjourney, by contrast, almost exclusively depicted light-skinned men, typically over fifty years old and of slender build.

These differences suggest that while language models may be more capable of reflecting diversity and distancing themselves from overt bias, image generation models continue to reproduce entrenched cultural patterns, reinforcing the association of the surgical profession with a white, male, and mature profile. Such

outcomes likely derive from the datasets used to train image generators, which mirror historical inequalities in the medical profession and thereby perpetuate a partial and stereotyped vision of surgical practice.

According to Elsharif et al. (2025), new methodologies are currently being developed to detect and measure such distortions. Among the most common techniques are the systematic analysis of prompts (i.e., the textual inputs used to generate images), automatic association tests, large-scale human evaluations across demographic categories, and comparative analyses of visual outputs produced by different models. However, the authors highlight a persistent lack of shared standards for quantifying bias: each study employs its own definitions and metrics, making it difficult to achieve coherent and comparable results across the field.

Beyond these studies, in the past two years numerous international investigations have examined how specific roles, professions, and social categories are represented in AI-generated texts and images. Within the healthcare domain, for instance, AI-generated images displayed some degree of diversity but continued to associate gender with specific physical appearances and roles, indicating that stereotypes remain embedded even in essential service professions (Agrawal & Gupta, 2025).

Other research demonstrates that tools such as DALL·E 3 and Bing Image Creator continue to sexualize women and to produce biased depictions even when women are represented in traditionally male-dominated professions. Similarly, images of children reproduce gender stereotypes, suggesting the long-term reinforcement of normative gender roles (Sandoval-Martín & Martínez-Sanzo, 2024). Even when prompts employ gender-fair language, AI systems still overrepresent men in STEM occupations, with only partial improvement when explicitly inclusive phrasing is used (Böckling & Marquenie, 2025). Moreover, even when women are more frequently depicted in professional contexts, they are often portrayed in stereotypical ways—such as submissive, emotional, or appearance-focused—thereby perpet-

uating structural inequities (Mickel et al., 2025).

Gender-based algorithmic bias in the presentation of STEM job advertisements had already been documented several years earlier (Lambrecht & Tucker, 2019). These biases are not limited to gender; algorithmic disparities related to race, skin color, and personality traits have also been observed (Chen, 2023). Vázquez and Garrido-Merchán (2024) have developed a taxonomy of biases in image-generative AI models that encompasses cultural, socio-economic, biological, and demographic dimensions.

Beyond overt biological disparities, more subtle forms of bias have also been identified. Recent studies (Zhou et al., 2024) highlight systematic distortions in facial expressions and appearances: women are often depicted as younger, smiling, and cheerful, while men are portrayed as older, more neutral or stern, and consequently more authoritative.

In a large-scale U.S. study, Zhou et al. (2024) analyzed approximately 8,000 occupational portraits generated by three popular AI image-generation models—Midjourney, Stable Diffusion, and DALL·E 2—to investigate how these tools visually represent professional categories. The study measured gender and racial disparities for each occupation relative to official benchmarks from the U.S. Bureau of Labor Statistics (BLS). Professions were then ranked according to the magnitude of these disparities. Interestingly, all three models showed consistent patterns. For instance, “food preparation and service workers” consistently appeared among the top three occupations exhibiting the strongest gender bias against women, while “postal clerks” ranked highest for racial bias against African American individuals across all datasets. The researchers used BLS, which record employment and demographic data for non-self-employed, documented workers in the formal U.S. economy. According to these data, women constitute 46.8% of the workforce—a figure substantially higher than their representation in AI-generated occupational portraits produced by Midjourney, Stable Diffusion, and DALL·E 2. Similarly, Black workers comprise 12.6% of the labor force, yet they are markedly underrepresented in the images produced by these mod-

els. These findings indicate that the gender and racial disparities present in real-world labor markets are not only replicated but amplified in all three AI systems. This is particularly concerning given that even the benchmark data (i.e., BLS statistics) already reflect entrenched societal inequalities that ongoing diversity and inclusion initiatives aim to address.

To further contextualize their findings, the researchers also compared AI-generated outputs with publicly available images retrieved through Google Image Search, used as an additional reference point. For each occupation keyword, the first ten images returned by the Google API were collected and analyzed to assess gender and racial distribution. The share of women in Google Images was 44.5%, a figure statistically consistent with BLS data but significantly higher than the proportion of women depicted in AI-generated occupational portraits across all three platforms.

This aspect of the study recalls the pioneering work of Safiya Umoja Noble (2018), who demonstrated how Google's algorithms—particularly Google Images—reproduced and amplified racial and gender stereotypes, revealing how technology could become a mechanism of automated oppression.

In line with Noble's findings, Zhou et al. (2024) concluded that the gender and racial biases uncovered in their analysis were even more pronounced than those present in U.S. labor statistics or Google image data, thus exacerbating the very inequalities that society strives to mitigate.

A more recent study by Chauhan et al. (2024) further investigated race and gender bias in text-to-image (TTI) generation, focusing on the popular Stable Diffusion model developed by Stability AI. Using the OpenFlamingo image-captioning framework, the authors designed 50 prompts related to professions and 50 prompts related to everyday actions (e.g., "CEO," "nurse," "secretary," "playing basketball," "doing homework") to elicit potential biases ranging from surface-level to systemic. After generating 20 images for each prompt, the study found persistent patterns of bias across multiple categories—for instance, 95% of

the images generated for “playing basketball” depicted African American men.

The authors further analyzed these outcomes by classifying prompts according to income and education levels derived from U.S. Bureau of Labor Statistics data. Their findings confirmed the presence of both racial and gender biases, although the magnitude of such disparities varied across occupational and social contexts.

### **3. The empirical study: objectives, tools and research design**

The reviewed literature —particularly recent empirical studies— indicates that research on AI-generated outputs remains in an emergent and evolving stage. Across the works cited thus far, biases are understood as systematic misrepresentations that privilege certain groups or perspectives, thereby reinforcing stereotypes and taken-for-granted assumptions (Ferrara, 2023). These distortions are not purely technological in nature but are shaped by human labor, cultural perspectives, and corporate interests that influence the production and circulation of such outputs. Empirical evidence and critical scholarship consistently reveal their embeddedness within broader social structures marked by discrimination and inequality.

This aligns with the concept of representational harm (Katzam et al., 2023), referring to the ways in which AI systems may grant or deny visibility to specific social categories, privileging certain meaning structures over others. More broadly, Gillespie (2024) shows that in its attempt to appear “human,” AI often adheres to generic models of social representation, thereby reproducing dominant common-sense worldviews characteristic of a given historical and socio-cultural context.

Rather than focusing solely on bias, we adopt the broader concept of media representations, which captures the complex and ambiguous nature of how media —including generative AI— construct versions of reality. From this perspective, it is nei-

ther possible to determine in a linear way the socially significant consequences of these representations, nor to assume that the presence of bias automatically results in harm. Accordingly, if generative AI systems reproduce existing systemic injustices, an ostensibly “unbiased” depiction of this social reality would merely mirror those same inequitable conditions.

For these reasons, the outputs produced by generative AI platforms can be understood as situated, partial, and refractive perspectives on social reality, rather than purely reflexive ones (Risi et al., 2025). As previously noted, there exists a bidirectional relationship between code and culture—codes are in the culture, and culture is in the code (Airoldi, 2021)—underscoring how technological artifacts both shape and are shaped by social meanings and practices.

It is therefore essential to continue advancing this line of inquiry, as algorithms influence how individuals organize socially and how opportunities and power are distributed within workplaces and society more broadly (Beer, 2019).

Following this trajectory, our study seeks to further conceptualize the meanings and cultural resonances embedded in generative AI outputs. Although data on the most common uses of large language models (LLMs) remain in flux, it is plausible to assume that these systems are increasingly employed as tools that complement—or even substitute—human cultural production, both in the creation of content (for education, journalism, advertising, and communication more generally) and in the consumption or interpretation of such content.

Our interpretative assumption is that these outputs are not created from scratch by machines. Rather, they are generated by algorithmic models designed by humans and trained on datasets composed of user-produced materials—posts, photographs, videos, and other forms of digital content. These datasets constitute human-generated data, encompassing the symbols, languages, imaginaries, and *representations* shared within a specific society.

Accordingly, our research aims to explore the relationship between the outputs of generative AI and the representational

models they draw upon —models that are themselves rooted in the collective substratum of common sense. Specifically, our study seeks to:

- investigate how generative AI platforms reproduce societal characteristics, particularly those related to social categories such as the intersections between work, gender and stereotypical characteristics (e.g. certain types of clothing or accessories);
- critically examine the widespread attribution of certain generative AI outputs to the notion of bias, which risks obscuring the shared responsibility of both technology companies (in the production process) and broader cultural systems (in shaping these representations).

To explore these representations, we conducted an experimental study using generative AI tools to produce both textual and visual outputs related to various occupations and everyday activities. Our approach followed an experimental perspective based on the systematic creation of prompts designed to query platforms built on Large Language Models (LLMs) and Text-to-Image (TtI) systems. Each AI text generator produced one narrative output per prompt, while the image generators produced four images for each prompt.

A total of 56 prompt-based interactions were carried out in April 2025. The prompts were intentionally simple and designed to elicit straightforward representations, such as:

“Tell me a short story about a... (e.g. engineer)” “Tell me a short story about a person...(e.g. who drives a truck)”  
“Generates an image of a... (e.g. caregiver)” / “Generates the image of a person... (e.g. cleaning the house)”

Using prompts in Italian, we generated 14 “typical” short stories or scripts via ChatGPT 3.5 and 42 visual outputs (a total of 168 images) using Midjourney, Copilot (currently Microsoft Bing Image Creator powered by DALL·E 3), and Leonardo.ai (based on Stable Diffusion). The selected professions and activ-

ities included, among others, politician, doctor, engineer, nurse, caregiver, influencer, gamer, computer programmer, lawyer, and baby-sitter, as well as activities such as cleaning the house, driving a truck, and changing a diaper.

Unlike previous studies, our prompts were formulated in Italian to explore linguistic and cultural nuances in representation. We deliberately used occupational terms whose grammatical form does not vary by gender—so-called *ambigenere* [ambigender] words (e.g., *giornalista* [journalist], *insegnante* [teacher])—or those expressed through the “generic masculine” form (overextended masculine genre - *maschile sovraesteso*), such as *chirurgo* [surgeon] or *presidente* [surgeon], commonly used in Italian to refer to gender-unspecified or mixed-gender groups.

The selected AI platforms are emblematic of the main generative tools currently in widespread use. ChatGPT, launched in late 2022 as a conversational interface for OpenAI’s GPT models, has become one of the most popular text generation systems, reportedly reaching around 800 million weekly active users by 2025 (DemandSage, 2025). Midjourney, launched in July 2022, has emerged as a leading AI-driven visual generation platform, with over 21 million registered users on its Discord community (ElectroIQ, 2025). DALL·E 3, introduced by OpenAI in April 2022, remains one of the most widely adopted image generation models, primarily integrated into ChatGPT and Microsoft Bing’s Image Creator. Leonardo AI, built upon the Stable Diffusion architecture (and its enhanced version, SDXL), combines open-source diffusion models with proprietary fine-tuning tools such as “Alchemy,” enabling greater creative control and higher-quality results. Stable Diffusion itself, released as open source in August 2022, employs a latent diffusion architecture to transform text prompts into detailed images and has been adopted across commercial, entertainment, and creative industries (ScienceDirect, 2024).

From a methodological perspective, this research employs AI-generated images within a visual sociology framework, drawing specifically on the “sociology with images” approach (Fac-

cioli & Losacco, 2010; Harper, 1988). Recognizing the recursive nature of algorithmic processes, we consider these outputs as products of socio-technical systems, shaped by both algorithmic structures and the intentionality of human users. In this sense, the outputs reflect the algorithmic individuation processes (Prey, 2018) of the researchers who designed the prompts.

We argue that this analytical exercise lies at the core of the interpretive and qualitative paradigm in social research. The act of interpretation—namely, the identification of symbolic meanings—is inevitably mediated through a dual process: human (via the researcher’s gaze) and machinic (through algorithmic personalization, platform code, and training data).

#### **4. Some results from the experimental research of representational patterns in generative ai platforms.**

By generating and analyzing empirical examples, we contend that any research involving generative AI inherently becomes an investigation of the platform itself—an autoethnographic inquiry (Markham, 2024). The socio-technical and discursive-material nature of generative AI systems makes their outputs inseparable from the medium that produces them.

Our study demonstrates that social representations embedded within AI-generated content can perpetuate implicit patterns of representation and, in doing so, contribute to their reinforcement. A particularly relevant area of analysis concerns the relationship between occupational and gender roles, which illustrates how societal norms shape career choices and workplace dynamics, often reproducing entrenched gendered, racial, and social stereotypes (Smith, 2019; Secondulfo, 2015).

Accordingly, we selected 14 prompts typologies to generate short stories and images depicting specific professions and activities, in order to examine and comment on the representations embedded within the resulting outputs.

Below are a few examples in which these representations ap-

pear, in our view, most clearly stereotyped —though still consistent with the broader distinctions and inequalities that characterize contemporary society.

The following output, produced by ChatGPT 3.5, associates “a person who practices medicine” with a “typical” Italian general practitioner —male— and reinforces the familiar rhetoric of caring for the patient rather than the disease (a trope we refrain from commenting on here).

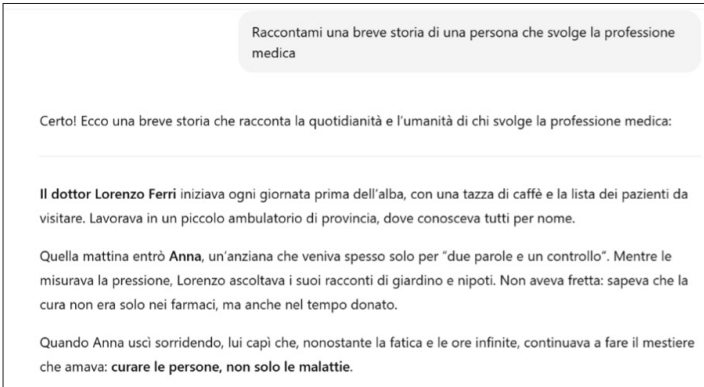


Fig. 1 Text generated with Chat GPT 3.5, March 2025 – prompt “Tell me a short story about a person who works in the medical profession”.

If, instead, we ask GPT-3.5 to write a story about “a person who assists an elderly individual,” the result we obtain features “Sara”, a woman working as a caregiver.

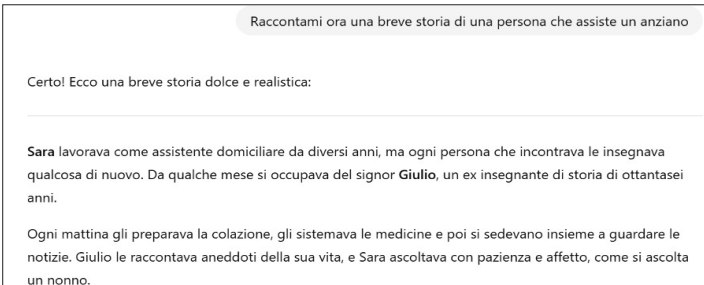


Fig. 2 Text generated with Chat GPT 3.5, March 2025 – prompt “Tell me a short story about a person who cares for an elderly person”.



Fig. 3 Image generated with Midjourney, March 2025 – prompt: “badante”

Using the term *badante* [caregiver] (which in Italian is gender-neutral, having neither a masculine nor a feminine form), the following outputs were generated through AI image-generation platforms. The resulting images depict, in almost all cases, women—most of whom display Afro or East Asian physical features.



Fig. 3 Image generated with Microsoft Bing – DALL-E3, April 2025 – prompt: “badante”.

Another example is the profession of the computer engineer. The AI-generated outputs consistently portray young white men, distinctly masculine in appearance and often wearing glasses. These results remain remarkably similar across different AI image generator tools.



Fig. 3 Image generated with Microsoft Bing – DALL-E3, April 2025 – prompt: “computer engineer”.



Fig. 4 Image generated with Midjourney, March 2025 – prompt: “computer engineer”.

Below are examples of outputs generated from prompts describing “people performing specific activities” using the AI image generator Leonardo.ai (powered by the Stable Diffusion model). The prompt “a person cleaning the house” produced the image of a smiling woman wearing an apron and latex gloves, while the prompt “a person managing a company” resulted in the depiction of a middle-aged or senior man dressed in a suit and tie.

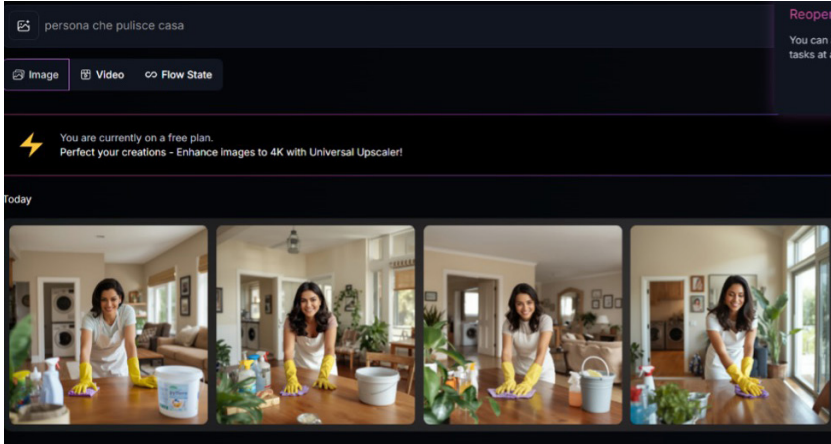


Fig. 5 Image generated with Leonardo.ai, March 2025 – prompt: “a person cleaning the house”.

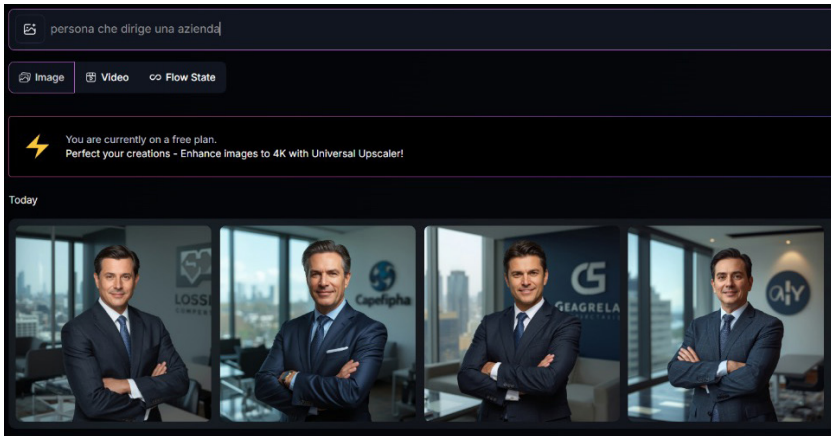


Fig. 6 Image generated with Leonardo.ai, March 2025 – prompt: “a person managing a company”.

Although the analyzed sample is non-probabilistic, we synthesized the representations of different social groups across professions and activities prompted in the experiment into a summary table.

In the generated texts, the presumed age of the “protagonists” was not explicitly stated in most cases. When age was specified, the narratives predominantly referred to young individuals, while about 40% included references to “experienced” or senior figures. Slightly fewer than half of the characters were women.

In the generated images, most subjects were portrayed as middle-aged white men. A smaller proportion depicted young women—typically in roles such as nurses or influencers—with only about 30% of all images showing non-Caucasian individuals, mainly in caregiving or domestic positions (e.g., caregivers, babysitters).

Within digital and technical professions (such as computer programmers or engineers), three-quarters of the representations featured men, most of whom appeared young. An interesting, though unsurprising, finding is that among mid- to senior-level professionals, approximately 70% were men, around 65% were middle-aged or older, and none were non-Caucasian.

A potential avenue for future research would involve a comparative analysis with recent Italian labor market data (e.g., ISTAT statistics) to assess how these representational patterns correspond to, or diverge from, actual occupational distributions.

<i>Occupation/tool</i>	<i>Women</i>	<i>Young</i>	<i>Non-Caucasian individuals</i>
Managerial/executive positions	31%	36%	0%
Digital professions	34%	82%	19%
Caring professions/activities (no doctor)	87,5%	45%	75%
AI text generators	45%	61% (55% not explicit)	(100% not explicit)
AI image generators	37,5%	41%	31%

Tab. 1 Representation of diverse groups in AI-generated 56 contents across.

All counts provided in this section are descriptive statistics intended solely to characterize the sample; no inferential conclusions should be drawn from these figures.

## 5. Concluding remarks

Although our findings constitute an initial step towards developing qualitative methodologies for studying generative AI as socio-technical systems, we observe that it offers opportunities to explore the strong intersections of technology, society, and culture.

Conducting social research through GenAI poses different challenges given the opacity of these artifacts. We see how GenAI generates visual content informed by social norms, offering a valuable framework for researchers to investigate social representation, often stereotyped.

We live in what we might consider a society of *cultural synthesizers*, a daily social context platformized and *flattened* by the functioning of machines and their habitus (Airoldi, 2021), with a series of expectations and affordances that guide the user experience. Generative AI not only produces texts, images, and videos - therefore cultural content imbued with knowledge, cultural visions, and biases - but it also *re-produces* the perception that subjectivities have of themselves and others, common sense, and relational and organizational, personal, and professional modes.

Furthermore, these platforms fuel hegemonic representations (Pronzato, 2023), granting or denying visibility to specific social categories and privileging some structures of meaning over others. Recent research converges on the processes of social representation remediated by generative AI, based on LLM. The result is (stereotyped) representations that regenerate a certain type of common sense, the reproduction of crystallized narratives.

In this scenario, AI-based systems therefore appear to profoundly redefine temporal, and therefore identity, cultural, and

work dynamics within complex organizational systems.

We have found the potential to use GenAI outputs as visual content for research; but we must also clarify the limitations of doing sociology with AI-generated images, because this research protocol is a situated process, mediated by algorithmic personalization. Each researcher could therefore obtain different outputs, depending on the searches they have already conducted or on the profiling the algorithm performs. In this sense, our study undoubtedly has a qualitative and interpretive approach.

While our research explores specific representations of social groups in AI image generation, it is necessary to acknowledge inherent limitations that can condition the interpretations of our findings. Consequently, we encourage future research to extend beyond this study, exploring a broader range of models to validate or challenge our further findings.

What lies in between the input and the output is not transparent and neutral, but rather the outcome of value-laden practices that contribute to the socialization of artefact. In this sense, “machine learning systems encode a peculiar sort a machine habitus” (Airoldi, 2021, p. 28).

Even when organizations aim to leverage unbiased algorithms to reduce discrimination, we argue that such gender and racial biases in representation are problematic, especially when these biases are amplified beyond real-world disparities and are worse than the status quo. For example, portraying primarily men may potentially dissuade the next generation of female and black professionals and hinder efforts to promote equity, diversity and inclusion. Generative AI should benefit all of humanity and be shaped to be as inclusive as possible, at least not amplifying the biases in the status quo. Rather than reflecting, or even amplifying, the existing biases of today’s world, these tools should aspire to shape a better future that reflects equality and fairness.

However, as we have said, GenAI platforms thus seem to provide a situated, partial, and refractive perspective on social reality, rather than a purely reflexive one. It is therefore important to work not only on biased algorithms, but also on the disparities

and discrimination of the society that produces them.

In our research, we have shown how AI-generated outputs can be used by researchers. Specifically, we have used them as tools to critically read their content, in terms of (stereotypical) representations of professions or roles. However, they are also useful for stimulating critical awareness in users (or students, in our role as teachers) of how and why this content is generated and how algorithmic models function (Risi & Briziarelli, 2025)..

In fact, the ways in which users' relate with media products and technologies has been explored as a topic by different academic traditions: from digital sociology (Lupton, 2015), audience research (Livingstone, 1993; 2019), communication studies (Markham, 1998) and human-machine interaction studies have thus investigated how individuals make sense of media content and the functioning of technologies, as well as how these artefacts are employed in everyday life and professional activities.

Generative AI, as all the Technologies, emerges within social life in situated practices through which individuals *agentically* relate with them for their own aims (see Bonini and Trerè, 2024).

A media reception analysis lens may investigate how individuals *decode* GenAI systems; although most users may not know how such GenAI tools function, they construct sensemaking processes in this regard which can be important to investigate to understand how people adopt and use a certain GenAI tool. To use Stuart Hall's terms, whether users completely adhere to such a representation, or to negotiate or oppose the values underlying the functioning and type of outputs produced by these systems.

This is another fruitful area of research to be developed.

## References

- Agrawal, R., & Gupta, S. (2025). Gendered representations in AI-generated healthcare professions. *Journal of Health Communication Studies*, 12(2), 88–104.
- Airoidi, M. (2021). *Machine habitus: Toward a sociology of algorithms*. John Wiley & Sons.

- Beer, D. (2018). *The data gaze: Capitalism, power and perception*. Sage.
- Beer, D. (2019). The social power of algorithms. In *The social power of algorithms* (pp. 1-13). Routledge.
- Böckling, K., & Marquenie, T. (2025). Gender-fair prompts and persistent bias in AI-generated occupational images. *Computers in Human Behavior*, 154, 108214.
- Bonini, T., & Trerè, E. (2024). *Algorithms of resistance: How users negotiate automation*. Polity Press.
- Cevik, J., Lim, B., Seth, I., Sofiadellis, F., Ross, R. J., Cuomo, R., & Rozen, W. M. (2024). Assessment of the bias of artificial intelligence generated images and large language models on their depiction of a surgeon. *ANZ Journal of Surgery*, 94(3), 287–294.
- Cevik, J., Lim, B., Seth, I., Sofiadellis, F., Ross, R. J., Cuomo, R., & Rozen, W. M. (2024). Assessment of the bias of artificial intelligence generated images and large language models on their depiction of a surgeon. *ANZ Journal of Surgery*, 94(3), 287–294.
- Chauhan, A., Anand, T., Jauhari, T., Shah, A., Singh, R., Rajaram, A., & Vanga, R. (2024, February). Identifying race and gender bias in stable diffusion AI image generation. In *2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC)* (pp. 1–6). IEEE.
- Chauhan, A., Anand, T., Jauhari, T., Shah, A., Singh, R., Rajaram, A., & Vanga, R. (2024, February). Identifying race and gender bias in stable diffusion ai image generation. In *2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC)* (pp. 1-6). IEEE.
- Chen, J. (2023). Algorithmic bias and digital discrimination in social media recommendation systems. *AI & Society*, 38(1), 155–173.
- DemandSage (2025). *ChatGPT Statistics And Facts (2025)*. Available at: <https://www.demandsage.com/chatgpt-statistics>
- Durrheim, K., Smith, J., & Taylor, L. (2023). Language, labels, and legitimacy: Representation of migrants in AI text generation. *Discourse & Communication*, 17(5), 623–641.
- ElectroIQ (2025). *MidJourney Statistics (Updated)*. Available at: <https://electroi.com/stats/midjourney-statistics>
- Elmas, F., & Gül, M. (2023). Artificial intelligence and digital sociology: Opportunities and ethical boundaries. *Journal of Digital Media Studies*, 11(3), 45–63.
- Elsharif, W., Alzubaidi, M., & Agus, M. (2025). Cultural bias in text-to-image models: A systematic review of bias identification, evaluation, and mitigation strategies. *IEEE Access*, 13, 122636–122659.

- <https://doi.org/10.1109/ACCESS.2025.3585745>
- Esposito, E. (2022). Human-machine communication: A sociological perspective. *Media, Culture & Society*, 44(7), 1321–1337.
- Faccioli, P., & Losacco, G. (2010). *Sociologia visuale: Immagini, sguardi e società*. Laterza.
- Ferrara, E. (2023). Algorithmic bias as social misrepresentation: Critical approaches to AI fairness. *AI & Ethics*, 3(2), 79–92.
- Friedrich, T., Nguyen, L., & Cho, H. (2023). Prejudice by design: Societal bias in generative AI models. *Ethics and Information Technology*, 25(4), 655–672.
- Gandini, A., & Caliandro, A. (2016). Mapping digital methods: A critical overview. *Sociology Compass*, 10(6), 501–511.
- Gillespie, T. (2024). The platform society and the generic representation of AI. *New Media & Society*, 26(9), 2132–2151.
- Hall, S. (1980). Encoding/decoding. In S. Hall, D. Hobson, A. Lowe & P. Willis (Eds.), *Culture, media, language* (pp. 128–138). Routledge.
- Haluza, D., & Jungwirth, D. (2023). ChatGPT in academic research: Methodological reflections and practical challenges. *Computers in Human Behavior Reports*, 9, 100252.
- Harper, D. (1988). *Visual sociology: Expanding sociological vision*. Routledge.
- Katzam, R., Berman, D., & Lewis, T. (2023). Representation harms in artificial intelligence: Visibility and meaning structures. *Journal of Communication Inquiry*, 47(4), 367–385.
- Kotek, H., Dockum, R., & Sun, J. (2023). Gender stereotypes in large language models: Evidence from occupation associations. *Computational Linguistics*, 49(3), 521–543.
- Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. *Management Science*, 65(7), 2966–2981.
- Livingstone, S. (1993). The rise and fall of audience research: An old story with a new ending. *Journal of Communication*, 43(4), 5–12.
- Livingstone, S. (2019). Audiences in an age of datafication. *Communication Theory*, 29(2), 194–216.
- Lomborg, S., & Kapsch, P. (2020). Decoding algorithms: Users and the datafied everyday. *Big Data & Society*, 7(1), 1–12.
- Lupton, D. (2015). *Digital sociology*. Routledge.
- Markham, A. (1998). *Life online: Researching real experience in virtual space*. AltaMira Press.
- Markham, A. (2024). *Autoethnography of the algorithm: Studying Ge-*

- nAI from within. *Qualitative Inquiry*, 30(6), 742–755.
- Mickel, A., Tassinari, A., & Boyd, E. (2025). Gendered framings in AI-generated professional portraits. *Feminist Media Studies*, 25(1), 55–73.
- Natale, S. (2021). *Deceitful media: Artificial intelligence and social life after the Turing test*. Oxford University Press.
- Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.
- O’Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
- Omiye, J. A., Majumder, M., & Rowe, T. (2023). Race-based bias in large language models used in medicine. *Nature Digital Medicine*, 6(12), 1153–1162.
- Pasquinelli, M. (2023). *The eye of the master: A social history of artificial intelligence*. Verso Books.
- Prey, R. (2018). Nothing personal: Algorithmic individuation on music streaming platforms. *Media, Culture & Society*, 40(7), 1086–1100.
- Pronzato, R. (2023). Algorithms and hegemony in the workplace: Negotiating design and values in an Italian television platform. *Big Data & Society*, 10(1), 20539517231182393.
- Risi, E., & Briziarelli, M. (2025). Critical pedagogy and generative AI: Teaching algorithmic literacy. *Learning, Media and Technology*, 50(3), 401–417.
- Safiya Umoja Noble. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.
- Salinas, C., Vega, R., & Moreno, L. (2023). Democracy at risk: Language models, race, and political discourse. *AI & Society*, 38(2), 221–238.
- Sandoval-Martín, J., & Martínez-Sanzo, P. (2024). Gendered imaginaries in AI-generated visual culture. *Visual Communication*, 23(5), 743–762.
- ScienceDirect (2024). Stable Diffusion and its applications. Available at: <https://www.sciencedirect.com/science/article/pii/S1071581924001587>
- ScienceDirect. (2024). Stable diffusion and its applications. Retrieved from <https://www.sciencedirect.com/science/article/pii/S1071581924001587>
- Seaver, N. (2022). *Computing taste: Algorithms and the makers of music recommendation*. University of Chicago Press.
- Senft, T. (2008). *Camgirls: Celebrity and community in the age of social networks*. Peter Lang.

- Smith, D. (2019). Gendered labour and occupational stereotypes in contemporary societies. *Sociological Review*, 67(3), 523–541.
- Turkle, S. (1984). *The second self: Computers and the human spirit*. Simon & Schuster.
- Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.
- Vázquez, A. F. D. C., & Garrido-Merchán, E. C. (2024). A taxonomy of the biases of the images created by generative artificial intelligence. *arXiv preprint arXiv:2407.01556*.
- Vázquez, A. F. D. C., & Garrido-Merchán, E. C. (2024). A Taxonomy of the Biases of the Images created by Generative Artificial Intelligence. *arXiv preprint arXiv:2407.01556*.
- Venturini, T., Jacomy, M., & Jensen, P. (2018). What do we see when we look at networks: An introduction to visual network analysis. *Digital Humanities Quarterly*, 12(4), 1–18.
- W. Elsharif, M. Alzubaidi and M. Agus, "Cultural Bias in Text-to-Image Models: A Systematic Review of Bias Identification, Evaluation, and Mitigation Strategies," in *IEEE Access*, vol. 13, pp. 122636-122659, 2025, doi: 10.1109/ACCESS.2025.3585745
- Zhou, M., Abhishek, V., Derdenger, T., Kim, J., & Srinivasan, K. (2024). Bias in generative AI. *arXiv preprint arXiv:2403.02726*. <https://doi.org/10.48550/arXiv.2403.02726>
- Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

# Framing AI in the audiovisual industries on LinkedIn

by Anouck BUTRAUD-ASSATHIAN<sup>1</sup>, Jaércio DA SILVA<sup>2</sup>, Cécile MÉADEL<sup>3</sup>

## Introduction

Amid an abundance of discourse on the effects of artificial intelligence (generative, algorithmic, and robotic), oscillating between alarm and enthusiasm (Bender & Hanna, 2025) how do professionals in the culture and creative industries navigate these often polarised positions? In this sector, and in the audiovisual field particularly, debate over the use of such tools has been intense. Actors and screenwriters, for example, have voiced acute concerns about automation and digital replication in the film industry, anticipating that these developments may proceed without effective oversight or financial compensation. These arguments were articulated, notably, during the 2023 Hollywood strike (Halperin and Rosner, 2025) and, more recently, in controversies surrounding the animated film *Critterz*<sup>4</sup>, slated for release in 2026 (which OpenAI reportedly intends to submit to Cannes). Such episodes have brought to the fore wide-ranging questions about the future of work in cinema and audiovisual production, as generative AI tools, such as ChatGPT and DALL·E, become readily accessible to the public and continue to improve in output quality.

These anxieties, often expressed as fears of the “replacement

---

1 Chaire PcEn/CES, Paris 1 Panthéon-Sorbonne University

2 Carism, Paris-Panthéon-Assas University

3 Carism, Paris-Panthéon-Assas University

4 Toonkel, Jessica. ‘Exclusive | OpenAI Backs AI-Made Animated Feature Film.’ Wall Street Journal, 8 September 2025, consulted on 11 October 2025 at the following address: <https://www.wsj.com/tech/ai/openai-backs-ai-made-animated-feature-film-389f70b0>.

of humans by machines”, partly stem from the imaginary of science fiction, and imaginary, that in a revealing paradox, is invoked precisely to assert its own obsolescence: “AI is no longer a science-fiction technology”. Such apprehensions attribute to these tools an almost autonomous, if not magical, capacity to act. The sociology of translation (actor-network theory) has long deconstructed the notion of machine autonomy, showing instead the “Seamless-Web Systems” (Hughes, 1979) through which innovation takes shape in a process that inseparably intertwines the technological and the social, an instance of “heterogeneous engineering” (Law, 1986). These approaches have also illuminated the gap between the programmes of action inscribed in machines, what Madeleine Akrich (2006) terms their scripts, and the dynamic, transformative character of their actual uses. Information and communication technologies, in particular, reveal a persistent tension between the intentions of designers and the open, plural and unpredictable environments in which these tools circulate, contexts marked by differentiated practices, repurposings, and appropriations, as well as formats and situations that, in turn, act upon the instruments themselves (Chateauraynaud and Lamy, 2025).

The integration of AI into work routines has provoked almost dramatic questioning across a wide range of sectors and industries (Brey, 2017). Virtually all forms of salaried labour are now reflecting on the implications of delegating tasks, modes of organisation, and even decision-making processes to artificial entities (Attencourt et al., 2025). This research<sup>5</sup> focuses on the cultural and creative industries, and more specifically on the audiovisual sector. It examines the narratives that professionals construct around generative artificial intelligence tools within the context of their professional practices. The analysis draws on publicly expressed discourses concerning these issues, situated within a

---

5 Research conducted as part of Styx, a project targeted by the PEPR ICCARE. Hosted by the PcEn chair at Paris-Panthéon-Assas and Paris1-Panthéon-Sorbonne universities, Styx receives government financial support under the France 2030 programme (ANR-23-PEIC-0006).

broader climate of change, emulation, and unease.

Our article identifies these narratives within a professional-oriented social networking platform: LinkedIn. This choice is grounded in the fact that the platform has established itself as a key space for public expression, where users engage through a diversity of formats, primarily text, but also videos, photographs, and graphic creations, and interact with one another on current issues, often to enhance or extend their visibility (Cardon, 2009, 2008). LinkedIn represents a professional communication arena, notably characterised by the overrepresentation of highly educated users (Bastin, 2015; Bastin and Francony, 2016), where various objectives intersect: networking, skills promotion, and continuous learning (Bridgstock, 2019). The platform is particularly widespread in environments where digital technologies are highly prevalent and intensively used (Rajkumar, 2022). This multi-purpose dimension makes LinkedIn a relevant field for observing how visibility discourses are constructed around technological innovations such as artificial intelligence (Dupont and Perticoz, 2016). Through LinkedIn, users articulate discourses of expertise, experiential narratives, job and service offers, news and event announcements, and professional imaginaries, thereby contributing to the shaping of representations of generative artificial intelligence, putting it into words, questioning it, and expanding its perceived possibilities within their respective fields of activity.

Given the diversity of tools and the breadth of professions and practices involved, this chapter, as previously noted, focuses on a specific segment of the cultural industries: the audiovisual industry, whose characteristics appear particularly relevant to our research question. It constitutes, first and foremost, a notably complex example of a market economy, characterised by high production costs, a continual pursuit of economies of scale, and a close interconnection between creation, technological innovation, and commercial logics. The audiovisual field also encompasses a wide range of professional groups —producers, screenwriters, writers, camera operators, editors, sound engineers, among oth-

ers— making it a fertile ground for analysis due to the multiplicity of perspectives it offers. Finally, this industry maintains a long-standing and intimate relationship with digitisation, which continually transforms and reorganises its production methods, modes of distribution, and economic models (Farchy, 2022).

### **Methodology**

Our analysis is based on a French-language corpus comprising 11,526 posts and 9,235 comments on those same posts, collected from LinkedIn using keywords related to artificial intelligence and audiovisual production<sup>6</sup>. Among these, 95% of the posts contain more than 100 characters, with an average length of around 1,400 characters, and none exceeding 3,000. It is not surprising that, within the specific “vernacular” of this platform (Gibbs et al., 2015), texts are not subject to formal constraints and tend to prioritise responsiveness in exchanges, hence being generally longer and more developed than those found on X/ Twitter. However, the topic of generative artificial intelligence appears to encourage authors to produce particularly elaborate and argumentative posts. The comments themselves can also be relatively substantial and, far from being limited to automatic reactions, go beyond mere expressions of approval or thanks (see Figure 1): nearly half exceed 50 characters, often including a link, a remark, a question, a piece of advice, or a contact request; and nearly 15% consist of several sentences forming a paragraph of at least 200 characters.

---

6 After exploratory work, the following equations were examined for 2024 and 2025: Audiovisual AI, AI cinema, AI video editing, AI screenwriting, AI creativity, AI digital platform, AI dubbing, AI video dubbing, AI content creator, AI-augmented creativity, AI film scriptwriting, AI visual storytelling, AI sound design, AI audiovisual transformation, generative animation, AI video, AI imaging, audiovisual innovation, cinema innovation, AI filmmaking, AI technology for creators.

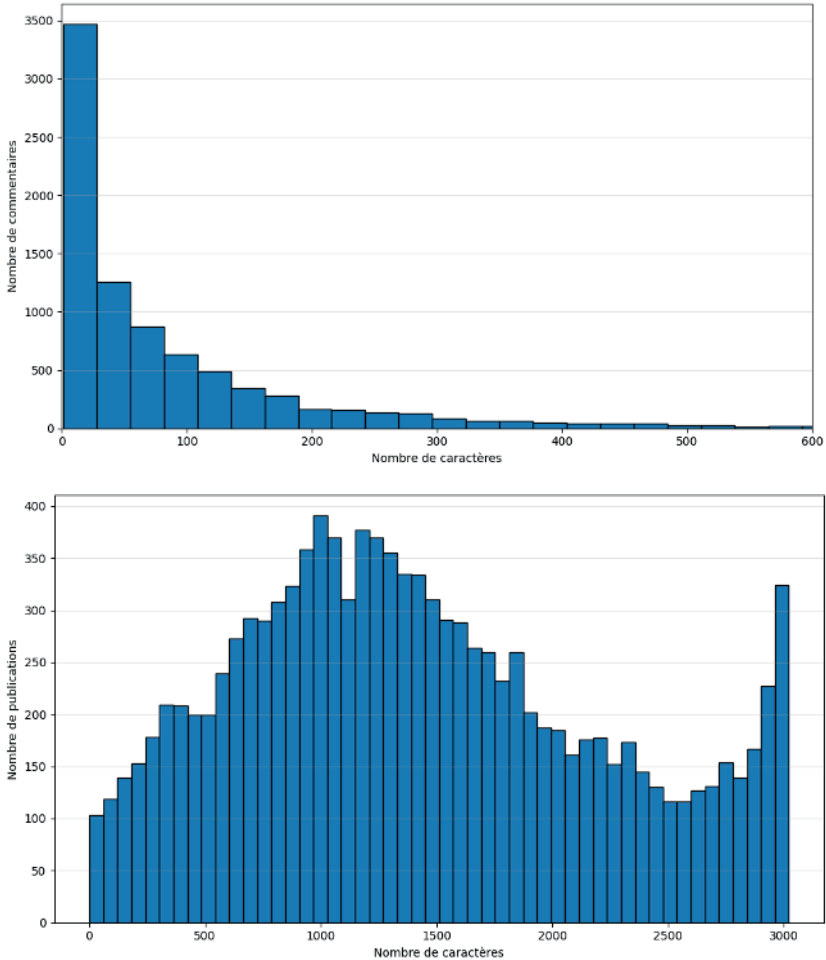


Fig. 1 Distribution of the number of characters per comment (1) and per post (2)

In total, 7,656 authors produced the posts included in the corpus. In 85% of cases, these authors contributed only a single post. The authors of both posts and comments were also analysed: we collected 8,298 “biographies,” that is, the self-descriptive texts users include on their LinkedIn profiles as personal introductions. These self-definitions vary widely, from simple professional titles such as “Computer Engineer”, or the more general

“Technology and AI specialist”, to brief mentions of their company, such as “PricewaterhouseCoopers (PwC)”. Others provide more elaborate descriptions of their occupations, skills, services, or professional trajectories, often blending French and English in what Laurence Allard (2017) terms a “hybrid digital writing style”, informally combining text, images, symbols, hashtags, emojis, and/or GIFs:

Webmaster, growth-hacker, **old web hand** 🤖♂ & creator of cash-machines. I boost revenues 📈 with #CRO, #SEO, #SEA, #SMA and plenty of tricks. #DigitalMarketing.

For some, it is not a self-presentation but rather a service offer:

A Commitment: 30% mini fewer problems in 3 to 6 months ✅  
 70/30 = A proven managerial balance ✅ A Human approach to Operational Excellence ✅ 200+ sites supported ✅ Founder of 7030.fr®.

The data on posts and author biographies, summarised in Figure 2, were collected in November 2024 and again in mid-May 2025. The posts were published between 19 October 2022 and 15 May 2025. In 2025, the number of comments collected is significantly lower, while the number of posts is higher; this discrepancy is due to LinkedIn’s policies, which restrict large-scale scraping. The collected data were pseudonymised in accordance with the GDPR: profile names and URLs were deleted after the collection phase, and only the information necessary for analysis (skills, experience, posts, and comments) was retained.

Categories	Total number
Collected posts	11 526
Comments collected from posts	9 235
Unique author biographies collected	8 298

Fig. 2 Summary of *corpus* data

The approach adopted follows the framework of *Digital Methods* (Rogers, 2013; Marres, 2017), which entails considering digital platforms both as research objects and as empirical fields, while taking into account their specific technical constraints and operational logics.

In the first stage, an initial corpus was constructed based on the search equation “AI+ audiovisual”. Qualitative analysis of this preliminary set then allowed the identification of other keywords recurrently used by LinkedIn users, which in turn broadened the scope of collection and nearly doubled the size of the original corpus. We observed that the number of posts retrieved per keyword varied greatly, with a median of 204 posts per term.

For data extraction, we used *Phantombuster* and *TexAu*, automation tools that execute scripts (known as *Phantoms*) to interact with web interfaces. One of these scripts was configured to systematically collect posts and their associated comments. However, automation does not imply that data collection is straightforward. LinkedIn enforces numerous technical restrictions: the maximum number of posts accessible per search URL is limited to one thousand. To address this limitation, we adopted the following procedure: for each selected keyword, we copied the corresponding LinkedIn search URL, connected a LinkedIn account through the automation tools, and executed the query. In some cases, the collection process extended over several days, as the first thousand results were retrieved daily. The data were then cleaned and structured using Python libraries (*Pandas* and *nltk*). Thematic analysis was conducted using a *Latent Dirichlet Allocation (LDA)* topic modelling approach with the *Scikit-learn* machine learning library, after text vectorisation with *CountVectorizer*. The results were visualised with *pyLDAvis*. Additionally, we examined word frequency distributions (using *Pandas* and *collections.Counter*) and generated *bigrams* and *trigrams* to analyse word co-occurrences.

Associated professions	Categories
artist, author, creative, content creator, design, digital design, music, video	Creative and artistic roles

project manager, director, boss, marketing, multi-activity	Managerial and strategic roles
teacher-researcher, coach-trainer-consultant, student	Academic and educational roles
communication, community manager, journalist	Communication and media roles
tech professions, webmaster	Technical and digital roles
administration, law, finance, HR	Administrative, legal and financial roles

Fig. 3 Summary of professional categories of post authors in the *corpus*

To understand who expresses themselves on LinkedIn, we grouped similar profiles according to the terminology they used, thereby identifying the professional roles associated with the authors of the posts and organising them into six categories (see Figure 3): creative and artistic; managerial and strategic; academic and educational; communication and media; technical and digital; administrative, legal, and financial. Given the diversity of self-presentation styles on the platform, these categories remain partially heterogeneous; nonetheless, they provide a useful overview of the professional landscape of users discussing generative artificial intelligence on LinkedIn. For the annotation of biographical descriptions, an initial sample of over one thousand profiles was manually classified into several thematic categories corresponding to their professional sector or group (see Figure 3), thus forming a training dataset. Based on these annotated data, we trained an automatic classification model using *DistilBERT* (from the *Transformers* library), which was subsequently employed to automatically assign a category to all other biographies in the corpus.

We also used the *text-mining* software Gargantext<sup>7</sup> (Delanoë

et al., 2023), which, through natural language processing and complex network analysis operations, enables the structuring of thematic clusters by analysing term co-occurrences—that is, the simultaneous presence of two or more words within the same statement. The tool identifies thematic sets (clusters) in the form of a network of co-occurring terms, optionally adding a spatial or temporal dimension.

Finally, we calculated the *Herfindahl–Hirschman Index (HHI)* to assess the degree of concentration of certain themes within the discourse, measuring whether a small number of actors or terms dominate the discussions. This approach does not rely solely on quantitative or computational analysis: the posts, comments, and biographies were also subject to close qualitative reading, allowing for a deeper understanding of the uses, perceptions, and obstacles observed within the corpus. Overall, the analysis emerged from an iterative process between qualitative readings and automated results.

## Results

The analysis of posts discussing artificial intelligence in relation to the audiovisual sector on LinkedIn reveals a discourse that is broadly positive, and in some cases even enthusiastic. Far from the forms of moral panic observed in other media spaces (Crépel and Cardon, 2022), these discussions are framed within a logic of valorisation: the term “*innovation*” ranks among the most frequently used, followed closely by “*creativity*.” One of the most recurrent arguments concerns the efficiency of AI tools, commonly presented as “*an innovation that enables greater speed and reduces repetitive tasks*,” thus “*freeing up time*” for activities perceived as more creative or strategic.

In the comments, specifically, in the subset of 62 comments that explicitly mention ChatGPT, enthusiasm is not universal: roughly one-third express concerns about data confidentiality, model efficiency, or the potential replacement of professionals.

Such concerns can be illustrated by the following comment: “*we are not all destined to become AI labourers.*”

Generative artificial intelligence thus appears in the posts as an operator of generalities: on LinkedIn, everyone who writes or reads these posts and comments is presumed to use it, mobilising it for a wide range of diverse and often diffuse purposes. This is, at least, the implicit assumption of the authors, who refer to artificial intelligence across highly varied topics, even when it is not central to their discussion. At the same time, LinkedIn itself functions as an operator of generalities insofar as it encourages overarching discourses in which authors propose a global view of the domains they address. Posts on the platform thereby become spaces through which users display their expertise, demonstrate their understanding of issues surrounding AI and, more broadly, technological innovation. Generative AI also seems to operate as a connector, a thematic anchor through which users expand their networks and increase both their readership and the visibility of their participation, as reflected in performance indicators such as likes, comments, views, and shares.

To complete the picture, it is worth noting the experimental nature of artificial intelligence. OpenAI, for instance, took fourteen months to publish its *Prompt Engineering Guide*, a manual designed to optimise results obtained from language models — a delay that no large-scale industrial product could afford without provoking criticism of irresponsibility (Legrand and Boullier, 1991). Other technologies have likewise been released to society for large-scale testing without their consequences or risks being fully anticipated, nor lingering doubts resolved. Ibo van de Poel (2017) uses the expression “learning-by-experimentation” to refer to the knowledge that emerges continuously during the experimentation of a technology. The discourse on AI observed on LinkedIn must therefore be situated within this unstable context of ongoing experimentation.

Within this context, marked by the experimental status of technologies, digital platforms tend to encourage ambiguous or simplified imaginaries (Bucher, 2018). The interventions of



The first two poles concern content creation. In the first (Pole 1), generative AI appears as a resource for content creation, encompassing discussions of tools and creative processes, the formats involved, and the various stages of production. The second (Pole 2) places greater emphasis on the economic dimension of content creation, focusing on major industry actors, spaces of circulation, and policy-related issues —such as progress, environmental considerations, inclusivity, and even ethics.

In a rather unusual way for this type of analysis, a third pole rearticulates and interconnects the themes of the first two in an entangled structure (see Figure 5). This cluster is heterogeneous, addressing simultaneously professions (*social media manager, webmaster*), content creation (*creator, creation, content, videos, image*), technologies (*Google, ChatGPT, deep learning*), economic aspects (*management, marketing, entrepreneurship, reduced cost*), forms of critique (*problems, bullshit, resistance*), and training (*learning, workshop, webinar*). It appears emblematic of the way LinkedIn users discuss generative artificial intelligence tools in relation to the cultural industries.

We hypothesise that, within the audiovisual field, these discourses reflect a moment of appropriation of a still unstable technology —a phase characterised by enthusiasm, curiosity, and a will to experiment, but also by uncertainty regarding the uses that are emerging or yet to be defined. While generative AI is broadly perceived as an “*unavoidable opportunity*,” its practical contours remain to be explored.

With this initial overview established, our results can be organised around three main axes.

First, we observe a strong preference for widely accessible, mainstream technological tools, as illustrated by the omnipresence of ChatGPT and Google. This indicates that users primarily discuss general-purpose solutions already integrated into numerous creative processes.

Second, in a context where academic training programmes in AI remain limited, and where the alarmist discourse of some academics regarding students’ use of generative AI, largely amplified and circulated by the media, continues to spread, new forms



## 1. The dominance of market leaders

In the field of audiovisual software, users tend to turn towards products offered by dominant market players, and only rarely highlight more specialised tools (see Figure 6). The authors mention a variety of AI systems in their posts. Unsurprisingly, ChatGPT overwhelmingly dominates in terms of the number of mentions, with its presence further increasing between 2024 and 2025—likely due to the wide range of functionalities it provides. Claude and Gemini are also cited with growing frequency, whereas Copilot remains relatively stable.

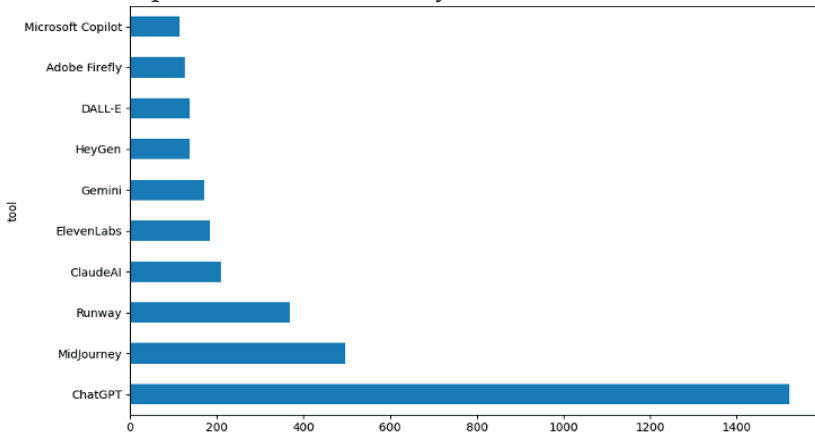


Fig. 6 Percentage of AI mentions in posts

Among more specialised AI tools, particularly those used for image or video generation, MidJourney is the most frequently mentioned, although its prominence is declining. It is followed by Runway, DALL·E, and HeyGen, which remain less well known to the general public.

Overall, these professionals, who identify themselves as specialists in audiovisual production, AI, or computing, refer primarily to the most popular and general-purpose tool, ChatGPT, and only marginally to more specialised ones<sup>9</sup>.

When looking more specifically at computing-related terms, we observe that programming languages such as Python top the list of mentions, alongside HTML, which, although not a programming language, remains the most commonly used language for creating web pages, particularly among non-specialists. More recent or specialised languages, such as JavaScript, Julia, or Rust, are mentioned far less frequently. This indicates a preference for versatile, widely used, and general-purpose tools (see Figure 7).

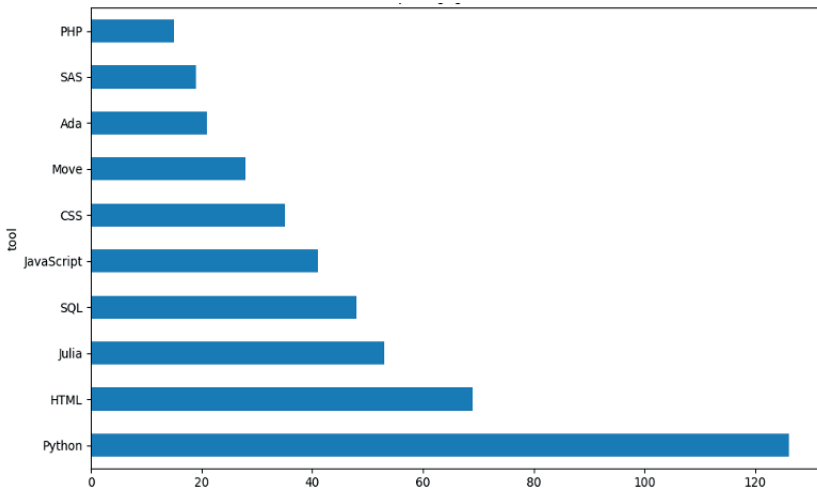


Fig. 7 Total number of mentions of programming languages in the posts

In conclusion, the strong presence of accessible, mainstream

---

Cinema and the Moving Image (CNC) reveals that the adoption of artificial intelligence tools in the audiovisual industry remains limited and uneven. Among professionals who use these tools, ChatGPT is by far the most widely used (82%), confirming its position as the main gateway into the world of AI. On the other hand, specialised tools, whether dedicated to writing, post-production, image generation or voice, are only used by 10 to 20% of respondents. Furthermore, nearly 45% of professionals say they have never tested an artificial intelligence tool, illustrating a still cautious adoption and a curiosity focused on the most accessible uses rather than on technical or business applications. Source: CNC Artificial Intelligence Observatory, document consulted on 11 October 2025 at the following address: [https://www.cnc.fr/professionnels/observatoire-de-lintelligence-artificielle\\_2390539](https://www.cnc.fr/professionnels/observatoire-de-lintelligence-artificielle_2390539).

tools in the analysed posts highlights a process of appropriation largely driven by non-specialised services. Ease of use, simplified access, and brand recognition appear to take precedence over the promotion of more technical solutions tailored to specific needs, particularly those related to audiovisual professions or emerging technological challenges. This trend can be partly explained by the high visibility of these tools, their wide online availability, their versatility, and their freemium access models, not to mention the abundance of tutorials that facilitate their adoption. However, it also indicates that tools specifically designed for the audiovisual field, such as *Genario*, mentioned only about thirty times, have yet to achieve widespread use or full legitimacy. Finally, this conclusion must be considered in light of the medium through which these discourses unfold: does LinkedIn itself not encourage the expression of more accessible, generalist approaches? The self-presentation of authors, which we now turn to analyse, seems to reinforce this hypothesis.

## 2. Lay knowledge and the evangelisation of AI

A textual analysis of how authors present themselves on the platform reveals a strong concentration of terms such as “digital,” “marketing,” “communication,” “manager,” and “expert” (see Figure 8). These terms correspond to generalist or overarching professional positions, with very few highly specialised experts, even though the cultural and creative industries (CCIs) are typically characterised by a great diversity of professions and service providers. The professionals in our corpus who express themselves on LinkedIn thus tend to adopt hybrid positions, combining elements of technology, leadership, and management. In the still highly unstable field of generative AI applications to the CCIs, those who speak out occupy multiple, overlapping roles that resist clear categorisation, relying instead on broad, cross-cutting skill sets. This reflects a broader pattern of career hybridity common in these sectors (Baillargeon & Cout-

ant, 2019), where professional qualifications are fluid and difficult to classify (Menger, 2014).

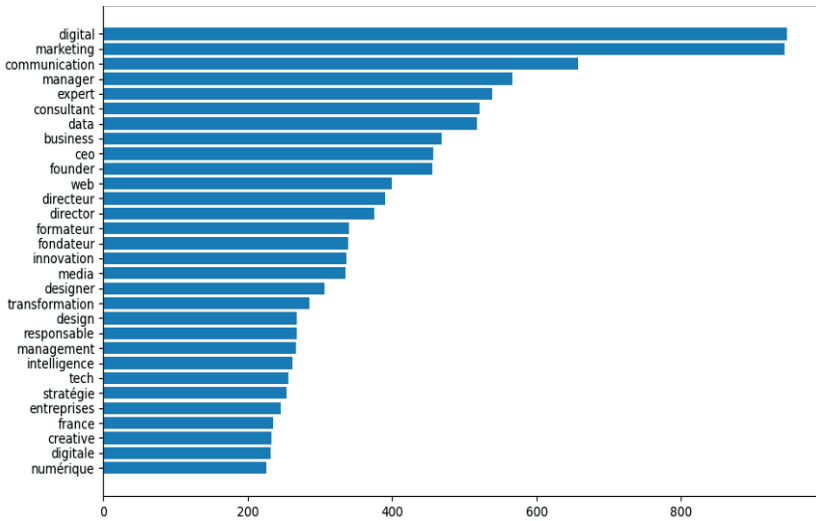


Fig. 8 Most frequently mentioned words in user biographies derived from the analysed posts

The messianic rhetoric surrounding the digital sphere finds a distinctive expression on LinkedIn through the promotion of informal, informal training programmes (Hoblingre & Audran, 2017). In a context where academic provision remains limited, and where the integration of generative AI (GAI) into traditional curricula is still weak, albeit widely debated, an alternative, less formal model is emerging online. This model introduces a new mode of legitimising expertise, bypassing traditional educational institutions and embedding itself within dynamics of visibility, engagement, and influence-seeking.

In our corpus, several private institutions, such as *DIXIT*, *Tech School of Business France*, *5 Formation*, *The Media Faculty*, and *Series Mania Institute*, promote their training offers (*workshops, thematic courses, conferences, study days, MBAs, etc.*) as rapid solutions to the urgent need to learn about AI for those working in audiovisual production. Among such initiatives, film schools like the

CLCF (*Conservatory of Film and Fiction (France)*) have introduced master's programmes focused on "*the opportunities opened by generative AI*," promising to "*understand, master, and exploit generative AI in your projects*", sometimes in the space of a single day. There are also public actors, such as *INA Campus*, which offers professional training sessions to help screenwriters familiarise themselves with *Genario*.

These programmes, often taking the form of short, intensive introductions (spanning a few days), masterclasses, or hands-on workshops, but also sometimes longer online modules lasting several weeks, share a common feature: the promotion of immediately applicable, practice-oriented skills:

What if you took your content to the next level? At DigiWeb, we help you make the most of HubSpot's full video potential.

Four weeks to explore the art of prompting, productivity, creativity—and to build your own AI agents.

Here are 9 game-changing AI tools that will save you time and boost your productivity.

These initiatives promote a learning logic centred on flexibility and immediate applicability, illustrating an approach to learning and digital acculturation that operates through synthesis and popularisation—that is, through the rapid digestion of complex technical content. For instance, ESAIP (*Engineering School for a Responsible Future, (France)*), a recognised but relatively low-profile engineering school, announces that it can, in a single day, "*reveal the inner workings of this revolutionary technology*." Others, such as La Salle in Metz, offer longer courses, though within more generalist educational settings. At the individual level, some of the LinkedIn authors in our corpus present themselves as initiators aiming to introduce professionals to the digital sphere through tutorials, thematic newsletters, and experience sharing:

As a professor of marketing and event communication, I am convinced that the integration of generative artificial intelligences

(AI) can transform the way we work. That's why, in my Web 4.0 courses, I explain how to use these tools intelligently—both effectively and creatively.

Join the newsletter to keep up with everything about graphic design, art, and mastering the Affinity software (you'll find the link in the comments). You'll get tips, tutorials, and plenty of inspiration to boost your creativity!

Last month, I revealed the scraping method used by Heads of Growth who follow Jordan Chenevier. Today, I'm offering the next phase for free (AI Qualification).

The professionals offering such training are not directly connected to, or at least do not identify with, the traditional professions of the audiovisual sector (such as screenwriter, editor, camera operator, production manager, or sound engineer). Instead, they emphasise the versatility of their career paths and the diversity of their skill sets, presenting themselves, for example, as marketing and innovation consultants, AI facilitators, or specialists in communication strategy and digital marketing.

We are thus confronted with forms of primarily operational knowledge, which appear to function without reference to any theoretical foundation, regardless of the task envisaged for generative AI. This raises the question of whether LinkedIn is becoming a fertile ground for the emergence of "evangelists" contributing to the construction of a messianic rhetoric surrounding digital technologies (Errecart, 2015). In our previous mapping of the development of the *metaverse* in the French public sphere, we highlighted the emergence of "consulting agents", self-proclaimed experts who seek to establish their legitimacy by positioning themselves as intermediaries between technological innovation and the general public, using social media and other dissemination tools such as podcasts to assert their authority and visibility (Da Silva and Méadel, forthcoming). These consulting agents appear to be highly active on LinkedIn, consistently adopting the role of promoting innovation through unquestioned technological tools and approaches.

### 3. AI as a gateway

In the posts from our corpus, AI is presented as a kind of bait with strong attractive power, as it is said to facilitate access to digital services that normally require demanding specialised skills. This is particularly noticeable in the case of desktop publishing (DTP) tools such as Adobe Photoshop, Illustrator, and Premiere Pro, which are among the most frequently mentioned tools in the posts analysed. Figure 9 shows the number of mentions of these art direction tools in the corpus. It was built from a database listing the commonly used DTP tools in artistic and graphic creation.

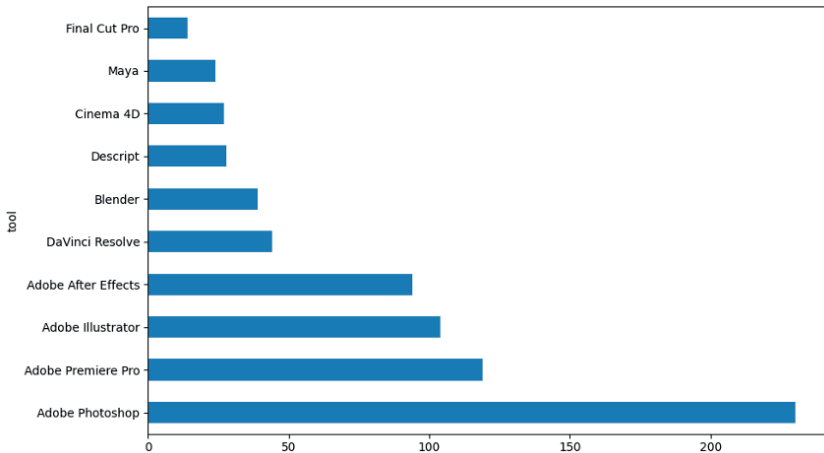


Fig. 9 Total number of mentions of art direction tools in the posts

Mastering software such as Adobe Photoshop or Adobe Illustrator requires specialised technical knowledge, involving not only specific training but also advanced computing skills and a form of craftsmanship. This mastery can be further strengthened by even basic knowledge of art history and visual arts. The posts in our corpus argue that generative AI tools (GAIs) facilitate access to these software environments and help lower the skill threshold required. They simplify access to functionalities that were once reserved for well-defined professional groups (camera

operators, photographers, editors, special-effects creators, etc.) working in more traditional sectors such as film, video games, animation, and television.

However, even though short-form videos have become an omnipresent format on social networks, their production has not erased the boundary between amateur and professional creation: full mastery of video production still requires specific skills in editing and post-production. Yet, tools integrating AI promise to reduce these entry barriers, granting access to advanced functionalities for less technically skilled users.

The automation of tasks such as image correction or format resizing is part of a broader trend towards the simplification of creative processes (Dubois and Bobillier-Chaumon, 2009), where users can delegate certain technical operations to artificial intelligence. This dynamic also contributes to the success of tools such as Canva and Adobe Express. Hutapea *et al.* (2024) show that, in the field of education, Canva functions as a tool of digital literacy, facilitating guided production of infographics, posters, presentations, animations, and videos.

This evolution aligns with broader transformations in information and communication technologies (Méadel, 2019), particularly those associated with the smartphone (Nova, 2019). The latter, by concentrating functionalities for recording, processing, publishing, and archiving images, sounds, and videos, has become a mobile and personalised production studio. The integration of AI into smartphone operating systems—such as Gemini for Android—illustrates this logic: these assistants allow users to automatically edit their photos through simple commands. AI thus becomes a marketing argument, by making previously restricted functionalities widely accessible.

Moreover, users who discuss AI on this platform tend to broaden the scope of use beyond the audiovisual sector alone, embedding their reflections within a wider framework of multimedia content creation using diverse digital tools. The analysed posts frequently reference professions such as community manager, graphic designer, or web designer, revealing a gradual shift

in the stakes of AI towards practices of visual communication, online community management, and digital design.

Finally, it is worth noting that the third cluster in our mapping also addresses voice cloning, in a similarly simplified and accessibility-oriented approach.

## **Conclusion and discussion**

This chapter examines how audiovisual professionals discuss and position themselves toward generative AI on LinkedIn. First, we showed the predominance of accessible, mainstream tools over specialised audiovisual systems, reflecting a logic of general technological literacy rather than sector-specific expertise. Second, we highlight the rise of hybrid professional identities and informal learning circuits, where self-proclaimed experts, consultants, and training providers contribute to the evangelisation of AI through short courses, tutorials and promotional narratives. Third, we observed how generative AI is framed as a gateway to advanced creative tools, lowering perceived entry barriers to software historically associated with specialised audiovisual crafts. Across these sections, a common dynamic emerges: generative AI is mobilised less as a deeply integrated production technology than as a symbolic resource for visibility, legitimacy, and adaptability within the sector.

Posts related to AI in the audiovisual field shared on LinkedIn are used primarily by users to signal their “professional skills” and their ability to keep pace with digital transformation. Users align their digital identities with widely recognised and accessible technological tools (such as ChatGPT and Python) rather than emphasising the transformations, contributions, or challenges associated with specialised software or creative projects. Overall, this gives predominance to well-known, easily communicable, and operational skills. It is therefore difficult to determine whether generative AI in the audiovisual sector is being concretely integrated into innovative projects, or whether its mention merely

reflects a performative stance, adopted by professionals wishing to display both their opinions and their up-to-date competencies. LinkedIn thus appears as a social networking platform for technological self-promotion, shedding light on broader cultural practices that shape professional identities today.

The discourses observed are generally positive in tone, sometimes tinged with concern, yet with a limited presence of critical or highly specialised perspectives, often giving way to simplified approaches and an implicit technological imaginary. The content of these posts oscillates between two main discursive regimes.

The first, generalist in nature, addresses generative AI in broad terms, associating it with promises of transformation or injunctions to “get involved”, on the model of “you have to dive in,” “don’t wait”, within a discursive register that often borders on the “technosolutionism” described by Morozov (2013). These discourses, frequently vague, tend to conflate diverse technologies, uses, and issues, reflecting a high level of generality and a certain confusion characteristic of the early stages of collective appropriation of a new technology.

The second regime, much more circumscribed but also rarer, focuses on a specific tool, API, or concrete application, often presented through training offers, tutorials, or service proposals. *Dubbing*, the subject of nearly 800 posts, constitutes a particularly revealing example: it is the focus of posts that detail usable applications, share experiences, address technical, legal, and social challenges, highlight advantages and drawbacks, and explain training modalities.

Taken together, these findings suggest that LinkedIn currently functions less as a space documenting mature adoption of generative AI in audiovisual workflow than as a platform where professionals publicly perform technological awareness and adaptability. The symbolic value of AI as a sign of creative agility, digital literacy, and professional modernity precedes its stable and consolidated integration into audiovisual production practices.

## References

- Afchar, D. (2023). *Interpretable Music Recommender Systems* [Doctoral dissertation, Sorbonne Université].
- Allard, L. (2017). "Creative sharing: Self-stylization and artistic experimentation." *Communication & Langages*, 194(4), 29–39.
- Akrich, M. (2006). *Les utilisateurs, acteurs de l'innovation*. In M. Akric, M. Callon, & B. Latour (Éds.), *Sociologie de la traduction* (p. 253-265). Presses des Mines.
- Aubert, A. (2021). "Information videos on social networks: Representing society through view metrics. A case study of videos from the media outlet Brut." *Questions de communication*, 2(40), 257–282.
- Baillargeon, D., & Coutant, A. (2019). "Atypicalities, hybridities and temporalities." *Communication & professionnalisation*, 7(1).
- Bartolome, A., & Niu, S. (2023). "A literature review of video-sharing platform research in HCI." *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, ACM, New York, Article 790, 1–20.
- Bastin, G., & Francony, J.-M. (2016). "Inscription, masking and data: The datafication of the web and interpretative conflicts around data in an invisible laboratory of the social sciences." *Revue d'anthropologie des connaissances*, 10(4), 505–530.
- Bastin, G. (2015). "Analyzing journalists' careers in the worlds of information." In C. Leteinturier & C. Frisque (Eds.), *Les espaces professionnels des journalistes. Des corpus quantitatifs aux analyses qualitatives* (pp. 203–228). Paris: Université Panthéon-Assas.
- Beaudouin, V., Bloch, I., Bounie, D., Cléménçon, S., d'Alché-Buc, F., Eagan, J., Maxwell, W., Mozharovskyi, P., & Parekh, J. (2020). "Flexible and context-specific AI explainability: A multidisciplinary approach." *SSRN Electronic Journal*, 1–66.
- Beaudouin, V., & Velkovska, J. (2023). "Investigating the ethics of AI." *Réseaux*, 240(4), 9–27.
- Bender, E. M., & Hanna, A. (2025). *The AI Con : How to Fight Big Tech's Hype and Create the Future We Want*. The Bodley Head.
- Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., & Ross, K. (2009). "Video interactions in online video social networks." *ACM Transactions on Multimedia Computing, Communications, and Applications*, 5(4), Article 30, 25 p.
- Bucher, T. (2018). *If... Then: Algorithmic Power and Politics*. Routledge.
- Cardon, D. (2008). "The design of visibility." *Réseaux*, 152(6), 93–137.

- Cardon, D. (2009). "Identity as a relational strategy." *Hermès, La Revue*, 53(1), 61–66.
- Chateauraynaud, F., & Lamy, J. (2025). "Algorithms and their ecologies: A sociology of the digital attentive to the materiality of the devices on which computational loops depend." *Socio*, 20(1), 17–40.
- Chateauraynaud, F. (2019). "A brief treatise on counter-artificial intelligence: A sociological look back at digital experiments." *Zilsel*, 5(1), 174–195.
- Da Silva, J., & Farchy, J. (forthcoming). "Transparency and explainability of AI: Social demands and challenges in the case of cultural industries." *Les Enjeux de l'information et de la communication*.
- Da Silva, J., & Méadel, C. (forthcoming). "Is technology soluble in podcasting? A case study on the metaverse." In *Le podcast natif: comprendre son essor, interroger son avenir*. Paris: Les Éditions Panthéon-Assas.
- Delanoë, A., & Chavalarias, D. (2023). *GarganText, collaborative and decentralised LibreWare*. ISCIPLIF GitLab repository, Complex Systems Institutes of Paris Île-de-France. (Online).
- Dubois, D., & Bobillier-Chaumon, M. (2009). "Technology acceptance: An analysis of determining factors." *Le Travail Humain*, 72(4), 305–327.
- Gibbs, M., Meese, J., Arnold, M., Nansen, B., & Carter, M. (2015). "#Funeral and Instagram: Death, social media, and platform vernacular." *Information, Communication & Society*, 18(3), 255–268.
- Halperin, B. A., & Rosner, D. K. (2025). "AI is soulless: Hollywood film workers' strike and emerging perceptions of generative cinema." *ACM Transactions on Computer-Human Interaction*, 32(2), 1–27.
- Hughes, T. (1979). "The electrification of America: The system builders." *Technology and Culture*, 20(1), 124–162.
- Hutapea, N. S., Manullang, Z. P. J., & Hartati, R. (2024). "Enhancing student engagement and academic performance through digital literacy: A transformative approach in Canva application." *Fonologi: Jurnal Ilmuan Bahasa dan Sastra Inggris*, 2(4), 154–170.
- Law, J. (1986). *Power, Action and Belief: A New Sociology of Knowledge?* Keele: Sociological Review Monograph.
- Legrand, M., Boullier, D., Sechet, J.-L., & Benguigui, C. (1991). *Between Humans and Machines: The User Manual*. Paris: IRIS.
- Marres, N. (2017). *Digital sociology: The reinvention of social research*. Polity Press.
- Martijn, M., Conati, C., & Verbert, K. (2022). "Knowing me, knowing

- you: Personalised explanations for a music recommender system." *User Modeling and User-Adapted Interaction*, 32(1–2), 215–252.
- Méadel, C. (2019). "A history of the user of information and communication technologies (ICT)." *Le Mouvement Social*, 268(3), 29–44.
- Menger, P.-M. (2014). "The expansion of artistic and cultural professions: Categorizations and mechanisms." *L'Observatoire*, 44(1), 7–19.
- Morozov, E. (2013). *To save everything, click here: The folly of technological solutionism*. PublicAffairs.
- Nova, N. (2020). *Smartphones: An Anthropological Inquiry*. Geneva: MédisPresses.
- Rajkumar, K., et al. (2022). "A causal test of the strength of weak ties." *Science*, 377, 1304–1310.
- Rogers, R. (2013). *Digital methods*. The MIT Press.
- van de Poel, I. (2017). "Society as a laboratory to experiment with new technologies." In D. M. Bowman, E. Stokes, & J. Rip (Eds.), *Embedding New Technologies into Society: A Regulatory, Ethical and Societal Perspective* (p. 404). Stanford in the Vale: Stanford Publishing.

# AI Ethnography: A Methodological Proposal for the Analysis of Vernacular Prompting Practices

by Gabriella TADDEO

## 1. Artificial Intelligence and the Methodological Turn

Artificial intelligence now occupies a central position in the infrastructure of everyday life and in contemporary knowledge production devices. This widespread presence requires the social sciences to rethink the analytical categories they use to describe agencies, mediations and processes of signification. Algorithms are increasingly understood as complex socio-technical assemblages, in which human and non-human actors, regimes of knowledge, organisational practices and material devices that support and direct action are intertwined (Gillespie 2016; Seaver 2017; Lange, Lenglet and Seyfert 2018). In this context, the methodological question becomes immediate: how can we observe and interpret social life when its expressions are filtered, produced or co-produced by computational procedures that often operate below the threshold of ordinary attention?

A well-established line of thinking emphasises the structural opacity of algorithms. The literature has shown that data selection chains, model optimisations and distribution strategies are difficult to understand for those who are not involved in their design. Often, observers can only access inputs and outputs, while the intermediate zone remains inaccessible or poorly documented, with significant consequences for the verifiability of explanations and the accountability of the actors involved (Pasquale 2015; Burrell 2016). In the face of this opacity, limiting analysis to observable performance alone risks reducing the phenomenon to a superficial level. The challenge is to make visible

the translation steps that underpin the functioning of the system.

Seaver (2017) observed that much of the work on AI tends to reproduce conceptual pairs that separate people and machines or culture and technology. These polarities unduly simplify ecologies of practice where actors influence and reshape each other (Airoldi 2021). A more useful approach considers the connections and synergies between heterogeneous elements, avoiding oppositional frameworks and valuing the relational character of socio-technical configurations. Latour's proposal goes precisely in this direction: to follow substitutions, associations and shifts of interest along the chains that make up scientific and technical objects, paying attention to moments when stabilisations are still in progress, when devices are not yet fully consolidated and their perimeter remains negotiable (Latour 1987). This type of approach explicitly places artefacts within circuits of actors, institutions, infrastructures and knowledge that support their production, circulation and legitimisation.

Within this horizon, ethnography returns to occupy a central role. Quantitative methods and large-scale measurements, based on computational analysis, provide useful images of phenomena, but struggle to convey the situated contexts and practical grammars through which people and systems encounter each other (van Voorst and Ahlin, 2024). Recent literature shows how overly positivist approaches tend to elide conditions of use, power asymmetries, and value regimes that structure interactions with AI (Rahwan et al. 2019; Adadi and Berrada 2018; Marda and Narayan 2021; Poell et al 2021). Ethnography, with its focus on practices, relationships and interpretations, allows us to piece together the picture, as it focuses on the local meanings and affective economies that emerge in the ordinary use of systems (Sartori and Theodorou 2022) and provides the tools to focus on the vernacular, relational and affective dimensions of algorithmic interaction (Marda and Narayan 2021; Sartori and Theodorou 2022; Barassi 2024).

Ethnographers have therefore focused on both the *production* and *reception* of algorithmic systems. Research on the production

side has shed light on the influence of professional norms, organisational configurations and work cultures which, especially in technology districts, guide design processes and define what is considered a good technical result. Studies on companies and development communities have highlighted models of horizontal organisation, project centrality, self-realisation rhetoric and competitive dynamics that shape metrics, objectives and engineering solutions (Noble 2018; Turner 2009; Marwick 2013). On the reception side, numerous contributions have shown how users develop adaptive practices and vernacular representations to navigate algorithmic logics (Bishop 2019; Bonini and Gandini 2019; Siles et al 2020; Ziewitz 2016). Awareness of opacity sometimes generates frustration, as in the context of platform work, where the mechanisms for assigning tasks and profile visibility are difficult to decipher and directly affect economic opportunities, to the point of creating experiences of automated management that are perceived as punitive (Rosenblat 2018; Bonini and Trerè 2024). At the same time, shared interpretative repertoires emerge, such as algorithmic imaginaries that guide expectations about the logic of the system (Bucher 2016; Baym 2018) and word of mouth among peers that consolidates operational advice and visibility tactics (Gillespie 2016).

A related line of inquiry concerns forms of relational interaction with conversational agents and generative systems. Several studies have therefore been conducted on so-called *social AI*, dedicated to analysing the relational affordances of generative models such as ChatGPT, but also dedicated conversational AI tools, such as Replika, Character.ai, and voice assistants (Depounti and Natale 2025). In these environments, users engage in exchanges that activate specific projections, roles and codes of conduct, with effects that touch on the emotional and identity spheres.

Christin (2020) proposes three operational steps suitable for using ethnography in algorithmic contexts: observing how algorithmic mediation reorganises social and institutional relationships; comparing sectors and platforms to isolate recurring

characteristics and relevant differences; using the same algorithmic tools to broaden access to the field and support forms of theoretical sampling, for example by exploiting recommendation systems to explore networks of content or connected actors. These strategies offer concrete ethnographic resources for circumventing opacity and connecting areas of the field that would otherwise remain invisible.

To situate this approach more precisely within the broader landscape of digital research, it is useful to draw on the tradition of *digital methods* as developed by Rogers (2017). This perspective emphasizes the methodological potential of studying digital media environments as epistemological infrastructures that can be repurposed for research. In this framework, the operational logics of the web—such as recommendation algorithms, tracking protocols, and profiling systems—are treated as native methodological resources. By aligning inquiry with the affordances and constraints of these systems, digital methods propose an embedded mode of investigation that leverages the medium's own mechanics while also rendering them visible and open to critique.

While digital ethnography focuses primarily on digital environments as contexts for social interaction, and platform ethnography studies the infrastructural and regulatory logics that govern digital platforms, the AI ethnography outlined here focuses on the semiotic, affective, and interpretative interaction between users and generative models. Incorporating this orientation into AI-focused ethnographic research allows for a more grounded understanding of how algorithmic systems structure user behavior and generate meaning. At the same time, it invites reflection on the methodological implications of adopting tools and procedures shaped by the same logics—raising questions about bias, visibility, and the conditions of individual creativity within computational environments. AI ethnography shares the focus on material and infrastructural conditions, but invests above all in the act of observing the practices as interpretative operations. The writing of the prompt into the AI systems, for example, in-

tertwines intentions, linguistic resources, aesthetic expectations, prior knowledge of the system, and hypotheses about the future behaviour of the model. Observing the work that takes place at this meeting point allows us to see how users articulate cultural sensibilities and regimes of meaning, and how these sensibilities are reorganised through interaction with AI. In this sense, Marda and Narayan's (2021) emphasis on the importance of linking statements to actions finds a meaningful application in the observation of prompting practices, where thoughts are materialized through the technology of writing and shaped by it—often more so than by the other affordances embedded in generative AI platforms.

This is the context for the AI ethnography proposal developed here within the “Prompting reflexivity” project. It is an experiment dedicated to observing how users of museums and cultural institutions use text-to-image AI to reflect on, define and expand the scope of their creative practices. The research proposal focused on the practices with which people use generative models, on negotiation with interfaces, and on the processes by which users define and redefine their creativity and reflexivity in the operational space opened up by generative text-to-image AI environments.

In the field of studies on creativity and computational art, and more recently on art generated with the aid of generative AI tools, the literature often tends to evaluate the final products and their similarity to the canons of human creativity (Natale and Henrickson 2022; Mazzone and Elgammal 2019). Much scientific attention has therefore been devoted to the reception of AI-generated products: for example, with the development of various methods and tests to measure the level of creativity of AI-generated products, or the public's perception of the quality of these products compared to artefacts generated by humans alone (Grassini and Koivisto 2024). An important concern underlying these works is to verify whether and how machines can be perceived as substitutes for human creativity, and what, therefore, their role is within creative economies and contemporary aesthetic practices.

Compared to this approach, which focuses mainly on advanced artistic fields and professional creative practices, my research has instead concentrated on observing vernacular practices, i.e. those of users who do not have a particular artistic background, are not involved in professional cultural production circuits and therefore use generative AI in conditions of spontaneous exploration, everyday micro-tasks and leisure time.

The objective of this inquiry was to develop a methodological approach capable of closely attending to the ways in which users interacted with generative text-to-image AI environments, with particular attention to the interpretative labor that surrounds each generated output. This includes the lexical adjustments made during prompting, the anticipations projected onto the system, the moments of disappointment or surprise, and the negotiations between the user's intended style and the aesthetic conventions inscribed in the model's training data.

Central to this investigation was the attempt to trace how individual creative trajectories unfolded in relation to, and were progressively shaped by, the specific affordances and constraints of generative systems. Rather than viewing outputs as static endpoints, the focus was placed on the dynamic interplay between user intention, system feedback, and the evolving semantics of the prompt—a process through which meaning is iteratively constructed within a technologically mediated creative space.

From this perspective, contributions on human-object interaction provided with an anchor that allowed me to consider generative systems as actants with which users entertain pragmatic and symbolic ties (Latour 2007; Pink et al. 2016). Callon's (1986) concept of enrolment has also proved particularly useful. Each prompting session involves alignments, translations and trials of strength that define roles and responsibilities among the people, models, interfaces and institutions involved. Describing these steps means following the negotiations that allow a certain arrangement to stabilise, sometimes only temporarily, and to produce results that are recognised as legitimate or desirable.

The aim of my ethnographic investigation was to explo-

re the interaction with generative AI along two interconnected dimensions. On the one hand, I conducted close observation of the participants' behaviors to capture the semiotic, operational, and emotional nuances of their engagement: subtle gestures during prompt formulation, moments of hesitation in response to unexpected outputs, recurring strategies of refinement, expressions of trust or suspicion toward the system's interpretative capacity, and the ways in which results were accepted, rejected, or negotiated in light of personal desires and imaginaries. In parallel to this observational work, I collected and analyzed the material traces of the interaction—namely, the prompts produced at each step of the process. These textual artifacts were treated as situated expressions of thought and intention, through which participants translated and reformulated their ideas in dialogue with the system. Developing an interpretive methodology for this corpus allowed me to trace the evolution of user reasoning, affect, and symbolic positioning as they unfolded across successive iterations. The prompts thus served both as analytical evidence and as a window onto the shifting semantics of creativity in algorithmic environments.

## **2. The AI Ethnography Setting: Participants, Methods, and Analytical Design**

The fieldwork that informs this reflection was carried out through a series of laboratory-based workshops conducted between 2022 and 2024 in schools, museums, and cultural institutions across northern Italy. In defining the characteristics of the participant sample, and in accordance with the ethical protocols established for the study and agreed upon with participants, all biographical and socio-demographic data were anonymized. Only general information—such as age, gender, and professional role—was recorded, to provide minimal but meaningful context for interpreting the observed practices.

A total of fifty-two participants, ranging in age from eleven to

fifty-five, took part in short, intensive sessions designed to explore the creative and interpretive dimensions of human–AI interaction. The sample was deliberately heterogeneous, including students, museum visitors, and educators, all of whom took part on a voluntary basis. Of the participants, twenty-six identified as female, twenty-four as male, and two as non-binary.

The ethnographic design adopted the model of “short ethnography,” as outlined by Pink and Morgan (2013): a temporally condensed yet deeply reflexive fieldwork strategy that prioritizes intensity and depth over duration. Each workshop became a dense ethnographic field in itself—a space where sensory, cognitive, and emotional aspects of human–machine interaction could be observed in real time. The limited timeframe heightened emotional involvement and made visible the improvisational nature of participants’ engagements with AI, an element often overlooked in long-term procedural studies. Despite the value of extended fieldwork, shorter periods are not inherently less insightful. As Pink and Morgan (2013) note, short ethnographies can yield moments of intense meaning and valid insight (Marcus and Okely, 2007; Vad Karsten, 2019).

Each workshop session lasted approximately two hours and followed a three-part structure. Participants were first introduced to the open-source text-to-image AI tool *Easy Diffusion*. They were then prompted to respond to a stimulus word—such as *otherness*, *happiness*, *anger*, *future*, or *friendship*—by generating images reflecting their personal interpretations of such concept. They were free to modify, iterate, or abandon their prompts as they wished, producing as many outputs as time allowed.

The workshop framework used text-to-image AI tools like *Stable Diffusion*, for activating creative elaboration and cultural reflection on a given topic. Text-to-image AI systems enable users to convert textual instructions (i.e., prompts) into images. These images are influenced by several parameters, including style descriptors (e.g., *gothic*, *brutalist*, *Art Nouveau*) and software settings that determine fidelity to the original prompt and level of detail.

While such AI tools are revolutionizing image production in both professional and artistic domains (Manovich and Arielli 2024; Oppenlaender 2023; Khutsishvili 2024), far less is known about how non-expert, everyday users engage with them. Therefore, the workshops employed image generation and visual creativity as means to activate visual thinking and foster identity- and culture-based reflection on selected themes.

Each session was thus dedicated to exploring a culturally sensitive concept and how it was interpreted and experienced by each participant. The use of AI was adapted by integrating the method of *photo-elicitation* into the participatory and creative context of AI-generated imagery. Participants' textual interpretations of each concept—iteratively reworked through sequences of prompts—served as a new method for collecting data and reflexivity, emotional involvement, and cultural negotiation about a sensitive topic.

The confrontation with the visual outputs generated by the system functioned as an additional stimulus, acting as visual elicitation to further deepen participants' reflections on the topic. For example, one workshop explored how participants imagine *the exotic*, *the distant*, or *the foreign*. Using both textual prompts and visual adjustments (style, lighting, color, etc.), participants were invited to generate and refine AI-produced representations of these issues. From this starting point, participants were encouraged to use text-to-image AI reflexively to generate contemporary visions of the exotic, co-constructed through dialogue with algorithms and mediated by interface affordances.

As Tota (2024) states, “we are what we see, and we see what we are.” Text-to-image AI enables us to imagine, create visual visions, and, through them, reflect on our present, our society, and ourselves. Participants were thus asked to produce personal visual elaborations of what they considered *the exotic*, *the other*, or *the distant*—using text-to-image AI iteratively, refining and reworking their images to sharpen their visions and deepen their critical engagement.

The aim of the workshops was not to promote AI as a tool for

trivializing or flattening cultural imagination—a potential risk if used uncritically—but rather to present it as a means for exercising cultural, identity, and creative agency through visual thinking (Arnheim, 2023). Visual language, in this context, enables forms of expressiveness not accessible through purely verbal or analytical reasoning (Ong and Hartley 2013).

Importantly, it is in the back-and-forth transition from text (prompt) to image -and back from image to revised prompt- that a new form of creativity and experience of the world can emerge. The images of the *other*, the *distant*, and the *exotic* generated during the workshops, for example, allowed participants to personally and contemporarily reimagine the concept. This facilitated critical reflection through both the exploration of media texts and inclusive creative elaboration, made accessible by the low technical barriers of AI tools.

In fact, comparing the results obtained allowed users to directly connect their imagination, defined and shaped by the words of the prompt, with the visual output, often the result of the incorporation of visual styles, stereotypes and cultural imagery absent from the user's original intent but present, and clearly visible, as evidence of the visual heritage and ideological affordances contained within the image datasets on which the AI model is based.

During the sessions, I alternated between roles—facilitator, observer, and interlocutor—collecting both textual traces of interaction (a corpus of 5,630 prompts) and field notes documenting gestures, reactions, and comments. At the end of each session, I held brief debriefing conversations in which participants reflected on their experiences, expressed moments of surprise or frustration, and described how they tried to “make the AI understand.” In each workshop, the generated images were shared and discussed collectively. I, as moderator, posed questions such as: *Did the images match your initial mental vision? If not, what was different or unexpected? Were there surprising or unintended visual elements that sparked new ideas or emotions?* This sharing and discussion process proved crucial for collecting interpretive data

about participants' textual and visual choices.

From a methodological standpoint, the analysis combined computational and ethnographic tools. On the quantitative side, the vast number of interactions was archived and systematized. Each user's textual flow of prompts was matched with the corresponding string of image outputs from the AI system. Using an open-source tool like Stable Diffusion allowed back-end access and anonymous archiving of session data, comprising all the prompts generated, time, and also the technical parameters and settings activated in each step of prompting. These quantitative records were complemented by field notes of participant interactions, helping to clarify emotional responses or intentions.

The whole material was thus analyzed using a combination of methods. Quantitative tools provided schematic summaries of each participant's semantic "journey" across prompts and their evolution throughout the session. To trace semantic changes between prompts, I applied cosine similarity—a linguistic metric that quantifies the degree of continuity or rupture in meaning across texts. Scores range from 0 (no textual overlap among the sequence of prompts produced by the user) to 1 (identical text along the prompts of the session). By calculating this metric for each participant's session, I could map whether their prompting behavior remained stable, adaptive, or discontinuous. I used AI itself—namely GPT-4—to create Python scripts that automatically computed semantic similarity across prompt sequences. Cosine similarity provided a quantitative mapping of individual semantic trajectories, indicating the degree of continuity or discontinuity in the formulation of prompts. These numerical values were placed within a qualitative framework in which in situ observations, participant comments, and post-session conversations allowed meaning to be attributed to the patterns detected. AI was therefore employed also as an analytical tool, capable of supporting interpretation through the processing of large textual corpora—in this case, the full set of prompts generated during the workshop sessions. While helpful, these quantitative indices were not central to interpreting creative trajectories but triangu-

lated with qualitative data from field notes and participant interview: without this contextual insight, changes in textual prompts would remain opaque.

This methodological synthesis —of metrics, notes, and reflexive engagement— constitutes the foundation of what I call AI ethnography: a multimodal, cross-temporal method for studying creativity and meaning-making in AI text-to image environments. Figure 1 summarises the proposed methodological approach, from data collection to analysis and interpretative assets.

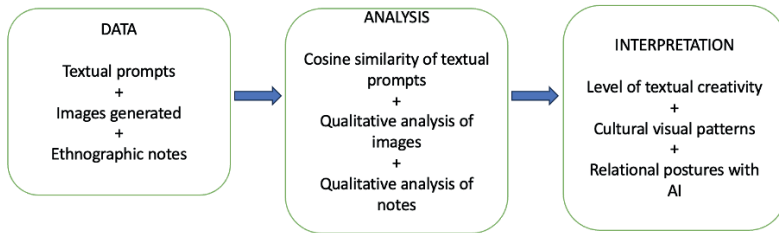


Fig. 1 Methodological path of AI ethnography

### 3. Observing the Process: An Ethnography of AI vernacular creativity

Analysis of the results revealed numerous useful elements for interpreting the socio-cultural trajectories of users. The workshops revealed a heterogeneous landscape of practices that defied neat categorization. Participants interacted with the AI in ways that reflected not only their technical skills, but also their broader cultural orientations toward technology, authorship, and creativity. Prompting emerged as a performative negotiation—part instruction, part conversation, part trial-and-error learning.

In some cases, participants approached the system with curious trust, using it as a playful generator of visual surprises. Children in particular often accepted the AI's suggestions without extensive modification, laughing at its bizarre combina-

tions or expressing delight in its unpredictability. Their prompts typically remained semantically stable, indicating a posture of acceptance toward the AI's interpretive logic. In one session, an eleven-year-old exclaimed, "It knows better than me what happiness looks like." This innocent yet telling remark encapsulates the displacement of agency characteristic of what might be termed *hegemonic decoding* (Hall 2007) -moments when users internalize the AI's aesthetic as natural or even superior to their own.

A passive acceptance of the visual outputs generated by artificial intelligence—characterized by minimal or entirely absent engagement in prompt refinement—was not limited to younger users. On the contrary, a similar orientation emerged across adult participants as well, pointing to the presence of a dominant interactional model in which the technological system is perceived as an autonomous and unquestionable authority in the image-making process.

In several cases, as also documented in field notes, adult users explicitly attributed the quality of the generated output solely to the software, displaying a tendency to delegate the creative initiative entirely to the system. This posture was often accompanied by a lack of interest in the expressive potential of prompt design, which was not approached as a space for intervention or co-construction, but rather as a technical step to be executed efficiently in order to obtain a visually satisfying product, without significant interpretive or experimental investment.

Such an attitude can be interpreted as a form of techno-cultural subordination, in which the algorithm's authority appears to neutralize the user's ability to critically negotiate the image-generation process. The underlying imaginary seems to reflect a conception of AI as a competent agent in its own right, before which the user's role is reduced to that of a mere activator or spectator of the result.

From an ethnographic perspective, this relational configuration between human and machine reveals not only a limitation in terms of creative engagement but also a culturally salient dynamic that intersects digital literacy, epistemic trust in technology,

and the internalization of externally driven logics of production. Identifying these behavioral patterns offers a means to critically interrogate the asymmetries that shape human-AI interaction and to explore the social and symbolic conditions that influence such orientations, beyond generational distinctions.

Both cases appear to suggest a possible correlation between certain age groups—specifically children around 11–12 years old and adults over 40—and lower levels of digital literacy, which are reflected in less articulated forms of interaction with generative systems. This condition seems to be associated with a less structured understanding of prompting as a creative practice, and with a more limited or less self-aware interpretation of one's role in the interaction with AI.

A reduced familiarity with the languages and operational logics of digital environments often translates into difficulty in recognizing the prompt as an active semiotic space through which users can exercise interpretive and design-oriented control over image generation. Within these age groups, the prompting experience tends to be configured more as consumption than as co-production, and the subjective contribution to the final output is often underemphasized or seen as secondary to the efficiency attributed to the system.

This observation does not imply a deterministic generalization but points to the importance of further investigating the intersections between age, digital competencies, and creative agency in AI interaction contexts. Such inquiry can illuminate how these factors shape users' creative posture and influence the ways in which they assign meaning and value to their participation in generative processes.

By contrast, many adolescents and some adults adopted more strategic, negotiating behaviors. They began with intuitive prompts, observed the resulting images, then revised their inputs to better match their internal vision. Some struggled to balance descriptive precision with creative openness-learning, for example, that overly specific prompts often produced clumsy, literal interpretations, while vagueness led to generic clichés.

This iterative movement between command and interpretation formed the dialogic heart of the workshop experience: a dynamic exchange of meanings across the semiotic threshold between human intention and algorithmic processing.

A particularly revealing moment occurred when a participant, aiming to represent the concept of freedom through the image of a street artist spray-painting a graffiti piece, noticed that all the AI-generated outputs featured a male protagonist. In response, he chose to explicitly specify a female subject in the prompt, deliberately “pushing” the system to produce a different kind of representation. As the participant explained, this choice was meant to evoke “an even greater idea of freedom,” since “you always see graffiti writers associated with men.”

In this brief statement lies a powerful cultural critique: an awareness that the model’s visual imagination reproduces gender bias embedded in its training data. Such episodes demonstrate how vernacular prompting can become a space for *micro-political reflection*, where users confront the cultural constraints of algorithmic representation and attempt -sometimes successfully, sometimes not- to resist them.

Other participants engaged in playful opposition. Rather than pursuing refinement or coherence through the prompts, they used the AI to create absurd juxtapositions: historical figures reimagined as superheroes, animals drinking beer, mythological scenes turned into cartoons. These acts of *semantic subversion* expressed a form of creative autonomy, asserting users’ freedom in the face of the model’s predictive logics. In these cases, the creative activity of prompting appears closely intertwined with the cultural consumption styles of the participants, particularly among younger users from Generation Alpha. The expressive strategies observed often involved a casual, sometimes deliberately incoherent use of language, evoking aesthetic tendencies linked to non-sense, digital Dadaism, and the hyper-saturated, fragmented dynamics of what is commonly referred to as *brain-rot humor* (Owens 2025). These forms circulate widely within social media environments frequented by this age group, where

absurdity, paradox, and visual excess constitute a recognizable cultural grammar.

The participants' approach to prompting suggests a culturally situated relationship with technology oriented toward open-ended experimentation rather than toward the production of structured outcomes. Interaction with AI took the form of a ludic and performative experience, where narrative coherence and productive finalization were often displaced by more exploratory, disjointed, and deliberately eccentric expressive modes. Within this framework, prompting functioned as a way to inhabit an alternative space—one where dominant logics of efficiency, optimization, and performance could be temporarily suspended. For these users, generative technologies offered an occasion to cultivate imaginaries shaped through excess, ambiguity, and the seemingly chaotic accumulation of visual and conceptual elements. Generativity was experienced as a process without fixed direction, where sense-making coexisted with paradox, breakdown, and surprise.

These dynamics reveal a different mode of engagement with artificial intelligence that resists standardized models of competence or technological agency. Prompting emerges as a situated expressive act, closer to symbolic play than to intentional design, and open to affective and semiotic experimentation that exceeds the boundaries of instrumental use. From an ethnographic perspective, this invites us to consider prompting not merely as a technical operation but as a cultural practice through which alternative relational modes with technology—and more broadly, with language and imagination—come into view.

Yet even in these moments of rebellion, the AI's aesthetic grammar remained present—hyperreal textures, cinematic lighting, symmetrical compositions—demonstrating that resistance was always partial, entangled with the very system it sought to escape.

Across all sessions, a clear continuum of co-agency emerged—a spectrum ranging from submission, to negotiation, to defiance—punctuated by moments of surprise, irritation, and discovery.

The creative process unfolded as a multilayered dialogue between competing imaginaries: the experiential world of the user and the statistical world of the algorithm and the often stereotyped imagery of the visual datasets. Ethnographic observation made visible the subtle ways users internalize, adapt to, or contest algorithmic aesthetics—how they recognize themselves (or fail to) in the images returned by AI systems.

The results observed during the workshops provide a detailed response to the initial research questions, highlighting how prompting practices are influenced by diverse relational postures towards AI (submission, negotiation, resistance), digital skills, and cultural attitudes towards technology. Diverse modes of creativity emerge: from passive acceptance of images, to strategic aesthetic experimentation, to playful subversion of the system's outputs. These dynamics allow us to understand how users attribute meaning to their creativity in an algorithmically mediated context.

The opportunity to directly observe participants during the various sessions revealed that the processes of dialogue, co-creation and technological reflexivity often took on an emotional dimension, activating dynamics of interaction and confrontation with the technological system that led to reflections on values, emotions and identity. The use of highly symbolic concepts within the workshops—such as happiness, friendship and the future—as an initial stimulus for creative production certainly facilitated the emergence of these dynamics. Added to this was a further component of involvement, attributable to the possibility of visualising representations of one's ideas in an immediate and accessible form through the technologies used. This combination of elements produced a significant emotional response across the board from participants, highlighting how interaction with automatic visual generation tools can trigger complex processes of personal and symbolic re-elaboration, even beyond strictly operational or functional purposes.

#### 4. Limitations and perspectives of the Method

AI ethnography is a research practice that, rather than offering definitive answers, opens up spaces for reflection on the ways in which people interact, attribute meaning and renegotiate their imaginaries in algorithmic environments. The reflective nature of this practice involves also the epistemic position of the researcher, who is called upon to confront his/her own interpretative tools, analytical categories and the very limits of understanding.

In this context, the use of quantitative metrics to analyse textual prompts—for example, through the analysis of lexical frequency, syntactic patterns or the semantic evolution of terms—can provide a useful contribution in heuristic terms. These tools make it possible to identify recurrences, formal variations and general trends within user-machine interactions, offering an initial mapping of the expressive behaviours and linguistic strategies employed. However, on an interpretative level, these data often prove insufficient to grasp the complexity of the transformations that take place in the creative process. In fact, metrics do not convey the situated meaning of semantic changes, nor are they able to account for the experiential and relational dimension of interaction, which develops over time and in the specific context of each workshop.

The integration of these analyses with qualitative methods of observation and listening—such as the analysis of field notes that emerged during the interviews and the participant observation—is therefore necessary in order to construct a more articulated understanding of the phenomenon. It is precisely in the triangulation between different levels of analysis that we can attempt to approach the complexity of prompting as a cultural practice, which cannot be reduced either to its technical component or to its textual surface. Prompting, in fact, takes the form of a space of negotiation between the user's intentions, the affordances of the system, and the unpredictability of responses, in which affective, aesthetic, and symbolic elements are deposited.

From this perspective, the use of quantitative tools should be

anchored to a theoretical framework capable of restoring prompting as an emerging, situated and relational process. Metrics, as in this case cosine similarity among the prompts, risk producing a flattened image of the creative process, eluding the deeper transformations that take place during the sessions: shifts in meaning, reconfigurations of identity, moments of impasse or revelation, which manifest themselves through language but cannot be reduced to it.

Furthermore, the laboratory environment in which the research takes place raises additional methodological issues. The structured and reflective nature of the setting can significantly influence the behaviour of participants, who tend to elaborate responses that are perceived as socially appropriate or culturally desirable. This type of reactivity, if not recognised, can lead to misleading interpretations. It is therefore essential to maintain a critical stance towards the methodological device itself, considering it not only as a space for observation but also as a context that co-produces the conditions of interaction and, consequently, the data collected.

A key element to consider concerns the position of the researcher, who played the dual role of workshop facilitator and ethnographic observer. This configuration has inevitable epistemological implications: the presence of the researcher and her guiding role may have influenced the course of the sessions and the participants' expressive choices. To mitigate this risk, a dialogical approach was adopted, in which moments of collective reflection and final debriefings allowed participants to verbalise their emotions, surprises and intentions. Furthermore, the ethnographic notes took into account the performativity of the context and the possible reactivity of the actors, considering them as an integral part of the data co-production process.

A possible development of the methodology could consist in broadening the field of observation to natural contexts of use, in which prompting practices emerge spontaneously and are not mediated by pedagogical or experimental purposes. In such environments, which are less constrained by the expectations of

the setting, different forms of relationship with the system may emerge, marked by other levels of urgency, ease or conflict. Access to these contexts, however, poses further ethical and logistical challenges related to privacy, platform variability, and the difficulty of tracing stable interpretative paths.

Ultimately, ethnographic research in algorithmic environments requires constant adaptation and methodological reflection, in which a variety of analytical tools—both quantitative and qualitative—are mobilised in order to achieve a critical and contextualised understanding of the cultural practices being observed. It is not a question of finding a definitive synthesis between computational models and cultural interpretation, but rather of inhabiting the field of tension between these two poles, accepting their incompleteness as an integral part of the cognitive process.

The proposed methodological framework, based on a hybridisation of brief ethnography, computational analysis and critical reflection, is designed to be adaptable to multiple research contexts. The ease of implementation of the tools used (open-source software, short workshops, participatory methodologies) makes it replicable in educational, museum, training and social research settings. Furthermore, the device can be extended to the observation of spontaneous practices in unstructured digital contexts, to explore everyday and unmediated uses of generative AI. This extension will require methodological recalibration—especially in terms of access and privacy—but opens up promising prospects for the study of algorithmic creativity as a widespread and situated cultural practice.

## References

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- Airoidi, M. (2021). *Machine habitus: Toward a sociology of algorithms*. Cambridge, UK: Polity.
- Arnheim, R. (2023). *Visual thinking*. University of California Press.

- Baym, N. K. (2018). *Playing to the crowd: Musicians, audiences, and the intimate work of connection*. New York, NY: NYU Press.
- Bishop, S. (2019). *Managing visibility on YouTube through algorithmic gossip*. *New media & society*, 21(11-12), 2589-2606.
- Bonini, T., & Treré, E. (2024). *Algorithms of resistance: The everyday fight against platform power*. Cambridge, MA: MIT Press.
- Bucher, T. (2016). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, 20(1), 30–44. <https://doi.org/10.1080/1369118X.2016.1154086>
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 205395171562251. <https://doi.org/10.1177/2053951715622512>
- Callon, M. (1986). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St. Brieuc Bay. In J. Law (Ed.), *Power, action and belief: A new sociology of knowledge?* (pp. 196–233). London, UK: Routledge.
- Christin, A. (2020). The ethnographer and the algorithm: Beyond the black box. *Theory and Society*, 49(5), 897–918. <https://doi.org/10.1007/s11186-020-09411-3>
- Depounti, I., & Natale, S. (2025). Decoding Artificial Sociality: Technologies, Dynamics, Implications. *New Media & Society*, 27(10), 5457-5470.
- Gillespie, T. (2016). #Trendingistrending: When algorithms become culture. In R. Seyfert & J. Roberge (Eds.), *Algorithmic cultures: Essays on meaning, performance and new technologies* (pp. 64–87). New York, NY: Routledge.
- Grassini, S., & Koivisto, M. (2024). Artificial creativity? Evaluating AI against human performance in creative interpretation of visual stimuli. *International Journal of Human–Computer Interaction*, 41(7), 4037-4048 <https://doi.org/10.1080/10447318.2024.2345430>
- Hall, S. (2007). Encoding and decoding in the television discourse. In S. Hall, D. Hobson, A. Lowe, & P. Willis (Eds.), *CCCS selected working papers* (pp. 402–414). London, UK: Routledge.
- Khutsishvili, K. (2024). AI creativity and human enhancement: The identity link. In M. Coeckelbergh et al. (Eds.), *Artificial intelligence, co-creation and creativity* (pp. 147–158). London, UK: Routledge.
- Lange, A.-C., Lenglet, M., & Seyfert, R. (2018). On studying algorithms ethnographically: Making sense of objects of ignorance. *Organization*, 26(4), 598–617.

- Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Cambridge, MA: Harvard University Press.
- Latour, B. (2007). *Reassembling the social: An introduction to actor-network theory*. Oxford, UK: Oxford University Press.
- Manovich, L., & Arielli, E. (2024). *Artificial aesthetics*. <https://manovich.net/index.php/projects/artificial-aesthetics>
- Marda, V., & Narayan, S. (2021). On the importance of ethnographic methods in AI research. *Nature Machine Intelligence*, 3, 187–189.
- Marcus, G. E., & Okely, J. (2007). How short can a fieldwork be? *Social Anthropology*, 15, 353–357.
- Marwick, A. (2013). *Status update: Celebrity, publicity, and branding in the social media age*. New Haven, CT: Yale University Press.
- Mazzone, M., & Elgammal, A. (2019). Art, creativity, and the potential of artificial intelligence. *Arts*, 8(1), 26. <https://doi.org/10.3390/arts8010026>
- Natale, S., & Henrickson, L. (2024). The Lovelace effect: Perceptions of creativity in machines. *New Media & Society*, 26(4), 1909–1926.
- Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York, NY: NYU Press.
- Ong, W. J., & Hartley, J. (2013). *Orality and literacy*. London, UK: Routledge.
- Oppenlaender, J. (2022, November). The creativity of text-to-image generation. In *Proceedings of the 25th international academic mindtrek conference* (pp. 192–202).
- Owens, E. (2025). “It speaks to me in brain rot”: Theorising ‘brain rot’ as a genre of participation among teenagers. *New Media & Society*. Advance online publication. <https://doi.org/10.1177/14614448251351527>
- Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Cambridge, MA: Harvard University Press.
- Pink, S., Ardèvol, A., & Lanzeni, D. (2016). *Digital materialities*. London, UK: Routledge.
- Pink, S., & Morgan, J. (2013). Short-term ethnography: Intense routes to knowing. *Symbolic Interaction*, 36(3), 351–361. <https://doi.org/10.1002/symb.66>
- Poell, T., Nieborg, D. B., & Duffy, B. E. (2021). *Platforms and cultural production*. Hoboken, NJ: John Wiley & Sons.
- Rahwan, I., et al. (2019). Machine behaviour. *Nature*, 568, 477–486. <https://doi.org/10.1038/s41586-019-1138-y>

- Rogers, R. (2017). Digital methods for cross-platform analysis. In J. Burgess, A. Marwick, & T. Poell (Eds.), *The SAGE handbook of social media* (pp. 91–110). London, UK: SAGE.
- Rosenblat, A. (2018). *Uberland: How algorithms are rewriting the rules of work*. Berkeley, CA: University of California Press.
- Sartori, L., & Theodorou, A. (2022). A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. *Ethics and Information Technology*, 24(1), 4.
- Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, 4(2), 205395171773810. <https://doi.org/10.1177/2053951717738104>
- Siles, I., Segura-Castillo, A., Solís, R., & Sancho, M. (2020). Folk theories of algorithmic recommendations on Spotify: Enacting data assemblages in the global South. *Big Data & Society*, 7(1), 2053951720923377.
- Tota, A. L. (2024). *Eco-thoughts: Conversations with a polluted mind*. London, UK: Routledge.
- Turner, F. (2009). Burning Man at Google: A cultural infrastructure for new media production. *New Media & Society*, 11(1–2), 73–94. <https://doi.org/10.1177/1461444808099575>
- Karsten, M. M. V. (2019). Short-term anthropology: thoughts from a fieldwork among plumbers, digitalisation, cultural assumptions and marketing strategies. *Journal of Business Anthropology*, 8(1), 108–125.
- van Voorst, R., & Ahlin, T. (2024). Key points for an ethnography of AI: an approach towards crucial data. *Humanities and Social Sciences Communications*, 11(1), 1–5.
- Ziewitz, M. (2016). Governing algorithms: Myth, mess, and methods. *Science, Technology, & Human Values*, 41(1), 3–16. <https://doi.org/10.1177/0162243915608948>



# **AI-Augmented Anticipatory Ethnography: Envisioning, Design Fiction and Generative AI for Co-Creating Eutopias**

by Agnese VELLAR, Matteo FOGLI

## **AI Disclaimer**

The authors use generative artificial intelligence broadly across research and writing activities, guided by two complementary dimensions: AI Literacy, as defined by the EU AI Act (understanding principles, ethics, and recognizing risks and opportunities), and AI Fluency, per Anthropic (strategic interaction through delegation, description, discernment, and diligence).

Generative AI has been employed as a co-creation tool to expand analytical perspectives, synthesize literature, refine arguments, and improve expository clarity. The authors maintain full responsibility for methodological, interpretative, and ethical choices, ensuring that every claim is validated through authoritative sources and critical reflection, in coherence with the creative co-intelligence approach proposed in the paper itself.

## **1. Imagining Preferable Futures in the Era of Technological Acceleration**

The advent of generative artificial intelligence has accelerated technological processes with significant societal impact, increasing complexity and volatility. Researchers, designers, entrepreneurs, and leaders in HR, transformation, and innovation now face new opportunities and challenges that require updated methods and skills. In an era where uncertainty is the only constant, the ability to imagine alternative scenarios and critically

interrogate the present is crucial for organizations and professionals who develop future visions.

Those who study cultures can become privileged observers of emerging dynamics. With the right tools, they can detect signals from the future. For sociologists and ethnographers, the task is no longer simply to observe the present, but to intercept weak signals, latent tensions, and unexpressed desires that prefigure possible futures and guide them toward preferable ones.

This role can be fulfilled by interdisciplinary groups that bring together design anthropology (Sampson, 2021), futures thinking (McGonigal, 2022), and design fiction (Dunne & Raby, 2013). Design anthropology helps design innovative, people-centered, inclusive solutions. Futures thinking imagines non-existent scenarios and invents new possibilities, while design fiction constructs worlds and narratives to help people experience them as if they were real.

From this perspective, the future becomes not something people wait for or endure, but:

a project [that starts from] recognizing underlying trends that require adaptation or transformation of human life through innovation. And acting accordingly. With Method (De Biase, 2024, p. 34).

For Luca De Biase, this method is Futures Design Thinking, born from the intersection of Design Thinking and Futures Thinking. It involves three stages (De Biase, 2024, p. 300):

1. analysis of current trends
2. multi-stakeholder discussion on possible ideas and their consequences
3. prototyping ideas, defining results, and collecting feedback

The ability to prototype future scenarios is enabled by speculative design and design fiction, which make the future not only thinkable but experienceable, stimulating public debate and opening spaces for co-creating preferable worlds.

Design thus transcends its purely functional dimension to assume a central role in social imagination and world-building. It becomes a collective process that interweaves strategy, technology, art, and futures studies.

How can we combine the depth of ethnographic observation with the imaginative power of fiction? What role can artificial intelligence play in this process? This contribution presents “AI-Augmented Anticipatory Ethnography” as a methodological frontier for social research and innovation. By integrating human creativity and artificial generativity, this approach enables:

- new ways of thinking to address uncertainty, complexity, and accelerating change
- new methods of empirical research and inclusive, participatory design suited to rapid technological innovation

The goal is to orient change toward desirable scenarios, leveraging all available agency or developing new agency through Futures Design Thinking methods. This multi-phase approach begins with cultivating AI Fluency and futures literacy and moves through speculative design to participatory imaginative co-creation.

## **2. Futures Thinking for a NAVI World**

### *2.1 AI Fluency and Futures Literacy for Navigating a NAVI World*

For decades we have lived in a VUCA world (Volatile, Uncertain, Complex, Ambiguous), which has evolved into NAVI (Non-linear, Accelerated, Volatile, Interconnected) as defined by EY (2025).

NAVI’s Non-linear dimension amplifies volatility through sudden tipping points that disrupt markets. Its Interconnection characteristic intensifies complexity: geopolitical crises, climate change, and technological disruptions act inseparably. The system is in constant Acceleration. Artificial Intelligence acts as a multiplier, catalyzing the transition from VUCA to NAVI. According to the European Commission (2024), an AI system is:

a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments (EU, AI Act, 2024, Art 3(1)).

AI systems are the engine of the NAVI world. They:

- accelerate volatility through continuous innovation and tipping points
- generate epistemic uncertainty due to their complex, opaque decision-making (black box)
- intensify complexity through human-AI interactions
- introduce semantic ambiguities in outputs

AI not only reflects VUCA characteristics but amplifies them into NAVI, posing new challenges for creators, researchers, and users.

The World Economic Forum (2023) identifies technological acceleration, economic uncertainty, and geopolitical fragmentation as forces redefining production processes and competencies. While AI, robotics, and automation increase demand for digital literacy, global instability makes soft skills essential: continuous learning, creative thinking, mental agility, and resilience. These are among the fastest-growing competencies expected by 2030.

Two complementary dimensions emerge: AI Literacy and AI Fluency. AI Literacy, according to the EU AI Act, comprises understanding AI principles, ethical impacts, and recognizing risks and opportunities:

AI literacy should equip providers, deployers and affected persons with the necessary notions to make informed decisions regarding AI systems. Those notions may vary with regard to the relevant context and can include understanding the correct application of technical elements during the AI system's development phase, the measures to be applied during its use, the suitable ways in which to interpret the AI system's output, and, in the case of affected persons, the knowledge necessary to understand how decisions taken with the assistance of AI will have an im-

pact on them. In the context of the application this Regulation, AI literacy should provide all relevant actors in the AI value chain with the insights required to ensure the appropriate compliance and its correct enforcement. Furthermore, the wide implementation of AI literacy measures and the introduction of appropriate follow-up actions could contribute to improving working conditions and ultimately sustain the consolidation, and innovation path of trustworthy AI in the Union.” (EU, AI Act, 2024, Art 4).

AI literacy is both technical and civic, protecting rights and promoting democratic control. But literacy alone is insufficient. AI Fluency, per Anthropic (2025), is a more advanced competency involving strategic interaction with AI, effective prompting, critical output evaluation, and responsible action. AI Fluency involves four capabilities:

- delegation: deciding when to involve AI while maintaining strategic control
- description: formulating clear, contextualized instructions
- discernment: critically evaluating AI outputs for errors and biases
- diligence: using AI ethically and responsibly

AI Literacy and AI Fluency together enable active digital citizenship and evolved professionalism, shifting users from passive to active roles while integrating rights protection and human agency.

Beyond these operational competencies, Futures Literacy (UNESCO) offers tools to imagine alternative scenarios. It invites exploration of different futures to break free from limiting narratives. Learning to use the future as a critical lens enables recognition of opportunities and alternatives that would otherwise remain invisible. Futures Literacy develops through participatory learning processes that explore multiple future narratives, fostering appreciation of diversity and comfort with uncertainty. It strengthens agency and empowerment, enabling individuals and communities to become active protagonists in constructing desirable futures.

## *2.2 The Role of Futures Studies and Different Types of Futures*

In a context of growing uncertainty accelerated by AI, developing the capacity to imagine multiple futures with an optimistic yet critical perspective is crucial. UNESCO emphasizes future-proof thinking literacy, particularly relevant for those developing innovative projects: public administrations, R&D companies, researchers, and designers.

Futures Studies, drawing from over fifty years of applied research, particularly from the Institute for the Future in California, offer rich methodologies for cultivating a future-positive mindset and co-designing desirable scenarios.

Three fundamental principles underlie Futures Studies: the future is plural, not singular; “futures thinking” can be exercised and learned; and everyone has agency to orient the future toward what they deem preferable:

Thinking about the future is also about imagining. It’s about transforming how we think. It’s about creating a map to the future and looking for the big areas of opportunity. We like to think about transformations, for example, in learning and work, and how they get connected and intertwined in various ways. And then we start thinking about zones of opportunity. How can we shape the future to make it more equitable? How can we amplify learning outcomes? What do we need to do to achieve these outcomes? The future doesn’t just happen to us. We have agency in imagining and creating the kind of future we want to live in, and we can take actions to get us there. (Gorbis, 2019, p. 30)

To map multiple futures, Hancock and Bezold (1994) propose the Futures Cones model with four types:

- possible: anything that could happen, without limits imposed by current reality
- plausible: what is realistic or credible, considering available knowledge and technologies
- probable: what is most likely to happen, based on current trends and data
- preferable: those we would want realized, as collective intention

McGonigal (2020) adds “preferred” futures, what an individ-

ual or group wants for themselves. While “preferred” futures concern individual or small team desires, “preferable” futures have a broader social connotation oriented toward common good, considering ethics, social values, and sustainability.

### *2.3 Developing Futures Thinking and Urgent Optimism*

Jane McGonigal (2020, 2022) centers her methodology on the vision that everyone can actively shape their preferred future. Her approach prepares people to face uncertainty and change, providing practical tools for personal transformation and developing a future-positive mentality. McGonigal has developed Futures Thinking tools that enable people and teams to exercise key competencies for navigating continuous change:

- **continuous Learning:** cultivating constant curiosity and actively seeking new knowledge and skills
- **agility:** adapting rapidly to changes and acting effectively in uncertain situations
- **creativity:** imagining what doesn’t yet exist and inventing new possibilities
- **optimism:** believing today’s actions can positively influence the future
- **critical thinking:** maintaining a lucid, realistic vision
- **empathy:** understanding people’s hopes and fears

McGonigal particularly explores Urgent Optimism as a mindset that combines future anticipation with motivation to act immediately, even amid obstacles:

Urgent optimism doesn’t mean staying up all night worrying. It means jumping out of bed each morning with fire inside, ready to act. Urgent optimism is knowing you have unique agency, talents, skills, and life experiences to create the world you desire. (McGonigal, p. 30, 2022)

Urgent optimism is a balanced sentiment recognizing future challenges while maintaining lucid confidence (McGonigal, 2021). It comprises:

- **psychological flexibility:** rapidly adapting to new sce-

narios, embracing change as constant

- **realistic hope:** confidence in influencing the future while recognizing risks and limits
- **future power:** agency, the conviction that one's actions can generate concrete impact

If urgent optimism is the prerequisite for thinking like a futurist, the mental process activates when “time spaciousness” emerges. The awareness of having sufficient time for what truly matters favors reflection, planning, and conscious action, enabling the shift from “first-person imagination” to “third-person imagination” projected into the future (typically 10 years). When episodic future thinking is active, the brain changes perspective and temporarily opens the mind to discovery. This isn't escapism but a way to interact more deeply with reality, exploring otherwise invisible risks and opportunities. It helps us ask crucial questions: *Is this the world I want to wake up in? What do I need to be ready? Can I change something today to make this future more or less probable?*

The next step is imagining a future scenario, a detailed description of a possible world where at least one element differs radically from today:

A specific story set in a future forecast. It describes what we might see, feel and experience if we woke up in that forecast. A scenario describes the future as if it were already real. It can take the form of a short story, a news article, a comic, a film, a documentary from the future.... any form you can use to tell a story, you can use to share a scenario. Scenarios are important because they help us imagine the future more vividly. They give us concrete possibilities to evaluate – do I want this future? What would I do in this future. (McGonigal, 2020)

Having developed and empirically tested methods and techniques for exercising a futures-positive mindset, urgent optimism, and developing future scenarios, Futures Thinking emerges as a fundamental discipline for addressing contemporary complexity, uncertainty and acceleration. Developing agency and orient-

ing visions toward desirable future scenarios means overcoming linear prediction and embracing multiplicity. Futures thinking becomes a design compass, orienting the future not only toward what is achievable but especially toward what is desirable.

### **3. Anticipatory Ethnography and Design Fiction**

#### *3.1 Speculative Design and Design Fiction: Imagining Utopia*

Futures Design Thinking is a multi-phase approach: in the first phase, methods and tools from Futures Studies cultivate futures literacy and urgent optimism, enabling agency, imagination and a future-positive mindset; in the second phase, the approach adopts Speculative Design methods to conduct research and design future scenarios. In Speculative Design, the designer's goal is not to create products and services that address specific needs or solve problems, but to facilitate collective processes and mediate among needs, values, and future visions. Anthony Dunne and Fiona Raby (2013), leading speculative design theorists, invite us to move beyond problem-solving toward problem-finding: design should not only respond to existing needs but also raise questions, challenge assumptions, and open spaces of possibility. Design practice becomes "social dreaming", creating a collective dream nourished by imagination, empathy, and the capacity to critically interrogate the status quo:

As we rapidly move toward a monoculture that makes imagining genuine alternatives almost impossible, we need to experiment with ways of developing new and distinctive worldviews... If our belief systems and ideas don't change, then reality won't change either. To be effective, the work needs to contain contradictions and cognitive glitches. Rather than offering an easy way forward, it highlights dilemmas and trade-offs between imperfect alternatives. Not a solution, not a "better" way, just another way. This is where we believe speculative design can flourish—providing complicated pleasure, enriching our mental lives, and broadening our minds... It's about meaning and culture, about adding to what life could be, challenging what it is, and providing alternatives that loosen the ties reality has on our ability to

dream. Ultimately, it is a catalyst for social dreaming. (Dunne & Raby, 2013, p. 189)

To open dialogue around possible worlds, we must make the future tangible and experienceable. Design fiction combines prototyping, narration, and critical reflection to transform future scenarios from abstractions into concrete experiences. Julian Bleecker (2009), a design fiction pioneer, defines it as the intentional creation of “diegetic prototypes”: objects, documents, artifacts that exist only within a narrative but invite observers to suspend disbelief and imagine the world they belong to.

These objects function as theatrical props: the point is not believing they’re real but allowing imagination to explore the social, cultural, and ethical implications of the innovations they represent. Exhibitions, installations, publications, videos, and digital content become spaces for dissemination and confrontation, where the future becomes an object of public discussion and collective negotiation. Design fiction is not only a tool for visualizing change but a device for enabling “social dreaming,” making reality more malleable and fostering multiple micro-utopias.

Design fiction draws on literature, cinema, art, philosophy, and social sciences to construct fictional worlds, utopias, dystopias, thought experiments, and counterfactuals. The intention is to “misalign” thinking, stimulate imagination, and spark public discussion. The ambiguity between real and unreal becomes a resource: keeping the field of possibilities open means avoiding both pure realism and pure fantasy to explore the gray zones and uncertainties that traditional methods often ignore.

Design fiction is useful for exploring and communicating future scenarios, but Dörrenbächer *et al.* (2020) note it often privileges dystopian or ironic narratives, risking reinforcement of negative, unconstructive imaginaries. In other cases, it closes into an elitist authorial dimension, with designers as sole protagonists and participants relegated to passive spectators. To overcome these limitations, Dörrenbächer *et al.* propose an approach oriented toward positive, participatory, contextualized design fiction. In this perspective, participants are no longer mere re-

ipients of pre-packaged scenarios but co-authors of utopias and “from within” evaluators of lived experiences. Dörrenbächer *et al.* propose a three-phase co-creation process:

- **imagining utopia:** participants project themselves into sustainable, positive scenarios, identifying emotions, values, and technologies that would make the future desirable
- **enacting utopia through co-construction:** materializing utopia through roles, contexts, and concrete actions. Participants act in the fictional world, negotiating conflicts and experiencing social dynamics
- **evaluating utopia from within:** evaluation happens “from within” the fiction. Participants reflect on their experiences, assuming characters’ viewpoints and commenting on technologies, emotions, and lived dynamics

This positive design fiction method enables collection of multiple insights describing multifaceted scenarios, useful for understanding people’s needs, expectations, and frustrations.

### *3.2 Anticipatory Ethnography: Observing Future Worlds*

If design fiction allows us to construct and immersively experience future worlds, the study of practices, emotions, and dynamics emerging in these worlds can be approached with innovative ethnographic methodologies. Dörrenbächer *et al.*’s proposal to evolve design fiction toward a participatory process moves in this direction. Parallel to this, Lindley *et al.* (2014) suggest Anticipatory Ethnography as an evolution of Design Ethnography that can unite design fiction’s imaginative freedom with ethnography’s methodological rigor.

Design Ethnography adapts classic ethnography to design needs, basing itself on situated observation of the present to generate insights for designing the near future. It anticipates future use scenarios but bases them on observable signals, behaviors, and emerging practices in the present. Design fiction, conversely, is free from temporal constraints but often lacks rigorous methods for systematic application in real contexts.

Anticipatory Ethnography emerges from the synergy be-

tween these practices: it extends ethnography's temporal scope, enabling observation and analysis of future worlds constructed through design fiction, while providing design fiction with methodological rigor. Lindley, Sharma & Potts propose three operational modes:

- studying the **design fiction** creation process by observing the creative team
- studying **public interaction with the fiction**, observing reactions and reflections
- studying the **fiction's content itself**, analyzing the narrative world as an ethnographic field; this mode involves direct ethnographer immersion in the artifact

Anticipatory Ethnography creates a “discursive space” where designers, researchers, and stakeholders can dialogue about preferable, plausible, and possible futures, generating actionable insights not only about the near future but also more distant temporal horizons. This perspective transcends predictive linearity and values the imaginative, participatory, and critical dimension of social research.

While traditional ethnography privileges the “here and now,” risking crystallization of the status quo, Anticipatory Ethnography overcomes this constraint. Through diegetic prototypes, design frees itself from temporal constraints and explores distant, plural futures. Ethnography provides rigorous methods for observing, analyzing, and evaluating interactions with fictional worlds, making design fiction more operational and relevant for real design. This perspective enables suspension of disbelief, generation of insight-driven future dialogues, and transcendence of traditional “situatedness,” enabling new temporal and conceptual freedom for ethnographic research.

#### **4. AI-Augmented Anticipatory Ethnography: The Orbyta Tech Method**

##### *4.1 The Futures Design Thinking Framework Developed by Orbyta Tech*

Developing technology is insufficient if it doesn't bring real value to people and organizations. With the spread of genera-

tive AI, global tech companies have pursued a techno-financial race focused on creating and dominating new markets, often neglecting to evaluate actual impact. In this context, local technology developers, though influenced by big tech logic, are called to question their role in guiding sustainable adoption, development, and diffusion of emerging technologies.

Orbyta Tech, a tech enabler supporting organizations in digital transformation now enhanced by AI, has questioned its role and the importance of innovating its approach to technology development. Through interdisciplinary dialogue led by the marketing department involving HR, design, business development, and technology delivery areas, a Futures Design Thinking method has been developed and tested to support organizations in developing a mindset suited to addressing challenges and imagining preferable scenarios for people and organizations. The heart of this method is envisioning processes, which enable people to imagine and concretely visualize possibilities, strengthening the intention to develop personal and team paths and innovation projects oriented toward preferable futures:

Envisioning is based on a very simple idea: it's much easier to achieve something if you can visualize yourself already achieving it (Tan, Goleman, Kabat-Zinn, 2012).

This pathway becomes preparatory for a more traditional Design Thinking process, one aimed at developing alternative services, products, and business models. But not before first unleashing new imaginative possibilities, disrupting the status quo of how organizations are managed and solutions are developed, and establishing a participatory process that brings key stakeholders together to envision and debate future scenarios. Only then can innovation, whether technology-enabled or not, truly take root and generate positive change.

For this reason, we developed an integrated Futures Design Thinking framework comprising three phases:

- **Futures Thinking:** the crucial, essential phase where co-design participants are challenged and guided to develop a

future-positive mindset through Futures Thinking and AI Fluency workshops. They use Orbyta Tech's "Postcard From Futures" toolkit to envision their own or their team's preferable future. This phase serves as a gym for exercising a new mindset and testing at the micro level (personal or small groups) the creative processes that will be developed at the macro level in the next phase. The final output is an artifact, the "Postcard From Futures", which participants take with them as a visual reminder of their intentions to embrace future with a positive attitude.

- **Futures Design Thinking:** the central phase where, applying Speculative Design and Anticipatory Ethnography, design fictions and artifacts of possible scenarios are imagined and enacted, then reactions are observed, emotional experiences are collectively reflected upon, and preferable futures are negotiated at the macro level. The final outputs are diegetic artifacts (videos, prototypes, fictional documents) that make future scenarios tangible and enable participants to experience and evaluate them from within, generating actionable insights for design.

- **Design Thinking:** following the imaginative opening of Futures Design Thinking, after unlocking mindset and choosing preferable futures, a more traditional design thinking phase can be initiated involving design of solutions or innovative projects. The outputs include prototypes, business models, and implementation plans that bring preferable futures to life.

Across all three phases of the framework, generative AI is used to amplify human capabilities and skills. It's simultaneously genuinely generative, giving participants demiurgic powers, and a tool for operational efficiency, compressing participation timelines and making it easier to keep working teams in a state of flow (Nakamura & Csikszentmihalyi, 2009). It enables mapping broader and more numerous perspectives than participants could generate alone, offering insights and directions that multiply possibilities and open up original, diverse scenarios. Beyond these uses, generative AI serves as an envisioning tool throughout all framework phases: starting from the development of preferable scenarios, AI-enhanced image, video, and digital interface

production tools allow for the creation of concrete artifacts that are visually and emotionally impactful. Thanks to generative AI, methodological tools from the Futures Thinking and Speculative Design traditions, like envisioning, design fiction, and artifacts, can be created rapidly, more affordably, and with greater effectiveness, expanding both their potential and contexts of use.

#### *4.2 Cultivate a Future-Positive Mindset: Unlock Your Mind and Envisioning*

The first phase of the Futures Design Thinking process aims to cultivate a future-positive mindset, an urgent optimism, including toward competencies needed to address change. Through envisioning exercises, storytelling, and AI-enhanced creation of metaphorical scenarios via “Postcard From Futures” artifacts, people deconstruct fears, reflect on their competencies, and imagine their futures. As McGonigal (2022) suggests, developing “urgent optimism” means training the capacity to imagine future scenarios and act immediately to make them possible. The future scenario addressed is personal or team-related, discussing both purpose and competencies for realizing it, particularly referencing key soft skills identified by the World Economic Forum (2023) and UNESCO (2024) that are exercisable through Futures Thinking: agility, continuous learning, creativity, empathy, critical thinking, optimism.

Operationally, this process occurs through Orbyta Tech’s “Postcard From Futures” toolkit. The toolkit involves a four-phase process with various analog and AI-enhanced creativity tools developed by Orbyta Tech:

- 1. Unlock Your Mind:** using Creative Cards, people confront images representing possible dystopias, interrogate their fears, and are called to act to overturn them. They discover that every dystopia can be unlocked and behind it lies a metaphorical representation of a preferable future.

- 2. Sketch The Storyboard:** using Creative Cards, The Storyboard canvas, and sketching techniques, people undertake an imaginary journey toward their preferable future, drawing on

the Storyboard (guided by Creative Cards) three key elements of the hero's journey they protagonist:

a. **The Place To Be:** the journey's destination, the preferable scenario representing their chosen purpose

b. **The Spirit Animal:** the guide animal they want to accompany them

c. **The Power Object:** the soft skills useful for best facing the journey

3. **(AI)Generate Your Envisioning:** starting from the storyboard and using generative AI image creation tools, people create "Postcard From Futures" that can be printed and become a personal or collective visual reminder of the preferable future.

4. **Send Your Intention:** on the postcard's back, people write their intention toward a future vision oriented to realizing personal or group purpose.

This toolkit can be applied in team building or AI Fluency workshops or as the initial part of innovation processes. In the latter case, the Futures Thinking phase continues with Futures Design Thinking and enables implementation of AI-Enhanced Anticipatory Ethnography.

#### *4.3 Experience the Fiction: AI-Enhanced Anticipatory Ethnography and Artifacts*

Within the Futures Design Thinking framework, people first develop a futures-positive mindset and exercise the envisioning mental process typical of Futures Thinking. From this mindset training, people are better prepared to engage in a participatory AI-Enhanced Anticipatory Ethnography process aimed at developing preferable scenarios related to an organization, market, product innovation, business model, or social impact project.

Preferable scenarios are not pure fantasy abstractions but are anchored to emerging possibilities because they're generated from signals in the present. For this reason, preferable scenarios are an alternative to both dystopias (dis-topos = "bad place") and "utopias" (ou-topos = "non-existent place") and can be defined as "eutopias" (eu-topos = "good place"). Eutopia construc-

tion starts from identifying signals of change, things happening today that could be clues to the future:

A signal of change is anything already happening today that could be a clue to the future. A signal shows how something could be different. It makes you say, “Aha! That’s new. That’s weird. I haven’t seen that before.” It sparks curiosity. It might be a new invention, product, business, behavior, the first successful demonstration of a new technology, the first major failure of an old technology, a new law, new kind of crime (verbatim from in person seminar, also in McGonigal, 2022 rephrased).

Eutopic scenarios and diegetic artifacts are built and created from collected signals: videos, installations, AI-generated prototypes, “future” documents. Following Dunne & Raby (2013) and Bleecker (2009), design fiction artifact creation is not only visualization but critical and social practice. Design fiction artifacts are narrative prototyping tools that suspend disbelief and test solution desirability from within the diegetic world, exploring social, cultural, and ethical implications of innovations. As proposed by Lindley, Sharma & Potts (2014), diegetic worlds thus created are ethnographic fields for observing future worlds, studying the creative process and public interaction with narrative content. Even more interesting is using AI to deliberately create “imperfect” or “broken” artifacts. Deliberately problematic AI-generated scenarios stimulate “anticipatory realism”: participants, confronting imprecise or dystopian representations, draw on their experience to correct, enrich, and make proposed scenarios more realistic (Pink *et al.*, 2025). Imperfection becomes a catalyst for critical reflection and co-creation.

In this process, ethnography plays a central role, offering rigorous tools for analyzing emerging dynamics that apply participant observation techniques and interviews with users, stakeholders, and involved communities. The ethnographic method applies both in the signal collection phase and in the design fiction field observation phase, aiming to identify frontier practices and habits, validate scenarios and prototypes through direct interaction, and collect feedback and insights to guide design.

Evaluation occurs from within the fiction and co-created scenarios. Participants reflect on their experiences, assuming characters' viewpoints and commenting on technologies, emotions, and lived dynamics. This approach enables collection of ambiguous and complex viewpoints, enriching future understanding.

The AI-Enhanced Anticipatory Ethnography method unfolds through workshops and observation, following this process:

- **Alternative Futures:** people question certainty elements of a market or topic, then overturn them by hypothesizing alternative futures
- **AI-augmented Signal Scanning:** for each alternative future, signals are sought in the present of that possible technological, regulatory, or social evolution. Signal research occurs through both desk methods augmented by AI (regulations, startup investments, patents) and field methods through ethnographic observation and in-depth interviews. Insights from ethnography are signals of change and thus empirical elements from which to build scenarios. Before doing so, participants are asked to discuss and select signals they imagine could lead to preferable futures
- **Eutopia Co-Design:** starting from selected signals, participants co-design eutopic scenarios with AI's generative support, while keeping humans in the loop for ethical considerations. They collaboratively constructs a future world where the identified signals have evolved positively
- **AI-Augmented Artifacts Creation:** using generative AI tools (text, image, video generation), diegetic artifacts are created that make the eutopic scenario tangible and immersive
- **Enactment and Anticipatory Ethnography:** participants experience the scenario, enacting roles and dynamics within the fiction. The ethnographer observes behaviors, reactions, and emerging interactions
- **From Within Evaluation:** evaluation occurs from within the fictional experience. Participants share reflections, emotions, and considerations about the lived scenario
- **Insights Synthesis:** insights collected are synthesized to

guide design of real solutions, products, services, or policies aligned with preferable futures.

This method enables rigorous exploration of future scenarios through an approach that combines imagination and empirical observation, leveraging AI as an enabling tool for creation and experimentation while maintaining humans at the center of the design and evaluation process.

## 5. Conclusions

The Futures Design Thinking method, and particularly the AI-Augmented Anticipatory Ethnography approach, represents an advanced methodological response to challenges posed by technological acceleration and the growing complexity of the NAVI (Non-linear, Accelerated, Volatile, Interconnected) world. By integrating futures thinking, speculative design tools, anticipatory ethnography, and generative artificial intelligence, this approach enables exploration of future scenarios in an articulated, inclusive, and participatory manner.

Epistemologically, the method breaks with the tradition of linear, deterministic prediction, embracing the plurality of futures as a resource for knowledge generation. As highlighted in Futures Studies and the Futures Cones model (Hancock & Bezold, 1994), the future is not singular but multiple: possible, plausible, probable, and preferable. AI-Augmented Anticipatory Ethnography operates precisely within this perspective, enabling eutopia co-creation through immersive and narrative practices.

Ethically, the centrality of “human in the loop” grounds decisions in participants’ experience and sensitivity. Scenario co-creation and “from within” fiction evaluation, as proposed by Dörrenbächer et al. (2020), enable a culture of shared responsibility, where social, cultural, and ethical implications of innovations can be discussed in protected, narrative environments.

Methodologically, Orbyta Tech’s framework is founded on synergy between AI and human intelligence, where AI acts as a

radar for identifying signals and patterns, while humans interpret and attribute meaning, as suggested in the Creative Co-Intelligence Manifesto (Meaningfool, 2025). This collaboration amplifies exploration and synthesis capacity while maintaining human control over value choices.

Finally, organizationally and socially, the method functions as a powerful transformative learning engine. The artifacts produced, from “Postcard From Futures” to design fictions, are not mere outputs but epistemic devices that activate new practices, languages, and visions. In line with McGonigal (2022), developing a future-oriented mindset and “urgent optimism” becomes fundamental for addressing uncertainty and generating positive impacts.

In summary, AI-Augmented Anticipatory Ethnography represents a methodological frontier capable of combining analytical rigor, design imagination, and ethical responsibility, offering concrete tools for exploring, co-creating, and testing preferable futures in participatory and creative ways. This approach provides concrete tools for addressing complex problems, anticipating changes, and generating positive impacts in society and organizations. Experimentation with participatory practices, construction of dialogue spaces, and valorization of diverse perspectives become key elements for promoting responsible and sustainable innovation.

## References

- Anthropic (2025), *AI Fluency: Framework & Foundations. Learn to collaborate with AI systems effectively, efficiently, ethically, and safely*, <https://anthropic.skilljar.com/ai-fluency-framework-foundations>
- Blecker, J. (2009), *Design Fiction: A Short Essay on Design, Science, Fact and Fiction*, Near Future Laboratory.
- De Biase, L. (2024), *Apologia del futuro*, Luiss University Press.
- Dörrenbächer, J., Laschke, M., Löffler, D., Ringfort, R., Großkopp, S., Hassenzahl, M. (2020), *Experiencing Utopia. A Positive Approach to*

- Design Fiction*, in CHI 2020 Extended Abstracts.
- Dunne, A., Raby, F. (2013), *Speculative Everything: Design, Fiction, and Social Dreaming*, MIT Press.
- EU, AI Act (2024), <https://ai-act-service-desk.ec.europa.eu/en/ai-act/recital-20>; [https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L\\_202401689](https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689)
- EY (2025), *What if disruption isn't the challenge, but the chance*, [https://www.ey.com/en\\_gl/megatrends/what-if-disruption-is-not-the-challenge-but-the-chance](https://www.ey.com/en_gl/megatrends/what-if-disruption-is-not-the-challenge-but-the-chance)
- Gorbis, M. (2019), 5 Principles for Thinking Like a Futurist, EDUCAUSE REVIEW <https://er.educause.edu/articles/2019/3/five-principles-for-thinking-like-a-futurist>
- Hancock, T. and Bezold, C. (1994), *Possible futures, preferable futures*, in Healthcare Forum Journal, Vol. 37/2, pp. 23–29.
- Lindley, J., Sharma, D., Potts, R. (2014), *Anticipatory Ethnography: Design Fiction as an Input to Design Ethnography*, in EPIC Proceedings.
- McGonigal, J. (2020), *How to Think Like a Futurist*, Stanford University Continuing Studies, <https://www.coursera.org/specializations/futures-thinking>
- McGonigal, J. (2021), <https://urgentoptimists.org/>
- McGonigal, J. (2022), *Imaginable: How to See the Future Coming and Feel Ready for Anything, even Things That Feel Impossible Today*, Transworld Digital.
- Meaningfool (2025), *Manifesto pratico della co-intelligenza creativa*, <https://www.meaningfool.it/>
- Nakamura, J., & Csikszentmihalyi, M. (2009), *The concept of flow*. In Snyder, C. R., & Lopez, S. J. (Ed.), in Oxford handbook of positive psychology. Oxford University Press, USA. 89-105.
- Pink, S., Korsmeyer, H., & Lyall, B. (2025), *Generative AI and Broken Futures*, in Qualitative Inquiry, 0(0). <https://doi.org/10.1177/10778004251358070>
- Sampson, O. (2021), *A Case for Design Anthropology for Creating Human-Centered AI*, in Cultural Anthropology, <https://www.culanth.org/fieldsights/a-case-for-design-anthropology-for-creating-human-centered-ai>
- Tan, C., Goleman, D., Kabat-Zinn, J., (2012), *Search Inside Yourself: The Unexpected Path to Achieving Success, Happiness (and World Peace)*, HarperOne
- UNESCO (2024), *Futures Literacy & Foresight*, <https://www.unesco.org/en/futures-literacy>

World Economic Forum (2023), *Future of Jobs Report*, [https://reports.weforum.org/docs/WEF\\_Future\\_of\\_Jobs\\_2025\\_Press\\_Release\\_IT.pdf](https://reports.weforum.org/docs/WEF_Future_of_Jobs_2025_Press_Release_IT.pdf)

**Alessandra Micalizzi**

PEGASO UNIVERSITY - SAE INSTITUTE (MILAN)

[alessandra.micalizzi@unipegaso.it](mailto:alessandra.micalizzi@unipegaso.it); [a.micalizzi@sae.edu](mailto:a.micalizzi@sae.edu)

Alessandra Micalizzi is an Associate Professor of Sociology of Cultural and Communication Processes at the Department of Psychology and Health Sciences of Pegaso University, where she directs the Research Center in Digital Humanities and its doctoral program. She also teaches at SAE Institute, where she lectures in Sociology of New Media. Her research interests include users' practices of technological appropriation, gender studies applied to the cultural and creative industries, and videogames as socio-cultural spaces of interaction and empowerment.

**Lara Balleri**

PEGASO UNIVERSITY - DIGITAL HUMANITIES CENTRE

[lara.balleri@unipegaso.it](mailto:lara.balleri@unipegaso.it)

Lara Balleri is a PhD student in the XXXIX cycle of the Digital Humanities program at the Unipegaso Telematic University. As a social-pedagogical educator and pedagogist, she is a member of several research groups and serves as a teaching tutor in General and Social Pedagogy at the IUL Telematic University of Studies. Her main research topic is narrative in its autobiographical form. She has authored several publications, available at: <https://orcid.org/0000-0002-1808-8640>.

**Leonard Busuttil**

UNIVERSITY OF MALTA

[leonard.busuttil@um.edu.mt](mailto:leonard.busuttil@um.edu.mt)

Prof. Leonard Busuttil (ORCID: <https://orcid.org/0000-0003-3779-891X>) is an academic and associate professor at the Faculty of Education, University of Malta, where he coordinates research and teaching in Computing education. His scholarly work spans generative AI in education, unplugged and constructionist approaches to computational thinking, and the pedagogical use of digital and game-based learning. His recent research explores how large language models reshape dissertation supervision practices, qualitative data analysis, and academic research cultures in higher education. Prof. Busuttil has extensive experience supervising postgraduate research, and supporting educators in navigating the methodological, ethical, and epistemic implications of AI-supported inquiry. He leads CPD initiatives for secondary teachers and university educators on the responsible integration of generative AI in teaching, learning, assessment, and research. He is also involved in institution-wide AI literacy programmes. His work foregrounds equitable, transparent, and critically reflective AI adoption in academia, advocating for supervisory practices that preserve interpretive ownership while leveraging AI's analytical affordances.

### **Rosienne Camilleri**

UNIVERSITY OF MALTA

[rosienne.camilleri@um.edu.mt](mailto:rosienne.camilleri@um.edu.mt)

Dr Rosienne Camilleri is a Senior Lecturer in the Department of Early Childhood and Primary Education at the University of Malta. Her research examines high ability and gifted education, inclusive pedagogy, childhood transitions, and learning identity in early childhood and primary education. More recently her scholarly work looked into digital practices and experiences in early childhood and explored how emerging technologies, including generative AI, shape research practices, academic writing, and supervisory pedagogy. Alongside her academic role, she is a warranted couple and family therapist. Her relational and systemic orientation informs both her teaching and her supervision philosophy, shaping a holistic approach to learner support, identity formation, and the development of researcher autonomy.

### **Giulia Coppo**

UNIVERSITY OF PADUA

[giulia.coppo@phd.unipd.it](mailto:giulia.coppo@phd.unipd.it)

Giulia Coppo is a PhD student in Sociology at the University of Padua. Her work addresses political communication and critical AI studies, with attention to professionalisation and AI-mediated practices in political marketing and social research. Previously, she has analysed local councillors' strategies, parties' digital media use, and politicians' online visibility. She is the author of "Le maschere degli eletti: la natura e le forme della comunicazione politica locale" (2022, Epokè edizioni).

### **Matteo Fogli**

ORBYTA TECH

[matteo.fogli@orbyta.it](mailto:matteo.fogli@orbyta.it)

Matteo Fogli is the Manager of the Digital Innovation Area at Orbyta Tech. Formerly founder, CEO and CTO of MODO, a web development and performance agency focused on fast, accessible and modern digital products, he brings a strong background in web performance, user experience and cross disciplinary product delivery. Over more than thirty years he has worked as product manager, project manager, business manager and creative developer, supporting organizations in designing high performing digital platforms and in adopting a performance driven culture. He is a pragmatic problem solver and a passionate technology consumer, committed to creating better online experiences through disruptive technologies and customer centric innovation.

### **Azaleah Mohd Anis**

RYSENSE LTD

[azaleah.anis@rysense.sg](mailto:azaleah.anis@rysense.sg)

Azaleah Mohd Anis is a qualitative researcher with 5 years' experience in qualitative research. Specialised in conducting in-depth interviews, she

has also conducted focus groups and run studies with various qualitative methodologies such as eye-tracking and online diaries. She also has experience with speaking to vulnerable and minority peoples such as at-risk youths and caregivers of special needs persons.

### **Nadia Olisa**

RYSENSE LTD

[nadia.olisa@rysense.sg](mailto:nadia.olisa@rysense.sg)

Nadia Olisa is currently Head of Qualitative Research & Business Partnerships at RySense Ltd. She set up the qualitative practice in RySense for running of focus group discussions as well as the set-up of new technologies such as eye tracking as complementary technological capabilities to qualitative research for the organisation. She currently leads the Qualitative team at RySense and trains qualitative researchers in moderation and running of focus groups, ethno-related and behavioural studies on social research. At RySense, she works with various government stakeholders to design and implement qualitative studies on public policy and public feedback in Singapore. Prior to joining RySense, Nadia hails from the private sector specialising in behavioural and public policy communications. She has more than 10 years in the industry having completed several qualitative projects with major commercial agencies and government institutions.

### **Elisabetta Risi**

IULM UNIVERSITY - MILANO

[elisabetta.risi@iulm.it](mailto:elisabetta.risi@iulm.it)

Elisabetta Risi is Assistant Professor at IULM University in Milan, where she teaches courses in social research methodology and in the sociology of cultural and communication processes. Previously, she taught at IUVE in Venice and Verona, and at NABA and IED in Milan. In 2007, she earned a PhD in "Information Society" from the University of Milan-Bicocca and subsequently held a postdoctoral fellowship and research grant at IULM University in Milan on projects dedicated to social research in the field of Internet Studies, with a particular focus on the socioeconomic processes arising from the spread of digital platforms and artificial intelligence.

### **Caterina Sapone**

PEGASO UNIVERSITY - SAE INSTITUTE (MILAN)

[caterina.sapone@unipegaso.it](mailto:caterina.sapone@unipegaso.it)

Caterina Sapone is a PhD student in Digital Humanities at the Department of Psychology and Health Science of Pegaso University. She is a member of LAHTI Lab in Milan. Her research is interdisciplinary, combining cognitive psychology, education, and user experience to explore human-technology interaction, with a focus on VR environment and GenAI tools. Her interests include motivation, literacy and socio-cultural factors that shape the interaction with technology.

### **Gabriella Taddeo**

UNIVERSITY OF TURIN

[gabriella.taddeo@unito.it](mailto:gabriella.taddeo@unito.it)

Gabriella Taddeo is Associate professor of Sociology of Cultural and Communication Processes at the University of Turin, where she teaches Digital Media Theory and Techniques, Sociology of Communication and Social Interaction Design. She is the author of over 50 publications in national and international journals, on the subject of digital media, social media studies and informal learning strategies through the digital spaces. She has also recently published: *Social. L'industria delle relazioni* (Einaudi, 2024) and *Persuasione Digitale. Come persone, interfacce, algoritmi ci influenzano online* (Guerini scientifica, 2023). On the subject of artificial intelligence, she has recently published: Taddeo G. (2024) "Artificial intelligence literacy: aspetti sociali e educativi di una nuova frontiera dell'educazione", in Ricucci R. and Rosa A. (eds.), *Didattica per Competenze e Orizzonti Educativi*, Pensa Multimedia, Lecce and "Prompting reflexivity: the use of artificial intelligence as a tool for identity and cultural exploration" (2025) in Novomisky and Le Voci Sayad (eds) *Alfabetización mediática y Informacional en la era de la Inteligencia Artificial*, Comunicar ediciones, Madrid.

### **Agnese Vellar**

ORBYTA TECH

[agnese.vellar@orbyta.it](mailto:agnese.vellar@orbyta.it)

Agnese Vellar is a Marketing Manager for the tech enabler Orbyta Tech. With a background in social research, community building and digital innovation, she helps people and organizations shape their vision and turn it into positioning strategies, marketing plans and community based innovation projects. She has carried out research and taught communication and innovation at Politecnico di Torino, Università di Torino and IED, and has worked with startups, unicorns and technology driven companies. She believes in the possibility of building preferable futures, in AI augmented creativity and in the value of emotions within business processes.





MEDIA, COMMUNICATION  
& SOCIO-CULTURAL PROCESSES

*Poche. La questione di genere nell'industria culturale italiana*  
a cura di Alessandra Micalizzi  
ISBN 9791255440130, prima edizione: giugno 2023, pagg. 296.

*Play seriously. The Transformative Power of Video Games*  
Preface by Fabio Viola  
edited by Alessandra Micalizzi  
ISBN 9791255440345, first edition: dicembre 2023, pagg. 204.

*Forme di produzione nelle industrie creative e culturali. Confini e significati*  
a cura di Rebecca Paraciani, Lorenzo Cattani  
ISBN 9791255440505, prima edizione: giugno 2024, pagg. 252.

*The Amplification of Sense*  
by Sascia Pellegrini  
ISBN 9791255440666, first edition: december 2024, pagg. 292.

*Digital Journalism. Transformations, Challenges, and Innovations*  
by Özgür Yılmaz  
ISBN 9791255440741, first edition: june 2025, pagg. 200.

*Artificial Intelligence and Social Research: Methods, Contexts, Imaginaries*  
edited by Alessandra Micalizzi  
ISBN 9791255440895, first edition: december 2025, pagg. 239.



Give a look to our books at





Finito di stampare da  
Services4Media Srl  
viale Caduti di Nassirya, 39  
70124 Bari