

Article

From Animated Paintings to Social Robots: Embodied Generative AI, Algorithmic Governance and the Data-Based Mediation of Cultural Heritage at the Alfieri Museum

Norberto Albano *  and Sandro Brignone * 

The Luciano Gallino Laboratory for Behaviour Simulation and Educational Robotics, Department of Philosophy and Education Sciences, University of Turin, Via Sant'Ottavio 20, 10124 Torino, Italy

* Correspondence: norberto.albano@unito.it (N.A.); sandro.brignone@unito.it (S.B.)

Abstract

Digital technologies and data-based approaches do not only increase the amount of information available to institutions and the public, they also transform the conditions through which social reality is classified, made visible, ordered, interpreted and governed. This article examines this transformation through AI4Alfieri, a project developed in Italy, at the Luciano Gallino Laboratory of the University of Turin in collaboration with the Fondazione Centro di Studi Alfieriani in Asti, and devoted to generative AI for the mediation of Vittorio Alfieri's life and work. Specifically, it presents a virtual conversational agent and one embodied in a social robot (Pepper or NAO). The project is analysed as a reflexive case study of a technology in the making, with autoethnographic elements. The empirical corpus comprises a development diary, code and design artefacts, prompts, logs, architectural diagrams and laboratory tests; the study does not evaluate effects on museum visitors. From analysis of the design process, the article advances the interpretive thesis that embodied generative AI makes cultural mediation a problem of distributed accountability: knowledge is governed by the corpus, speech by prompts and retrieval, action by deterministic policy, presence by the body, legitimacy by the institution, and auditability by infrastructures. The paper proposes an exploratory S0–S4 governance stack for embodied generative agents and formulates hypotheses for future research on how data-based cultural mediation may reallocate epistemic authority, institutional legitimacy and design responsibility.

Keywords: digital society; data-based approach; generative AI; social robots; cultural heritage; algorithmic governance; STS; digital sociology; accountability; autoethnography



Academic Editors: Felice Addeo and Gabriella Punziano

Received: 18 June 2026

Revised: 23 July 2026

Accepted: 25 July 2026

Published: 5 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction: When the Model Inhabits the Museum

Digital technologies and data-based approaches are redefining the fabric of contemporary societies, with effects that cross social practices, institutions, and the ways in which individuals construct knowledge, identities and relations [1–4]. The transformation does not concern only the quantitative increase in available data, nor the diffusion of more powerful computational instruments, but invests, in a pervasive manner, the forms through which social reality is classified, made visible, ordered, interpreted and governed. In digital societies, data-based processes do not merely represent the social world: they contribute to producing it, by establishing which entities count, which relations become relevant, which subjects are recognisable, and which decisions can be automated, assisted, or made plausible by computational infrastructures.

Within this general transformation, the recent development of generative artificial intelligence (generative AI), and in particular of large language models (LLMs), marks a further and decisive threshold. The novelty does not lie in embodiment as such: artificial agents endowed with body, voice and movement have populated interactional contexts at least since the wave of social robotics of the 1990s. What is new is the integration of foundation models into agentic systems: a generative capacity for language, reasoning and action that simultaneously redefines what an avatar on a screen and what a robot in physical space can do in the interaction with a human being. Along this threshold, a social space is taking shape in which the issue is no longer only the algorithmic mediation of contents, but the presence of situated artificial interlocutors, capable of speech and, when embodied, of gesture, posture and movement, perceived by users as actors within the interaction. What changes, sociologically, are the conditions under which knowledge is selected, ordered, presented and made credible within an institution.

The museum is one of the contexts in which these systems are introduced and tested. From a model mainly founded on conservation and static exhibition of collections, the contemporary museum tends to become an open space of experiences, relations and technological experimentation. Within an increasingly service-oriented vision of the museum, a museum-as-a-service logic, the technologies included in the RAISA acronym (robots, artificial intelligence and service automation) are adopted to reconfigure the visitor experience and include both virtual agents, such as avatars, digital humans and holograms, and social robots [5]. This article assumes the museum not as a mere application domain, but as a privileged observatory of the data-based transformation of the social mediation of knowledge. Cultural heritage, when translated into a computationally searchable corpus, is made operable by infrastructures of retrieval, classification, generation, validation and embodied presentation. The visitor's question becomes an input that is transcribed, vectorised, routed and compared with fragments of a corpus; the answer is no longer only text, but a situated utterance, pronounced by an agent that may have voice, posture, gesture and capacity for physical action in a shared space. Cultural mediation thus becomes a site in which data, algorithms, artificial bodies, platforms and institutions participate in the production of social reality.

In this perspective, the question from which we set out is not whether the robot is a good museum interface, it is more radical: what happens to the responsibility of cultural mediation when a language model enters a body, speaks within the institutional space of the museum, and acts on the basis of a certified corpus? The case examined in this article, AI4Alfieri, allows us to address this question because it places in continuity three forms of data-based cultural mediation devoted to the Italian writer Vittorio Alfieri: animated paintings at the Museo Palazzo Alfieri in Asti, a screen-based conversational avatar, and a Pepper/NAO social robot endowed with an LLM-powered agentic architecture.

The research question is therefore dual, because the same data-based transformation is observed from two sides of the artefact. On the first side, we ask how LLMs and social robotics are changing the way in which artistic and cultural heritage is valorised and experienced through embodied agents and LLM-powered interactive systems; in a more compact formulation, what happens when a language model stops living on a screen and begins to inhabit a body. This question projects simultaneously on several analytical planes. On the sociological plane, an anthropomorphic body activates relational expectations that differ from those produced by a textual chatbot: the human interlocutor attributes intentionality, authority and reciprocity in ways that redefine the dynamics of exchange and the social position of the non-human agent [6,7]. On the ethical and philosophical plane, an agent that speaks on behalf of a certified source and presents itself in physical space shifts the boundaries of epistemic responsibility, informed consent and identity

simulation, particularly when the agent speaks as, or appears to speak as, a historical figure [8–10]. On the plane of data governance, the construction of such agents requires explicit choices: which sources are introduced into the system, how answers are anchored to documentation, how grounded inference is distinguished from hallucination, and how bias, retrieval quality and traceability of robotic action are monitored [11,12]. On the plane of infrastructural governance, finally, the system depends partly on proprietary platforms, APIs and frontier models, which introduce exogenous constraints: costs, opacity, continuity of service, commercial policies, data processing and limits of auditability.

On the second side of the artefact, we ask how the same instruments, LLMs, coding agents and assisted development environments, are transforming the practices of the designers who build these agents. This movement is not an external addition to the first, but its complementary face. A consolidated tradition in the sociology of technology and in science and technology studies (STS)—from Zuboff’s study of how information technologies reshape work and skill [13] to the thesis that artefacts are never neutral but embody politics [14]—has taught us to consider technical devices as socio-cognitive media that redefine practices, competences and the division of labour. In Italy, this line was developed by the sociologist Luciano Gallino, after whom our Laboratory is named, in an early study of information technology and the quality of work [15]. The same attention must be applied reflexively to the tools with which the system is constructed. Already before LLMs, the programming of social robots such as Pepper and NAO was the object of sociotechnical analysis: Rudaz [16], for instance, shows how tools, exposed events, documentation and data hierarchies selectively constrain the forms of conversational competence that can be translated into robot behaviour. Our case observes the same platform in a later era, in which generative development tools participate in shaping both what the robot can become and the work of those who build it.

In sum, the overall question can be formulated as follows: how do embodied generative agents participate in the data-based reconfiguration of cultural mediation, and how do the same generative infrastructures transform the research and design practices through which such agents are constructed? We use the AI4Alfieri case to formulate an interpretive thesis, derived from the analysis of the design process rather than from measured visitor behaviour: when an LLM enters a robotic body and the space of a museum, cultural mediation becomes a problem of distributed accountability. This accountability is not concentrated in the model but runs along a chain that links the certified corpus and the prompts, the deterministic action policy, the robotic body, the institution that authorises the voice, and the external infrastructures on which sustainability and auditability ultimately depend. From this thesis, the contribution is threefold.

First, at the level of digital sociology and STS, the article interprets PepperAgent (the part of our AI4Alfieri project dedicated to social robots; see Section 3.1) not as a simple application of AI to heritage, but as a sociotechnical device through which epistemic authority, mediation roles and institutional responsibility are reallocated at the level of design. The robot does not merely deliver information: it prefigures a situated form of heritage interaction in which legitimate knowledge is selected, ordered, uttered and made credible through an artificial body. Second, at the methodological level, the article proposes a reflexive case study of a technology in the making. The empirical object is not the effect of the robot on real visitors, still not measured, but the sociotechnical process through which an embodied data-based agent is built, documented, corrected and governed. The development diary, design artefacts, code, prompts, validation policies and test logs are read as empirical materials of an auto-ethnography of design. Third, at the conceptual level, the paper proposes a five-level framework, S0 to S4, which interprets the accountability of an embodied cultural robot as distributed across different temporal,

technical, organisational and infrastructural scales, from the pre-cognitive presence of the robotic body (S0) to institutional and infrastructural governance (S4).

The remainder of the article is organised as follows. Section 2 reconstructs the state of the art on data-based systems for heritage mediation, bringing digital sociology, STS, datafication, virtual agents and museum social robots into dialogue, and ends by positioning our work within the field. Section 3 presents the AI4Alfieri case, discusses the methodological design as a reflexive case study of a technology in the making, describes the three forms of mediation explored by the project, and reconstructs the sociotechnical architecture of PepperAgent. Section 4 develops the discussion around power, institutional legitimacy and embedded algorithmic governance: it reads corpus governance as epistemic power, the robot's voice as borrowed institutional authority, the S0–S4 framework as a governance stack, and the use of coding agents as a transformation of the design process itself. Section 5 concludes with three theses and outlines future work.

2. From Datafication to Embodied Algorithmic Mediation

2.1. Datafication, Epistemic Authority and Algorithmic Relevance

The notion of datafication indicates the process through which practices, relations, experiences and cultural objects are converted into computable data, that is, into elements susceptible to storage, calculation, correlation, classification and intervention [4,17]. In digital societies, datafication does not coincide with simple digitalisation. To digitalise means to transform an object or a procedure into digital format; to datafy means to make it available to operations of measurement, extraction, indexing, profiling and governance. In this sense, datafication is a sociotechnical condition: it produces new possibilities of knowledge, but also new asymmetries of power, new criteria of visibility and new forms of infrastructural dependence [2,18].

Applied to cultural heritage, datafication has at least three effects. First, it converts documents, texts, images, works and biographies into computational bases that can be queried. Second, it transforms the relation between public and heritage into an interaction mediated by systems of search, recommendation, generation and visualisation. Third, it makes cultural institutions participants in a new economy of relevance: it is no longer sufficient to preserve and interpret heritage, it becomes necessary to decide how it must be structured so that an algorithm can retrieve it, order it and transform it into an answer. Algorithmic mediation does not operate in a social vacuum. Gillespie [19] has shown how algorithms participate in the production of public relevance, selecting what should appear meaningful in a given context; Striplhas [20] has proposed to read culture itself as increasingly crossed by computational processes, while Seaver [21] has invited us to consider algorithms not as isolated technical objects, but as situated cultural practices. These perspectives are decisive for reading PepperAgent: the system does not simply find information on Alfieri, it translates a certain idea of cultural pertinence, documentary authority and institutional sayability into procedures of retrieval and generation.

When heritage becomes a corpus for a RAG system, the sociological question is therefore not only whether answers are correct, but who defined the corpus, which texts were included, which remained outside, according to which criteria sources were segmented, which fragments emerge as relevant, and how the language model transforms them into discourse. Corpus governance is knowledge governance; the quality of retrieval is a technical question, but also an epistemic and institutional one; the mitigation of hallucination is, at the same time, an engineering problem and a form of cultural responsibility. This point connects the case to the broader field of digital sociology and digital methods, which has long stressed that digital data and computational infrastructures do not simply offer new sources for social research, but transform the conditions of observation, evidence and

interpretation [22–24]. In the case of embodied generative mediation, this transformation becomes visible in a particularly concrete manner because the corpus does not remain behind the interface: it is translated into a spoken answer, into a body, into a performance.

This argument calls for a more precise account of epistemic authority and trust, and we draw on three debates to sharpen the question that runs through AI4Alfieri. Fricker's account of epistemic injustice shows that credibility is not distributed neutrally, but socially allocated through power relations, testimonial expectations and institutional positions [25]: in a museum, an embodied agent may receive a surplus of credibility from the institution that authorises it, while visitors and human mediators have limited opportunity to contest or reinterpret what it says. The literature on trust in automation adds that trust is not a generic positive attitude but calibrated reliance on an imperfect system, where both over-reliance and under-reliance are failures of calibration [26]. Finally, the Responsible AI debate clarifies why accountability cannot be reduced to accuracy: Floridi and Cowls' principle of explicability links intelligibility to responsibility, requiring that users and institutions be able to understand how a system works and who answers for it [27].

2.2. Virtual Agents, Historical Figures and Synthetic Authenticity

The first family of comparable systems is composed of conversational agents, digital avatars and holograms of historical figures or cultural guides. This line has a history of almost two decades and a consolidated literature. For the pre-LLM framework, Machidon et al. [28] identify two recurrent vocations of the virtual human in cultural heritage: conversational agent and population of 3D environments. At the same time, they report the limited verbal communication capacity of these systems, often constrained by scripted dialogue or closed repertoires of answers. Contemporary LLMs have sharply reduced this bottleneck, while opening new problematic nodes: hallucination, authority, authenticity, corpus governance and identity simulation. Sylaiou and Fidas [29] propose a survey of virtual humans in museums and cultural heritage sites, showing a trajectory towards more sophisticated onsite installations and a growing investment in mixed reality, with behavioural realism as a critical dimension of engagement. Yan and Hashim [30] identify four prevalent roles of virtual avatars in digital heritage: cultural guide, narrative character, educational resource and social interface; among persistent issues, they highlight trust vulnerabilities, emotional inconsistency, uncanny valley reactions and new ethical questions connected to AI-generated representations.

A first line of conversational agents includes cases that have built the grammar of face-to-face interaction between user and digital agent. Max, installed at the Heinz Nixdorf MuseumsForum in Paderborn since 2004, is among the longest-lived embodied conversational agents in a real museum; Ada and Grace, photorealistic virtual humans at the Museum of Science in Boston, were developed for a public aged seven to fourteen and based on a pool of prerecorded answers selected through natural language understanding [31]. Within the same museum context, Tinker, the relational agent of the Museum of Science in Boston, has shown how affective and personalised interactional strategies can sustain engagement and learning over time, becoming a reference for the design of relational dimensions in cultural mediation [32]. These systems anticipate design choices on gesture, multimodality and small talk, yet they operate within a non-generative paradigm. A second line is that of authentic voice retrieval, in which the system does not generate speech but selects answers from an audiovisual corpus produced by the real person. The international ethical and methodological benchmark is probably Dimensions in Testimony of the USC Shoah Foundation, inaugurated in 2015 and now present in numerous museum institutions; in this case, prerecorded video testimonies of Holocaust survivors are queried through NLP, and the system recognises the visitor's questions and selects the corresponding filmed

answers. The ethical choice is explicit: the system does not generate words, but retrieves them, preserving the voice and presence of the witness and keeping the interaction outside the domain of generative simulation.

On the opposite side are deepfake or synthetic media systems. A paradigmatic example is *Dalí Lives* at the Salvador Dalí Museum, an installation that produces the effect of Salvador Dalí's revived presence through audiovisual generation techniques. Mihailova [9] reads it critically as a spectacular interface, in which authenticity of voice and image is put in tension with algorithmic fabrication. The agent speaks as the historical figure, but with a high degree of synthetic construction. The line more directly comparable to AI4Alfieri is the one of LLM plus RAG solutions on historical figures, emerging from 2023–2024. Wang and Adzharuddin [33] propose for this family the concept of synthetic authenticity: original cultural sources, but algorithmically produced content that emulates traditional forms while originating from computational processes. The Luigi Einaudi chatbot developed by Fondazione Einaudi with Reply, analysed by Natale et al. [10], is a particularly close case because of institutional affinity: a RAG system on a selected corpus of writings and biography, with a knowledge graph supporting retrieval, interrogated from the point of view of authenticity, historical accuracy and hallucination. DaCosta [34] similarly describes the construction of an AI-generated character of Joseph Lister, discussing accuracy, fidelity of voice, user discomfort and cultural biases inherited from the corpus. In the Turin context, *Tell Me More* integrates LLMs and web interfaces to support advanced exploration of cultural contents [35]; the case is not embodied, but it situates AI4Alfieri within an Italian cluster in which cultural institutions, universities and generative systems converge around the same question: how can a documentary heritage be transformed into data-based interaction without dissolving its regime of authority?

This panorama shows a fundamental convergence: LLMs make a more open dialogue possible, but move the problem from the fluidity of interaction to the legitimacy of generation. In a cultural context, the question is not only whether the agent is engaging, but whether it has title to speak, with which sources, with which voice regime and under which institutional responsibility.

2.3. *Museum Social Robots and the Intensification of Embodiment*

The second family of systems is composed of social robots, embodied agents designed to interact with human beings by following social norms, expectations and behaviours [36]. Museums, archaeological sites and heritage institutions have been a fertile field of experimentation for such systems because the task is not only to inform, but to orchestrate a cognitive and relational experience that involves attention, movement, proximity, surprise, trust and interpretation. The literature on robots in museums has progressively consolidated. Gasteiger et al. [37], in a quasi-systematic review focused on museum deployments, show that most experiments use the robot as a guide, while entertainment and educational functions are less frequent. Pepper and NAO by SoftBank Robotics have become de facto standard platforms in research on social robotics. Hellou et al. [38] identify five fundamental functions that a social robot must possess in order to interact as a museum guide: social navigation, perception, speech, gesturing and behaviour generation. The critical point is not the availability of each functional block taken individually, but here is their coordination in real time.

Historically, the trajectory of museum robotics can be reconstructed from pilot cases such as RHINO at the Deutsches Museum in Bonn [39] and MINERVA at the Smithsonian Institution [40], through longer-term deployments such as Lindsey at The Collection in Lincoln, where the robot worked as a tour guide for an extended period and made it possible to observe usage patterns over time rather than only on first encounter [41], up to

more recent integrations with Pepper. The Smithsonian pilot of 2018 highlighted how the robot tended to attract the attention of the visitor as an object of curiosity before functioning as a fully integrated informational resource [42]. The robot, in other words, does not simply enter a pre-existing interaction: it reorganises it around its material presence. In Italy, Pepper4Museum at the University of Bari integrated computer vision modules able to estimate age and gender of the visitor in order to recommend artworks [43], while the Virgil project at the Polytechnic University of Turin introduced forms of heritage telepresence at the Castle of Racconigi, allowing remote exploration of areas not accessible to the public under staff supervision [44]. Beyond museum settings, qualitative research on older human–robot interaction with Pepper has similarly shown how the materiality of the robot and its anthropomorphic corporeality participate in structuring the social relations established with it [45].

The work of Maniscalco et al. [7], also on Pepper in a museum context, constitutes the most mature reference for interactional anthropomorphism: a Perception–Understanding–Action architecture, RDF knowledge graph queried through SPARQL, management of main channel and backchannel, blinking, proximity, gestures and robot states. Precisely for this reason, it is a demanding comparator for PepperAgent: our system does not reach the same level of backchannel formalisation, but shifts the centre of gravity to the governance of LLM-driven action. The integration of LLMs into social robots marks a further passage. Recent experiments document the emancipation from conversational systems based on rigid structures and rules towards architectures powered by foundation models, with new capacities such as contextualisation of dialogue, more flexible turn-taking, dynamic narratives and personalisation of interaction. Nevertheless, the criticalities remain significant: latency, reliability, multilingualism, management of uncertainty, hallucination, physical safety, and perception of the robot as authoritative source or spectacular object.

On the LLM-robot side, Rojas et al. [46] evaluate Python code generation for Pepper-like robots on 720 tasks, showing both the potential of LLMs for general-purpose robotic tasks and the risk of leaving direct executable instructions to the model. Garelo et al. [47,48], with Alter-Ego at the Italian Institute of Technology, represent the most direct museum comparator: a GPT-4o mini system that uses function calling to trigger robot actions—moving to a location (`go_to()`) and ending the tour (`end_tour()`)—together with a dynamic prompt (position, progress of the visit, knowledge base and dialogue history), navigation based on Hector SLAM, AMCL and Move Base, and an evaluation in a real museum with 34 participants. In this comparison, PepperAgent does not claim superiority on autonomous navigation or user study; it claims an explicit separation between model proposal and deterministic authorisation of action.

This priority, the authorisation layer over any particular model, also governs how the underlying systems should be read. PepperAgent is deliberately model-agnostic, with interchangeable backends (see Section 3.3 and the discussion of modularity in Section 4.3), the specific model identifiers are a moving target rather than a design commitment. At the time of writing the system relied, on the OpenAI stack, on a higher-capability model for conversation and documentary retrieval (gpt-5.4) and a faster model for physical control and tool calling (gpt-4.1); transcription used an interchangeable backend, either a local WhisperLive 0.8.0 instance or the online gpt-4o-transcribe; scene description used gpt-5.4-mini, with a local RAM++ tagger for rapid object labelling; and retrieval combined FAISS 1.13.0 vector search over text-embedding-3-large embeddings. These choices have changed repeatedly during development, as providers released, renamed or deprecated models, and will continue to change; what the architecture fixes is not the model but the roles, constraints and validation through which any model is authorised to act.

The studies reviewed in this section make visible the specificity of social robots with respect to virtual agents. In avatars, the stake is above all the fabrication of image, voice and regime of presence within a screen. In social robots, physical embodiment introduces further dimensions: proxemics, gesture, materiality, risk of physical action, stronger attribution of agency and responsibility. A robot that moves can collide, choose a wrong trajectory, occupy space, interrupt, surprise, intimidate. The governance of interaction therefore does not concern only the linguistic answer, but situated action. Among the few studies that have attempted to measure the quality of the relation beyond the novelty effect, Iio et al. [49] propose an instructive case: the ASIMO experiment at the Miraikan museum showed, on returning visitors, a significant increase in perceived interpersonal closeness to the robot between first and second visit, with 94.74% of participants declaring that they wanted to use it again. This is one of the few works that documents, for a museum guide robot, something resembling a relation and not only a reaction, and it is therefore useful for calibrating the future evaluation ambitions of our case.

2.4. Gap and Positioning

Read together, the two families of systems, virtual agents and social robots, reveal, especially after the integration of LLMs, a significant convergence of problems. The common node is that LLMs, taken nakedly, are not suitable for cultural mediation: their propensity to hallucination and their lack of documentary grounding make them unreliable in a context in which the authority of the answer is constitutive of the value of the service. RAG, as anchoring to a corpus, is widely recognised as a minimal path towards greater reliability, but it opens shared problems: retrieval quality, coverage and bias of the corpus, selection criteria of sources, traceability of answers and regime of authenticity. At the same time, differences remain. Virtual agents are more scalable and less exposed to physical safety constraints, but they pose acutely the problem of audiovisual fabrication and impersonation. Social robots intensify relational expectations and introduce the exclusive problem of physical action in shared space. In one case, the question is: who speaks through the screen? In the other: who speaks and acts in the museum space? The literature raises two further difficulties that it rarely holds together. The very qualities that make an embodied agent engaging are those that make it persuasive: attention and involvement can amplify misplaced trust in answers that are fluent but imperfect. And the safeguards a designer can build—retrieval anchored to a corpus, prompts, deterministic policies—operate locally, while model providers, APIs and platform changes remain inside the governance chain and outside the designer's control. Openness of dialogue, credibility of voice and control over infrastructure thus pull against one another.

The AI4Alfieri case is situated at the intersection of these challenges. It integrates contemporary LLMs, RAG on an institutionally certified corpus, an embodied architecture for Pepper/NAO robots, deterministic policy guarding physical action, and a sociological/STS reflection on the construction of the system. The contribution is the proposal of an interpretive grid for understanding how an embodied generative agent participates in the data-based mediation of heritage and in the situated production of epistemic authority. This gap is also visible quantitatively. The broad bibliometric study by Şahin et al. [50] on tour guiding technologies reports that the field remains dominated, for more than 60% of publications, by computer science contributions, while tourism-specific research, and more generally sociological, ethical and relational research, accounts for about 8.6% of the total, with a scarcity of case studies, rare mixed-method integration and ethical considerations, on responsibility, biometrics and redefinition of professional roles, still largely to be developed. Our contribution is explicitly placed in this under-occupied cell.

The positioning of the case must therefore be read within a rapidly forming Italian ecosystem: Rosa et al. [51] on the R1 robot and 5G offloading, Natale et al. [10] on the Einaudi chatbot, Geninatti Cossatin et al. [35] on Tell Me More, Garello et al. [47,48] on Alter-Ego, and AI4Alfieri on the transition from animated painting to governed social robot. Within this ecosystem, the specific contribution of AI4Alfieri is not to introduce yet another cultural agent, but, through PepperAgent, to make observable the sociotechnical chain that transforms corpus, model, body, policy and institution into data-based cultural mediation, and it can be articulated through three architectural figures that we claim as proper to our case. The first is a dual routing between language models specialised by role, conversation and retrieval on one side, physical control and tool calling on the other, which translates into operational terms the dualism between automatic and reflective regimes of thought, System 1 and System 2 [52]. The second is a deterministic validation policy of action plans, which functions as a pragmatic bridge between the pre-LLM Ethical Layer of Vanderelst and Winfield [53] and the contemporary LLM-driven robotics literature on safety constraints [54,55]. The third is the S0–S4 framework as a grid of multi-temporal accountability for embodied cultural agents, which we propose in Section 4.3 and discuss as the principal conceptual contribution of the article. Table 1 summarises the most relevant comparators discussed in this section and the specific implication of each for the design of PepperAgent.

Table 1. Relevant comparators for positioning PepperAgent in the state of the art.

Work	System	Main Strength	Implication for PepperAgent
Rojas et al. [46]	Pepper-like robot, LLM generating Python code	Benchmark on 720 tasks, with 400 snippets passing runtime control	Shows the risks of direct code generation, PepperAgent uses typed ActionPlan and ex ante policy.
Maniscalco et al. [7]	Pepper, PUA architecture, RDF/SPARQL knowledge graph	Multimodal backchannel and more mature user evaluation	Reference for S0, PepperAgent is stronger on LLM/action governance and weaker on backchannel formalisation.
Rosa et al. [51]	R1, 5G, YARP/ROS, SLAM/AMCL	Offloading, autonomous navigation and test in two museums	Clarifies a design choice, PepperAgent does not compete on navigation but on governed embodied mediation.
Garello et al. [47,48]	Alter-Ego, GPT-4o mini, SLAM, function calling	Direct comparator, real museum, 34 participants, Q&A and autonomous navigation	Strong empirical benchmark, PepperAgent instead claims explicit LLM-policy separation.
Vanderelst and Winfield [53]	NAO, Ethical Layer, consequence simulation	Theoretical precedent for an ethical layer separated from the controller	ActionPolicy is a pragmatic bridge between pre-LLM Ethical Layer and LLM-driven robotics.

Table 1. Cont.

Work	System	Main Strength	Implication for PepperAgent
Rudaz [16]	Ethnography of Aldebaran/SoftBank, Pepper and NAO	Shows how programming tools shape the social behaviour of the robot	Legitimizes the auto-ethnographic reading of generative tools as co-actors of design.
Durante et al. [56]	Agent foundation model, vision-language-action	End-to-end multi-modal and generalist horizon	Useful contrast, PepperAgent privileges modularity, inspectability and distributed accountability.

3. The AI4Alfieri Case Study

3.1. Methodological Design: A Reflexive Case Study of Technology in the Making

The case mobilised to address the research question is the AI4Alfieri project, developed from 2023 at the Luciano Gallino Laboratory for Behaviour Simulation and Educational Robotics (Gallino Lab) of the Department of Philosophy and Education Sciences of the University of Turin, in collaboration with the Fondazione Centro di Studi Alfieriani in Asti. It will also include the experimental involvement of local educational institutions, such as the Liceo Classico Alfieri in Asti. The project concerns the application of generative AI to the valorisation, teaching and communication of the life and work of Vittorio Alfieri. It develops cumulatively along three lines, sometimes in succession and sometimes in parallel: animated paintings at the Museo Palazzo Alfieri, launched in 2023; the LLM and RAG-based screen avatar, developed between 2024 and 2025; and the Pepper/NAO humanoid robot with an agentic architecture and capacity for physical action, launched from 2025.

These three lines should not be read as a teleological progression towards embodiment, nor as the overcoming of one by another. The animated painting on the wall, the screen avatar and the robot in physical space are different strategies of mediation, each with its own constraints, affordances and regimes of interaction, which in a museum can coexist and complement one another. The third line, internally named PepperAgent, constitutes the principal empirical observatory of this article not because it is the climax of the trajectory, but because it is the case in which the sociological and ethical nodes of the embodied agent, anthropomorphism, attribution of agency, safety of physical action in shared space, legitimacy of institutional voice, dependence on generative infrastructures, present themselves in the densest form.

Methodologically, the article is a reflexive case study of a technology in the making, in the sense that STS and the sociology of technology have given to that expression [57–59]. The standpoint is, more precisely, that of analytic rather than evocative autoethnography [60]: reflexivity is turned on the authors' own design practice, in the service of theoretical understanding. The unit of analysis is not the robot as an isolated technical object, nor the empirical effect of the robot on visitors, but the sociotechnical process through which an embodied and data-based cultural agent is designed, documented, constrained and made governable. This methodological choice responds directly to the nature of the research object: if the argument is that accountability is distributed across corpus, model, body, policy, institution and infrastructure, the empirical material must make visible how such distribution is produced.

The empirical material is composed of five main sets, which Table 2 pairs with the analytical function each serves in the study. The first is the project's development diary:

a versioned Markdown file, kept in the same repository as the code and updated on an event-driven basis, with a new entry added whenever a non-trivial technical or design decision, a failure, a correction or a validation needs to be made traceable. Some entries are particularly relevant for this article: 14 April 2026 documents four failed attempts to obtain parallel tool calling and the transformation of the limit into a rule of method; 22 April 2026 records corrections to the RAG after a live session, including rules on discursive register; 26 April 2026 documents the separation between ActionPolicy and PlannerExecutor. The diary's structure is described below. The second set consists of the technical artefacts of the system: code, prompts, configurations, architectural schema and validation policies. These are not analysed to evaluate engineering performance in the narrow sense, but as traces of sociotechnical decisions: where a limit was placed, which risk was anticipated, what kind of user was implied, which form of agency was allowed or prohibited. The third set comprises laboratory tests, dry-run outputs and logs, together with simulated dialogues with a typical visitor of the Alfieri Museum; the system is currently in an experimental phase at the Gallino Lab, and deployment in the museum and with school classes remains future work, so this material documents the behaviour of the system before public use, not interaction with real visitors. The fourth set is the comparison with comparable projects, on the same or adjacent platforms, which places the architecture within the recent literature on social robotics, human–robot interaction and LLM-based agents. The fifth set, relevant to the methodological discussion (in Section 4.4), is the assisted development workflow itself: coding agents, repository instructions and the documentation produced during the work.

Table 2. Empirical materials and their analytical functions.

Empirical Material	Analytical Function
Development diary	Reconstructing decisions, discarded hypotheses, limits, rationales and verifications.
Code, prompts, policies and configuration files	Analysing how voice, safety, refusal, clarification and action constraints are inscribed in the artefact, and how responsibility is distributed between model, router, policy, executor and robot bridge.
Logs, dry-run outputs and laboratory tests	Observing the behaviour of the system before public deployment—routing, validation, interruption, latency—without real visitors.
Comparative literature and adjacent projects	Situating the case in relation to virtual humans, museum social robots, LLM-powered agents, STS and AI governance.
Assisted-development traces	Analysing how coding agents transform the practice of building, documenting and validating the system.

Entries in the development diary follow a recurrent, semi-structured pattern (not every field appears in every record): a dated title; the context or problem; the alternatives and attempts, including unsuccessful ones; the decision adopted and its rationale; the files affected; and, where relevant, verification results and open problems. Entries are retained even when the decisions they record are later superseded, so that the file preserves the sequence of alternatives and revisions. Its original function is project memory and operational traceability—for the authors and for the coding agents that read and update it during development—not the production of research data or support for a predetermined analytical thesis. Figure 1 reproduces one such entry.

```

## 2026-04-22 - RAG quality fix after first live session
### Context
First live session with RAG active on the Nao robot. Log analysis
revealed 5 problems: BM25 not loaded (wrong import), no page numbers
in the context, robot pronouncing raw file names
("Alfieri_base_avanzato.pdf"), internal language exposed
("fragments", "passages", "in this chat"), suggestions impossible
for a physical robot ("paste me the passage").

### What we did
A. Fix EnsembleRetriever import: in langchain 1.2.10 the module is in
langchain_classic.retrievers.ensemble, not in langchain.retrievers.
B. Added page to RAG metadata (PyPDFLoader already provides it).
C. (Deferred) Context carry-forward for conversational follow-ups.
D. Source-name mapping: RAG_SOURCE_DISPLAY_NAMES in settings.py
translates "Alfieri_base_avanzato.pdf" into "text on the life and
works of Alfieri".
E. Full rewrite of build_rag_instruction() with 4 rules:
- No file names, no technical jargon
- No "paste/upload/send a photo"
- Natural page citation
- Coherence with previous answers
F. RAG_CHUNK_MAX_CHARS from 2000 to~4000.

### Files touched
Modified: llm/engine.py, llm/prompts.py, config/settings.py

### Open problem
Conversational follow-up ("which source are you talking about?", "on
which page?") does retrieval with the new query, which semantically
does not match the original content. The~model sees different chunks
and contradicts itself. Requires context carry-forward or query
reformulation - to be planned.

```

Figure 1. An entry from the development diary (22 April 2026), reproduced in full and without editing and translated from the Italian original, with file names, paths and configuration constants preserved unchanged. The Italian original is provided in the Supplementary Materials. This is the entry from which the vignette in Section 3.3 is reconstructed.

The analysis did not take the form of a separate, systematic coding of the diary conducted after development. Reflection was, for the most part, contemporaneous with the work: the diary is the trace of decisions taken, questioned and revised while PepperAgent was being built, and the interpretive categories used in this article—corpus governance, voice regime, action safety, infrastructural dependence, reflexive design practice—are the concepts with which the design was already being reasoned. This reflection was itself partly distributed between the authors and their tools: the coding agents that maintain the repository hold the diary as context and resurface earlier entries during development, recalling that a given approach had already failed or already worked, while the authors re-read the diary and discuss its most salient points. Interpretation and decision, however, remained with the authors. The episodes discussed here are the ones that had already forced a durable change of design or method during construction. They were identified and foregrounded for the article in dialogue with the same assisted-development tools, and then re-read and checked against the corresponding code, prompts and policies. In this abductive movement between materials and sensitising concepts [61], the vignettes figure not as representative frequencies but as theoretically informative episodes; we make no claim to have coded the corpus exhaustively, and the evidential value of the diary rests on its having been written for the work, not for the argument. Such care is needed because the method is reflexive in a precise sense: we do not study a system built by others, but our own. We design PepperAgent and, in the same movement, make it the object of research. This coincidence between designer role and analyst role is the condition that makes autoethnography possible, because it gives direct access from inside to decisions and their discarded alternatives, but it is also the principal limit of the method: what is

gained in depth of access is paid for in critical distance. We therefore assume our analysis as a form of situated and partial knowledge, in the sense of Haraway [62]: not a view from nowhere, but a knowledge that openly declares its own position. The dual position of designers and analysts was treated as both an epistemic resource and a source of bias. Beyond the cross-checking already described, two further safeguards mitigate the risk of retrospective rationalisation. First, we deliberately included negative or failed episodes—such as the unsuccessful attempts at parallel tool calling discussed in Section 4.4—and not only successful design choices. Second, we separated in the writing the empirical level of what the system currently does in the laboratory from the theoretical level of what the design prefigures for future museum interaction. These safeguards do not remove the partiality of the account, but they make the position from which the analysis is produced more explicit and contestable.

Because this is a single case, the value of the analysis is not statistical but analytical. As Flyvbjerg [63] has shown, a well-chosen case does not necessarily serve to estimate frequencies, but to generate and test concepts. In this article, the case allows us to propose the S0–S4 framework as a grid of accountability and to read design choices as sociotechnical inscriptions. The case makes thinkable and transferable a grid of interpretation that subsequent research, on other systems and institutions, may confirm, correct or extend. It is equally important to clarify what this study does not do. It does not evaluate the efficacy of the robot with real visitors; it does not measure engagement, learning, trust or satisfaction; it does not quantitatively compare avatar, robot and animated paintings; it does not claim to demonstrate the impact of the system on museum experience. Its contribution is prior and complementary: it shows how such effects are prefigured, constrained and made governable through architectural, documentary, institutional and infrastructural choices during the construction of the system. Its natural complement is a study of reception with the museum's differentiated public, outlined as future work in Section 5, which would extend the approach of this article from the construction of the system to its appropriation by real audiences.

3.2. Three Forms of Data-Based Cultural Mediation

AI4Alfieri is an experimental project applying artificial intelligence to the specific case of Vittorio Alfieri, the Italian dramatist and poet born in Asti in 1749 and who died in Florence in 1803, in order to explore new perspectives for heritage valorisation, teaching and communication. Its general aim is twofold. On the applicative plane, AI4Alfieri aims to construct and test an integrated generative AI software system operating on a restricted and certified knowledge domain, through RAG, in support of the study, education and dissemination of Alfieri's life and work. The choice of a certified knowledge base, works by and on Alfieri selected by the Fondazione, is not only technical: it is also methodological, epistemic and ethical. It responds to problems of data quality and bias typical of data-driven systems, seeking to anchor answers to documented facts, reduce hallucinations and decontextualised or incorrect answers, and make the system's inferences traceable. On the methodological-scientific plane, the project proposes itself as a pilot system that is conceptually and technically replicable on other intellectual figures, other cultural heritages and other institutions; at the same time, it functions as a research device, an infrastructure for critical reflection on the social, ethical, cognitive and relational implications of the encounter between generative AI, social robotics and heritage.

Phase 1, launched in 2023, experimented with generative AI to animate some paintings displayed at the Museo Palazzo Alfieri in Asti. Short multilingual videos, activated by the visitor through QR code, return voice and movement to the portrayed figures, Vittorio Alfieri, his sister Giulia, his mother Monica del Maino and others, who tell their

stories in the first person. The aim was not to display a technological effect, but to break the passivity of the traditional relation between visitor and painting, testing a model of valorisation integrated between heritage care and digital mediation. This phase operates with pre-constructed narratives: AI is a production tool, not a runtime interaction system. The voice regime is mainly impersonating: the portrayed character speaks in the first person, according to a closed and controlled script.

Phase 2, implemented between 2024 and 2025, introduces open dialogue on a screen. An avatar of Vittorio Alfieri converses with the visitor through an LLM anchored to a knowledge base certified by the Fondazione through RAG. The initial corpus includes a specialist text on Alfieri's life and work. This choice is proportionate to the current phase: on a small, selected and well-controlled corpus, a relatively linear hybrid RAG works well, allows observation of the system's behaviour and reduces the risk of attributing to the model an ungoverned competence. It should not, however, be understood as the definitive architecture. The horizon of the project is to progressively extend the documentary base by using the organisation and structure of materials produced, preserved and selected over time by the Museum and the Fondazione, together with new material from the digitised Alfierian corpus validated with expert support. The purpose of the phase was to test the data-based core of the project: the capacity of the system to ingest information, organise it, establish relations between concepts, produce inferences and answer naturally, appropriately and coherently to the questions of the public, without dissolving institutional control into the language model alone.

Phase 3, launched in 2025, constitutes the passage from the avatar, a virtual body behind a screen, to the robotic body present in physical space. PepperAgent is an embodied conversational agent developed for the Pepper and NAO humanoid robots of SoftBank Robotics. It brings the LLMs and RAG system of Phase 2 into a robot endowed with posture, voice and gesture in space, capable of listening to the user, conversing in Italian, being interrupted during speech, executing physical actions guided by the conversational context, and maintaining coherent dialogue on the documentary corpus. The LLM+RAG core of Phase 2 remains the cognitive heart of the mediation; what changes is the body through which it presents itself and acts. In the passage from screen to robot, the same data-based device produces a different configuration of expectations, risks and responsibilities.

Phase 3 is the place where generative AI takes body and, by taking body, changes sociological status. The avatar can speak; the robot can speak and act. The avatar can simulate a presence; the robot physically occupies it. The avatar can be contained in a screen; the robot shares space with the visitor. In this passage the nodes of accountability become most evident. Each form of mediation produces a different regime of interaction and a different configuration of responsibility. The animated painting is closed, controlled, narrative, multilingual and autonomously activated; its strength is the quality of the script and the predictability of content, its limit is the absence of dialogue. The conversational avatar opens interaction, but remains placed in the screen; its strength is conversational flexibility, its risk is identity simulation. The embodied robot introduces physical presence, gesture, proxemics and action; its strength is co-presence, its risk is excessive attribution and the need to govern not only what is said, but what is done.

3.3. *Architecture as Sociotechnical Inscription*

PepperAgent realises the third phase of AI4Alfieri: an embodied conversational agent for Pepper and NAO social robots. The system combines two capacities that in the literature often appear separated: dialogue in natural language, including reference to certified documentary knowledge, and physical action in the space shared with the human interlocutor. Functionally, PepperAgent is a pipeline, that is, a chain of processing

that brings the user's voice through successive stages to verbal answer or robotic action: listening and transcription, interpretation of the turn of speech, decision and routing, possible documentary retrieval or action planning, safety validation, execution as voice response or movement (Figure 2). The system can operate in two regimes: connected to the real robot or in simulated mode, or dry-run, in which physical commands are logged but not executed. This latter mode is useful for development, testing and methodological observation. The system also adapts to two hardware profiles, Pepper and NAO, which differ in sensors, cameras and available postures.

A decisive material constraint is the obsolescence of the SoftBank platform: the NAOqi SDK 2.5.7.1, necessary to command Pepper and NAO, remains tied to Python 2.7 (here 32-bit CPython 2.7.18) while the contemporary AI ecosystem employed by the project (LLM orchestration, transcription, RAG, vision and observability) requires Python 3 (here CPython 3.9.23). This incompatibility produced a multi-process architecture with a TCP bridge between the modern controller and the legacy robotic subsystem. The software snapshot examined in this study corresponds to PepperAgent Controller v6.0 and Pepper Bridge v2.5.0. The constraint is not an implementation detail: it is an example of how material infrastructure inscribes possibilities and limits into the design of the agent. Interfacing through the ROS ecosystem, for which a NAOqi driver exists, would have been an alternative to the custom TCP bridge; it was not adopted because the Laboratory works in Python for its parallel activity in educational robotics with schools, and keeping a homogeneous Python stack lowers maintenance and onboarding costs. This choice is itself an instance of how the local material and organisational setting inscribes the architecture.

Sociologically, this pipeline is not a simple technical flow. It is a chain of data-based transformations: the visitor's voice becomes audio, audio becomes text, text becomes classified intention, intention becomes retrieval or action plan, the plan becomes a validatable object, the validated object becomes robot behaviour. Each passage incorporates choices on what counts as a question, as pertinence, as risk, as appropriate answer. For this reason, the architecture can be read as an infrastructure of accountability.

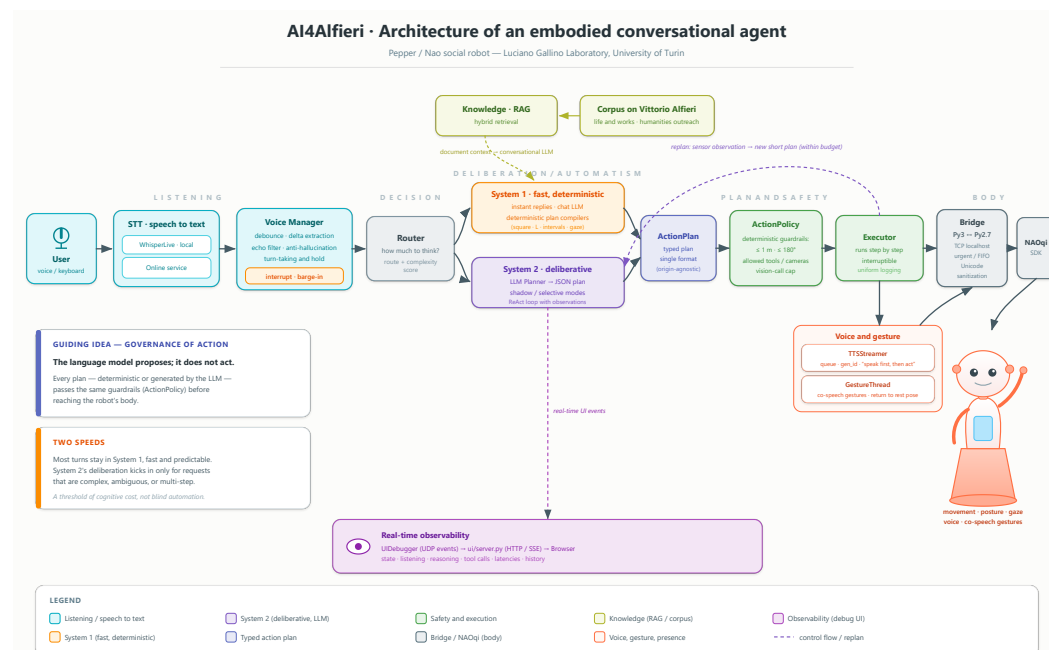


Figure 2. Architecture of PepperAgent: vocal pipeline, router, S1/S2 levels, ActionPlan, ActionPolicy, executor, robotic bridge and observability.

Imagine a user posing a question in front of Pepper in the context of the Alfieri Museum. The microphone captures the voice and the audio is converted into text by a speech recognition module with interchangeable backends behind a common interface: one local, based on WhisperLive, and one online, based on transcription services. The backend produces partial, provisional and updated transcriptions while the user speaks, and definitive transcriptions at the closure of the turn. The online backend also incorporates local detection of speech with adaptive threshold relative to ambient noise and volume normalisation before sending. Above the backend, a voice manager extracts clear speech turns from the transcription streams. It waits for stabilisation of the text, discards transcriptions that correspond to the robot's voice captured by the microphone, filters fragments that transcription models tend to generate on noise or silence, recognises interruption commands such as "stop", "fermati", "basta", and enables barge-in, that is, the possibility for the user to speak over the robot while it is functioning, stopping it and taking the floor.

These choices are technical, but not only technical. They govern the conversational turn: they establish when a noise becomes speech, when speech becomes command, when the robot must be silent, when the user can interrupt. In Goffmanian terms, the system participates in the interaction order [64], not simply in its transcription. Recent recommendations on spoken language interaction with robots and on general turn-taking models confirm that this infrastructure of listening, repair and interruption is a constitutive part of HRI, not a peripheral detail of the interface [65,66]. Once the text is obtained, a router classifies the turn and determines its treatment. It distinguishes conversational turns from those requiring physical action; it distinguishes simple cases from complex ones; it calculates a complexity score from explicit linguistic cues. Greetings and recurrent formulas receive ready answers; the presence of sequences, conditions, combinations of gaze and movement, or articulated requests, activates more complex paths. On this basis, the router chooses between a fast path and a deliberative path, according to a distinction inspired by the two regimes of thought, automatic and reflective [52].

A second level of routing concerns specialised language models by role. A faster model is used for physical control, tool calling and low-complexity answers; a model oriented towards linguistic quality is used for conversation, explanation and documentary retrieval. For informational questions on Alfieri, the system activates RAG: a hybrid retrieval that combines vector semantic search and lexical search on the corpus certified by the Fondazione. In the current architecture this solution is adequate because the corpus is small, curated and relatively homogeneous; the priority is not to maximise scale, but to preserve traceability, source control and pertinence to the domain. The system provides the model with pertinent passages including source and page indication, together with explicit instructions: adhere to sources, do not invent facts, cite naturally, do not expose file names or technical jargon to the visitor. Retrieval is selective: it is activated only when the request is effectively documentary. A short dialogue history maintains coherence between consecutive turns. RAG is, then, the point at which cultural mediation becomes governance of relevance. The system decides which fragments count as pertinent, in which order and with which weight; the Fondazione, by certifying the corpus, provides the institutional perimeter of the sayable; the language model transforms such fragments into an answer addressed to the visitor. The evolutionary perspective is to move from RAG as a simple top-k retrieval mechanism to a more agentic documentary infrastructure, in which retrieval becomes a tool of the system: structured memory, organised museum materials, digitised Alfierian corpus, expert validation, incremental update and explicit criteria of selection. Recent work on agent memory criticises flat RAG when materials become coherent, redundant and historically stratified, and proposes hierarchical and updateable structures based on decoupling and aggregation before retrieval [67].

An empirical vignette—a short narrative episode reconstructed from the development diary—clarifies the point. On 22 April 2026, after a live session with active RAG, the diary recorded problems that were not only technical but discursive: the robot pronounced raw file names, exposed internal jargon such as fragments or passages, and suggested actions impossible for a physical robot, such as asking the user to paste a passage or send a photograph. The correction produced four rules for the RAG prompt: no file names and no technical jargon, natural page citation, no photo requests incompatible with the robot's body, coherence with previous turns. Corpus governance thus translated into governance of register: it is not enough to retrieve the correct source, it is necessary to transform it into speech that is institutionally and bodily appropriate.

For physical turns, action is not left to the improvisation of the model. The fast path employs deterministic compilers that translate recurrent linguistic schemes, a path to follow, a movement at regular intervals, an orientation of gaze, into predefined action plans, without consulting the model and with constant outcome. When some important indication is missing, the compilers do not invent: they ask the user for clarification. They can also reuse the last executed plan when the user asks to do the same thing, possibly changing some parameter. When the request becomes more articulated, the deliberative path intervenes. An LLM-based planner compiles the user's intention into a structured plan, expressed in JSON, which lists sequences to execute, movements, rotations, sentences to pronounce, possible requests for clarification, with reference data, initial verbal confirmation and summary. This design choice is consistent with the broader pattern in the recent literature on language-grounded robotics, in which the model proposes high-level actions that are subsequently filtered through the affordances effectively available to the robot [68]: the deterministic compilers and planner therefore produce the same kind of object, namely a typed and uniform action plan. The planner can operate in observational mode, in which the plan is produced and logged but not executed, or in executive mode, in which the valid plan is executed. After an observation collected by the robot's sensors, it can also produce a brief update plan, again subject to the same controls.

The central sociotechnical point is the separation between generation of the plan and authorisation of action. The model can propose; it cannot decide by itself what the robotic body will do in shared space. Before execution, every plan, whatever its origin, is submitted to a deterministic policy. The policy verifies the admissibility and availability of tools on the robot profile in use, distances and rotation angles within predefined limits, head angles and velocities within conservative intervals, the validity of directions, maximum number of camera accesses per single plan, maximum length of sentences, and the absence of unexpected arguments. The cameras effectively allowed depend on the robot profile. A non-conforming plan is blocked before reaching the robot, which signals the impossibility and asks for a more precise formulation. Conforming plans are executed step by step and interruptibly by a single executor, independently from the component that produced them. The executor waits for the eventual verbal confirmation to be pronounced before starting movement: the robot, in other words, first says and then acts. At the end of a physical action, it returns limbs and posture to a neutral configuration.

A second vignette makes the design decision visible. On 26 April 2026, from real logs of controlled physical sequences, the need emerged to separate what produces the plan from what authorises action. The diary records the introduction of ActionPolicy, which validates ActionPlan before the tools, admitted names, availability on the robot profile, distances, angles, directions, length of sequence, and the extraction of PlannerExecutor, which executes every plan with the same interrupts, TTS barriers, logs and callbacks. The passage from textual answer to physical action plan is the point at which risk changes

nature; for this reason every plan, even when generated by an LLM, must cross a separate authorisation level.

The verbal answer is managed by a text-to-speech module that queues sentences and pronounces them as they are generated. A generation marker makes sentences queued but rendered obsolete by a user interruption be discarded rather than spoken. The module communicates to the voice manager what it is about to say, feeding the echo filter. While the robot speaks, a parallel process produces gestures coherent with the content of speech, chosen among categories such as explanation, reference to self or interlocutor, hesitation, assent, then returning the arms to a rest pose. In visual perception tasks, the system uses the camera and has complementary tools: rapid object labelling and narrative description of the scene in one or two sentences, supported by interchangeable vision backends. To orient the gaze, the robot first moves head and camera, resorting to body rotation only when the required angle exceeds the useful field of the head. At the stop word, the system interrupts speech and movement, empties pending queues, returns the arms to rest and pronounces a brief confirmation.

An observability module transmits in real time the internal events of the system, listening text, routing decision, tools invoked, validations, vision calls, latency, sentences pronounced, to a web interface accessible by browser. This dashboard makes it possible to follow the functioning of the system step by step and, in dry-run mode, to observe the entire behaviour even without the physical robot.

4. Discussion: Power, Legitimacy and Distributed Accountability

4.1. Corpus Governance and Epistemic Power

The tradition opened in Italy by Luciano Gallino showed how information technology reshapes work processes and forms of organised action, and called for a sociology attentive to the effects that machines produce on people [15]; already in the 1980s, Gallino had moreover proposed a model of the social actor engaging biology, culture and artificial intelligence [69], a line of inquiry recently extended, in the same Turin tradition, to the wider society of robots [70]. In a convergent line, the STS lenses already invoked in Section 3.1 provide the vocabulary to read an artefact such as PepperAgent as an actor within a network of humans, objects, institutions and procedures: Actor–Network Theory, which refuses an a priori split between human and non-human agents [71]; Suchman's account of agency as produced in interaction rather than possessed by the machine [59]; and Akrich's notion of the user inscribed in the artefact [58].

Applied to the case, these lenses focus on precise aspects. Datafication names what the project does to heritage before it does anything to the visitor: Alfieri's life and work are converted into a computationally queryable base, and the encounter itself, the visitor's question transcribed, vectorised and given to retrieval, enters a data-driven regime [18]. RAG and corpus are devices for selecting relevance: at every question, the system decides which fragments of heritage count as pertinent and in which order, exercising a politics of relevance that no neutral answer can fully hide [19]. This selection is not mere code: it is culture and situated practice, because it incorporates an idea of what is relevant, authoritative and sayable about Alfieri [20,21].

Following Suchman, the agency of the system is treated here not as an internal property of the robot but as an interactional accomplishment. In future museum use, it would depend on whether and how visitors attribute intention, competence, reciprocity and authority to the robot. The robotic body may amplify this dynamic: a textual chatbot can be consulted; a robot present in the space of the museum can be encountered, greeted, interrupted, observed, avoided or invested with social expectations. Taken together, these readings converge on an interpretive claim: PepperAgent may not only deliver information, but also

redefine the interaction order of cultural mediation by introducing a non-human participant into the museum space and redistributing the power to establish what counts as legitimate knowledge on heritage and who is authorised to utter it. In this sense we propose to read it as a site in which the social reality of the heritage encounter is produced, and not simply transmitted, in the datafied era.

The thesis articulated here is that each level of the system architecture, and each transversal theme that crosses those levels, represents a site of inscription [58]: a point in which a sociotechnical choice is translated into a material constraint of the system. Ethical nodes, in this reading, are not frameworks to add after design, but design choices observed from the side of their social and relational effects. The certified corpus inscribes an idea of authoritative knowledge; the prompt inscribes a voice regime; the safety policy inscribes a limit to action; the router inscribes a classification of turns; barge-in inscribes the user's right to interrupt; dry-run inscribes a principle of controlled experimentation; the audit log inscribes the possibility of reconstructing responsibility; modularity of backends inscribes a strategy for mitigating infrastructural dependence. Every technical element is also a decision on power, interaction and trust.

This is particularly evident in the governance of the documentary corpus and, inseparably, of the discursive register of the robot. The Fondazione Centro di Studi Alfieriani is not only a cultural partner: it is an epistemic actor. It decides, or contributes to deciding, which texts enter the corpus, which interpretations become available, which sources are legitimate and which remain outside. The choice of corpus is therefore an act of cultural consecration [72]: it does not merely guarantee accuracy, but participates in producing the legitimacy of a certain version of heritage. PepperAgent speaks on behalf of a corpus certified by a Fondazione, within a university project, in a museum. Its authority can therefore be interpreted as authority by association, a borrowed legitimacy from the institutions that guarantee it; the robot functions as a *porte-parole* [73], and its word may count because of the position it occupies more than because of what the model knows.

RAG does not eliminate the power of selection; it makes it operational. Every answer of the system is the result of a chain: selection of the corpus, the segmentation of documents, indexing, retrieval, ranking, prompt, generation, speech synthesis, gesture. Pertinence is operationalised through parameters, vector representations, criteria of similarity, semantic and lexical weights. What appears as a natural answer is the result of a procedure that makes some fragments more visible than others. For this reason, future enlargement of the corpus cannot consist of a simple quantitative increase in indexed documents. It must become an institutional process of organisation, validation and maintenance of cultural memory: materials produced and selected by the Museum, documentation of the Fondazione, digitised sources of the Alfierian corpus and expert competence must constitute the basis on which the system retrieves and reasons. In ethical terms, this architecture aims to relocate a greater share of responsibility and control from the model towards the institution: the model is not authorised to autonomously define the relevant heritage, but operates within a documentary memory whose form, extension and revision remain governed by responsible cultural subjects.

The power of enunciation is further transformed by embodiment. A written answer in a chat window does not have the same interactional status as an answer pronounced by an anthropomorphic robot in the museum space in real time. At the level of design, embodiment may translate epistemic authority into interactional authority: knowledge is a situated performance in which voice, posture, gesture, gaze and institutional context can participate in the credibility of the utterance. This claim remains an interpretive hypothesis until the system is observed with real visitors; what the present analysis can show is how the architecture prepares the conditions for such a transformation. Within this

interpretation, it is therefore useful to distinguish four forms of authority that converge in the case: epistemic authority, because the answer appears grounded in certified knowledge; institutional authority, because it appears authorised by the museum and the Fondazione; interactional authority, because the robotic body organises attention, turn-taking and proximity; infrastructural authority, because proprietary models and platforms condition what can be produced and how.

Assuming this perspective, the safety architecture described above, deterministic policy, non-coincidence between model and safety subsystem, observability and S4 governance (see Section 4.3), ceases to appear as a set of neutral technical precautions and can be read as a regime of algorithmic governance [74]: a device that regulates through code what the agent can say and do. Every regime of governance, however, responds to three questions, who governs, for whom, and to whom is it responsible, and in our case governance does not belong to a single subject. Engineers govern affordances, operational constraints and execution safety; designers govern the form of interaction and the regime of expectations; university researchers govern method, validation, documentation and scientific interpretation; the Fondazione and the museum institution govern cultural legitimacy, corpus and public conditions of mediation; providers model infrastructural constraints that must be managed but do not absorb the responsibility of local use. The visitor remains the addressee of governance, not yet co-author of the rules.

The limits of this self-governance must be openly declared. Audit logs, versioning of corpus and prompts, observability and stopping procedures are internal safeguards: they make the system legible to those who design it, but they do not automatically give the public a power of contestation [75]. The risk that the literature on human-machine interaction names automation complacency is particularly relevant here, because the very appearance of transparency, when it is taken as a guarantee of reliability, can paradoxically increase reliance rather than scrutiny: the more the system seems to expose itself to inspection, the more its outputs may be received as already legitimate, with a consequent erosion of the critical disposition that transparency was supposed to nourish. The adoption of an agentic documentary memory can reinforce these safeguards only if it remains subject to institutional curation: every update of the corpus, every new digitised source and every rule of aggregation must be traceable to a documentable decision, not to the model's optimisation alone. Moreover, the datafication of heritage carries with it a dimension of appropriation that critical studies define as data colonialism: the cultural record is transformed into a computational base that can be queried, extracted and reordered by technical infrastructures not always controlled by the cultural institution [2]. This is why corpus governance is also power governance, and why any claim of neutrality in data-based cultural mediation should be treated with caution.

4.2. Institutional Legitimacy and the Artificial Voice of Heritage

A crucial design choice concerns the regime of voice. The technically most spectacular possibility would be to make the robot speak as if it were Alfieri. The current choice is more cautious: Pepper speaks about Alfieri, not as Alfieri. This distinction is not a stylistic detail, but an ethical device. It reduces the risk of simulating a historical presence that does not exist, clarifies the mediating position of the robot and keeps open the connection with the institution that certifies knowledge. The problem of identity is not solved once and for all; it is made explicit as a specific design level. In the animated paintings of Phase 1, the first-person voice is scripted, closed and controlled; in the avatar of Phase 2, the dialogue is open and therefore the risk of synthetic authenticity and identity simulation increases; in the robot of Phase 3, the risk is intensified by physical co-presence. Saying that

Pepper speaks about Alfieri is therefore a way of aligning the voice of the agent with its epistemic position: it is a mediator, not a historical resurrection.

This choice also has implications for institutional legitimacy. As argued in Section 4.1, the robot's authority is largely borrowed from the museum, the Fondazione and the university that stand behind it; what the voice regime adds is that this borrowed legitimacy does not end with the selection of documents but extends to voice, register, refusal, clarification and declaration of limits. A robot that cites sources, asks for clarification, declares its limits (e.g., saying "I do not know", "the available sources do not allow me to answer", or "I can answer as a mediator, not as Alfieri"), or even allows itself to be interrupted and refers back to the institution, can be expected to support a different regime of legitimacy from a robot that simulates omniscience. In a context of cultural heritage, such humility can be understood as a governance mechanism, not merely as a conversational virtue.

The issue of legitimacy must also be connected to the redistribution of work between robot, human educator, cultural institution and designers. The introduction of PepperAgent does not necessarily imply replacement of the human mediator. Rather, it points towards a redefinition of the division of mediation work. The Smithsonian Pilot Pepper Robot Program summarises this posture well: Pepper functions realistically as an ice-breaker and transmitter of simple messages, not as a substitute for the human guide [42]. In this perspective, the robot could break the ice, answer frequent questions, offer multilingual access, activate curiosity and support introductory paths; the human mediator could retain, and perhaps see revalued, interpretive, contextual, pedagogical and relational competences. In Gallino's terms, the stake is less the replacement of mediation labour than its redistribution: which competences of the human mediator are revalued once the robot has broken the ice, and which risk being eroded if delegation becomes excessive.

As with the epistemic authority discussed above, these questions remain open and require empirical observation with real visitors and school groups. Yet design itself can already inscribe a response. The point is crucial for the ethics of generative AI. Many risks of LLMs derive not only from the possibility of error, but from the fluency with which error is expressed. A reliable generative system in a cultural context is not a system that never makes mistakes, but is a system that knows when not to answer, when to ask for clarification, when to declare uncertainty, when to stop. This is also why the limit of the system should not be hidden behind the marketing of capacity, but inscribed in the robot's own speech. If a sequence requires too much time to be controlled, if it does not pass safety checks, if the system does not have sufficient sources, the robot should say so.

This is also the point at which the literature on trust in automation becomes directly relevant. If trust is understood as calibrated reliance on an imperfect automated system [26], then the ethical problem is not to maximise trust in PepperAgent, but to calibrate it: a visitor should neither distrust the robot simply because it is artificial, nor rely on it as if institutional embodiment guaranteed infallibility. The features just described are, in this sense, candidate mechanisms of trust calibration. Their effectiveness, however, cannot be inferred from the architecture alone and must be tested in the planned user-facing phase of the project.

4.3. S0–S4 as a Governance Stack

The architecture of PepperAgent distinguishes bodily presence, fast routing, deliberation, voice regime and institutional/infrastructural governance. Read from an ethical and sociological perspective, these levels are functional layers and, at the same time, spatio-temporal scales of responsibility. The S0–S4 stack proposed here is an exploratory heuristic derived from our case study still under development: it is intended to make the distribution of accountability analytically visible and its robustness will have to be assessed through

future deployments and comparative cases. Extending the dualism between automatic and reflective thought, System 1 and System 2 [52], we propose to read the accountability of the robot as distributed across five levels, from S0 to S4. We adopt this dualism as a design and organisational heuristic, not as a claim that the robot reproduces human cognition: applying dual-process models to artificial systems is contested, and System 1 and System 2 here name regimes of control, not mental faculties of the machine. To the levels that govern behaviour in real time, we add S4, which operates on the long scale of institutional governance. Distributed accountability does not imply indistinct responsibility: each level has stakeholders with more immediate competence and authority to prevent, monitor or correct its risks. Responsibility is primary, not exclusive: engineers do not absorb the institutional responsibility of the museum, just as the Fondazione does not replace technical verification of policy, execution and physical safety. Table 3 articulates the five levels, with their temporal scale, central ethical question, implementation and primary responsibility. The last column identifies, for each level, the stakeholder who most directly possesses the competence and authority to prevent, monitor or correct the corresponding risk.

Table 3. Levels of accountability as temporalities of responsibility: extension of the System 1/System 2 framework to a five-level governance stack.

Level	Temporal Scale	Central Ethical Question	Implementation	Primary Responsibility
S0, pre-cognitive presence	milliseconds	relational legibility, non-spectacularity, non-deception	gaze, micro-movements, return to neutral pose, echo filter	interaction designers and HRI/robotics engineers; researchers for empirical verification of expectations
S1, fast deterministic routing	about one second	trust distributed by type of turn	router with complexity score, compilers for fixed plans	software/AI engineers and conversational designers defining router and turn patterns
S2, deliberation	seconds per turn	not delegating reliability to the model	LLM Planner, ActionPolicy, interruptible executor, replan	AI/robotics engineers and system integrators; scientific lead for risk thresholds and tests
S3, social norms and fidelity to the character	session	historical identity of Alfieri, ambiguity, repair instead of guess	RAG on certified corpus, role prompts, explicit clarifications, roadmap towards institutional agentic memory	Fondazione/curators and university researchers: corpus, voice, sources, interpretive boundaries and controlled growth of the documentary base
S4, institutional and infrastructural governance	days–months	who signs, who reviews, who can stop	audit log, versioning of corpus/prompt/model-id, stopping procedures, RAG life-cycle governance	museum institution/Fondazione, university laboratory and governance committee; Alfierian experts for validation of extended corpus

At the fastest extremes of the scale, the milliseconds of bodily presence and the second of conversational routing, the ethical stake is the calibration of expectation. S0 concerns everything that precedes explicit understanding: posture, gaze, micro-movements, proximity, gesture. An anthropomorphic body can make interaction more legible, but it can also produce excessive attribution, illusion of competence, undue trust and expectations superior to the real capacities of the system. The MuMMER project and studies on stakeholder expectations offer a warning in this direction [76,77]. Design must therefore pursue a presence that is legible but not spectacular. The relational value of multimodal interaction with Pepper in museums, documented by Maniscalco et al. [7], must be capitalised within a containment of register: the robot is socially legible, not socially realistic.

S1 concerns cases in which the system can respond rapidly and deterministically: greetings, confirmations, simple movement requests, recurrent formulas. Here the ethics consists in not feigning deliberation where it is not needed and in not involving a generative model when a stable rule is safer, more predictable and faster. Trust is distributed by type

of turn: not everything must pass through the LLM. Speed may therefore be treated as an ethical design variable, insofar as it can influence trust calibration: a response that is too rapid and formally confident can transmit an unjustified authority, amplifying epistemic risk [10].

At the centre of the scale, the seconds of the single deliberative turn, lies the architectural principle that we consider most relevant ethically: the non-coincidence between the language model and the safety subsystem. Taken up by a recurrent idea in the literature on LLM-controlled robotics [54,55], and anticipated in pre-LLM terms by the Ethical Layer of Vanderelst and Winfield [53], this principle is an explicit inscription in PepperAgent. The LLM-based planner proposes structured action plans, but a deterministic policy validates them, admitted tools, distances, angles, directions, sequence length, before they reach the executor, and a non-conforming plan is not executed. Reliability is not delegated to the model: it is pursued through separation, validation and interruptibility. Recent empirical work on adaptive autonomy in cultural robotics also shows that increasing the decisional autonomy of the agent can improve the perceived quality of the tour while at the same time reducing user satisfaction with the robot itself [78]: this asymmetry strengthens the argument for making refusals, validations and uncertainty visible to the user rather than concealing them behind apparent fluency.

S3 concerns the session scale and the norms of voice, social repair and fidelity to the character. Here the voice regime discussed in Section 4.2 governs the boundary between cultural mediation and identity simulation, in light of debates on synthetic authenticity, deepfake interfaces and historical characters [9,33,34]. S3 is also the level at which ambiguity should be repaired rather than guessed, a pattern that human–robot interaction research has shown to be more sociotechnically robust than silent disambiguation [79]: if the visitor’s question is unclear, the system should ask for clarification; if the corpus is insufficient, it should say so; if the answer requires interpretation beyond sources, the robot should make the limit visible. The corpus thus recurs at two distinct scales without contradiction: at S3 what is at stake is its curatorial use, its voice and its interpretive boundaries within the encounter, whereas at S4 it becomes a matter of life-cycle governance, namely versioning, validation and controlled growth over institutional time.

S4 is the long temporality of institutional and infrastructural governance: days, months, versions, updates, agreements, logs and dependencies. It asks who signs, who reviews, who can stop the system, and who is responsible when a provider changes a model or policy. PepperAgent builds many local safeguards: certified corpus, RAG, prompts, deterministic policy, observability, dry-run, logs, modularity. Yet a significant part of generative capacity still depends on frontier models and external APIs for transcription, language generation, multimodal vision or conversational orchestration. This dependency introduces costs, opacity, lock-in, privacy constraints, exposure to policy changes and limits of auditability. GDPR, the AI Act and ISO 31101:2023 are not treated here as objects of legal analysis, but as a normative horizon that makes visible the need to integrate risk management, human oversight and safety management systems for service robots [80–82]. This level also connects the case to the broader Responsible AI debate. In the terms proposed by Floridi and Cows [27], explicability combines intelligibility (the possibility of understanding how the system works) and accountability (the possibility of identifying who is responsible for its operation); S4 is precisely the level at which these two dimensions must be institutionalised, through versioning, audit logs, human oversight, stopping procedures, model documentation and public contestability. Modularity is therefore an important strategy, but not a solution. Making backends, models and vision components replaceable reduces dependence; it does not abolish it. The governance of an embodied cultural agent cannot stop at local action safety or corpus quality: it must include infrastructures, data in

transit, economic sustainability, contestability and institutional responsibility towards the public [83–85].

4.4. Designing Agents with Agents: Reflexivity and Methodological Transformation

AI4Alfieri, and in particular Phase 3 from which PepperAgent takes body, was developed in part with the support of generative coding agents, after years of manual programming in Python. By coding agent, we do not mean here a simple copilot suggesting local completions, but an agent capable of exploring the repository, reading and editing files, executing commands, linters and tests, incorporating feedback and producing patches or sequences of modifications oriented to an objective [86,87]. The workflow relied on two repository-level agentic clients, OpenAI Codex and Claude Code, used by the authors between February and June 2026. Punziano [88] invites us to rethink the boundaries between collection, analysis and interpretation of data in the presence of non-human agents endowed with generative capacities, underlining the co-production of meaning between researcher and system and the need for a reflexive, adaptive and ethically sensitive posture. Our case does not assume this framework as its main theoretical axis, but offers a situated version of it: it shows how human–machine co-production does not concern only the research process, but also the construction of an agent that will in turn mediate cultural knowledge towards the public. The theoretical framework remains that of the sociological considerations developed above, transposed to development tools. Our reflexive account suggests that coding agents, LLMs and assisted writing environments are not neutral instruments: they participate in the sociotechnical network, orient work, suggest solutions and make some architectural options more readily available than others. If Akrich shows how the user is inscribed in the artefact, here it is also necessary to observe how the designer is inscribed in the assisted development environment. In terms closer to recent software engineering, the case is placed within the trajectory of agentic software engineering: one does not design only software for human users, but also environments, specifications, constraints and procedures through which software agents can operate under human supervision [89].

The distinction proposed by this literature between the agent command environment and the agent execution environment is useful for reading PepperAgent. On one side there is the plane on which the researcher-designer formulates objectives, constraints, success criteria and stopping decisions; on the other side there is the operational plane on which the agent reads the repository, proposes modifications, runs checks and produces artefacts. These two environments are not external to the system: they become part of its design. The layered safety of PepperAgent thus finds a parallel in the process that generates it: assisted development also requires routing, validation, logs, review, dry-run and thresholds beyond which one does not proceed.

As described in Section 3.1, the assisted-development workflow also maintains the diary and generates the versioned traces used in the reflexive analysis. This integration of documentation and construction makes the workflow itself part of the empirical apparatus. This observation dialogues with studies on context files for coding agents: AGENTS.md, CLAUDE.md, repository prompts and operational instructions are becoming design artefacts, because they make context partially executable by the agent [90]. The critical lesson, however, is not that more context always produces better agents. Gloaguen et al. [91] show that generated or redundant contexts can increase costs and agent steps, and may even reduce task success. In our case, the diary has value because it is selective, situated, verifiable and connected to effective decisions, not because it generically accumulates memory.

A third vignette shows how a limit of the model becomes a norm of practice. On 14 April 2026, faced with the composite command “go forward thirty centimetres and then

turn ninety degrees”, the main model produced a single tool call instead of two. The diary records four configurations tried, explicit prompt, tool-choice parameters, reasoning settings and variants of the builder, all incapable of solving parallel tool calling without side effects on latency, preliminary text or completeness of arguments. The final decision was not to mask the limit, but to accept it consciously and transform it into an operational rule: after two failures on the same technical hypothesis, it is better to stop and recalibrate. The vignette is coherent with a broader tendency in the literature: agents can appear competent in the general understanding of an objective, but degrade when they must translate it into executive steps, intermediate constraints and verifiable actions; for this reason, works such as ToM-SWE insist on modelling user intent and memory, not only on patch generation [92]. Sociotechnically read, the limit of the artefact inscribes a norm in the designers.

Two observations emerge from this trace. The first concerns the symmetry between product and process. The mechanisms that PepperAgent inscribes in its operational levels, tool calling, ex ante validation, management of ambiguity, refusal of non-validated action, reappear in development practice: typed file operations, code review, internal documentation of decisions, acceptance of tool limits. Layered safety is not only a property of the built system, but a practice of its construction. The second concerns the transformation of design responsibility. When a piece of code, a refactoring or an architectural solution emerges in dialogue with a coding agent, we interpret responsibility as redistributed across the designer, the generative system as co-author and the traceability artefacts, while acceptance, integration and validation remain the designer’s responsibility. Recent empirical literature confirms this caution. Agents can reduce effort and expand the range of completable tasks [86], and in real workflows based on pull requests they can produce contributions frequently accepted, but often after human review [87,93]. At the same time, effects on quality are not automatically positive: causal studies observe persistent increases in static warnings and cognitive complexity after the adoption of autonomous agents [94]; agentic refactoring tends to produce local improvements more than deep architectural realignments [95]; subsequent maintenance of generated code continues to fall largely on human developers [96]. In this sense, development assisted by generative agents requires new practices of accountability: declaring AI use, documenting decisions, maintaining traces of changes, distinguishing generative suggestion and authorial responsibility.

We close this section with two notes of methodological honesty. The first is a limit: we do not have a precise metric of the fraction of code produced with generative assistance compared to code written manually. Ours is an observation of practice, not a measurement of output, and for this reason we do not advance a productivity thesis. The second is a reflexive implication: the same duality—AI as object and instrument—also informs the writing of this manuscript, which was in part conducted in dialogue with coding agents and generative assistants. If Punziano [88] proposes adaptive epistemology in order to account for the irruption of generative AI into social sciences, Lenzi et al. [97] reconsider the status of the research subject in digital social research when data become traces of persons, and Trezza et al. [98] thematise response quality of LLMs in sensitive contexts, our contribution adds a further level: when generative AI enters not only the analysis but also the construction of the object of research, the figure of the designer, and with it responsibility, must also be rethought. The question is not only whether coding agents write correct code, but also what second system of accountability must be constructed around them: living specifications, repository instructions, selective memory, tests, review, logs, sandbox and provenance.

5. Conclusions

The AI4Alfieri case, read through the sociological, ethical and methodological considerations developed above, allows us to return to the dual question that opened the article: how are LLMs and social robotics changing the way in which artistic and cultural heritage is valorised and experienced through embodied agents? And how are the same instruments changing the practices of the designers who build such agents? Our answer is articulated in three theses. Given the present research design, these conclusions concern the design process and the governance implications of an embodied generative agent under construction: they should not be read as measured effects on visitors, but as theoretically grounded implications and hypotheses generated from the AI4Alfieri case.

First, embodied LLMs may reconfigure cultural mediation as a situated production of epistemic authority. When a language model enters a social robot, cultural mediation knowledge is uttered by an agent endowed with body, voice, gesture and capacity for action. At the level of design, embodiment positions the answer as an interactional event. The robot is not a page, nor a search engine, nor a simple talking screen: it is designed to function as a non-human participant in the interaction order of the museum. Whether visitors so receive it, and how far this redistributes epistemic authority in the encounter, is a question for the planned reception study; but the design already distributes that authority across a network: the Fondazione certifies the corpus; the University designs the system; retrieval selects fragments; the model reformulates them; the robot pronounces them; the hosting institution confers context and legitimacy; and the generative platform provides part of the infrastructural capacity. The authority of the answer does not belong to a single actor, but to a sociotechnical network. What must be governed is this whole network, not the model alone.

Second, the accountability of an embodied cultural robot should be analysed as distributed across different temporal and organisational scales, as the exploratory S0–S4 framework proposed in this article sets out. Each level responds to a specific ethical question and requires a different form of implementation. This framework should not be understood as a definitive taxonomy, but as a provisional conceptual contribution. Its value consists in making visible that the reliability of an embodied generative agent cannot be sought in the LLM alone. Reliability is distributed: in the corpus, retrieval, prompt, body, policy, interface, logs, institutional review, possibility of stopping, declaration of limits and management of infrastructural dependence.

Third, local governance remains partial without a critique of infrastructures. As argued at level S4, local safeguards do not neutralise the dependence on frontier models and external APIs; modularity mitigates it without abolishing it. Moving towards local models or alternative providers can strengthen control and confidentiality, but introduces new technical responsibilities, hardware costs, maintenance and possible loss of quality. The governance of an embodied cultural agent cannot therefore stop at action safety or corpus quality: it must include reflection on the generative infrastructure on which the system depends.

For heritage institutions, the AI4Alfieri case suggests that the question often asked, avatar or robot, is probably the wrong question. The different modalities of embodiment, preconstructed narrative animation, screen-based conversational avatar, physical robot in the room, share the same generative AI and knowledge base, but respond to different needs and address different contexts: online reach of the avatar, physical co-presence of the robot, narrative integration of the animated painting. For the Alfieri Museum and similar institutions, the productive question concerns which combination to adopt, for which context, with which role for the human mediator and under which governance regime.

Future work follows three main directions. The first is empirical evaluation with real visitors of the Alfieri Museum and with school groups of the Liceo Classico Alfieri in Asti, attentive to dimensions of variation suggested by recent HRI literature, such as age, familiarity with robots and conversational AI, framing of the encounter and previous expectations. Conceived as the complement of the present design-stage study, this phase will combine, for example, observation of situated interaction, questionnaires, interviews, and focus groups with visitors, teachers and human mediators, and address the museum's differentiated public (from school groups to international visitors, older visitors and scholars), each bringing distinct communicative expectations and registers. It will examine how these audiences understand the robot's role, how they calibrate trust in its answers [26], whether they grant it institutional authority, how they read its declared limits, and how it reshapes the work of the human mediator. Quantitative measures of engagement and satisfaction may complement this corpus, but will not replace it. The second is the controlled expansion of the documentary infrastructure: a more explicit formalisation of S3, in dialogue with the curatorial criteria of the Fondazione and with the possible use of structured knowledge graphs, and the evolution of RAG from an adequate solution for a small controlled corpus to an internal tool of institutional agentic memory, built on materials produced and selected by the Museum, documentation of the Fondazione and digitised Alfierian corpus validated by experts. The third is the strengthening of external accountability: clearer procedures of audit, contestability by the public, governance of platform dependence, and possibly participatory forms of corpus governance.

The theoretical anchoring to the Luciano Gallino Laboratory is, in conclusion, more than an institutional note. The Laboratory looks at digital technologies with a sociological gaze, interrogating the effects that machines produce on people more than designing robotics in the strict engineering sense. In this perspective, LLMs and social robots have been treated here not as technical objects for their own sake, but as socio-cognitive media that may reconfigure cultural practices, modes of learning and relational configurations. And layered safety, in the light of the same gaze, is not only a property of the system built: it is a practice of its construction. With respect to the second part of the research question, our case suggests that generative tools transform design practice not only by accelerating code writing, but by requiring a redesign of the development process. The designer becomes also the curator of context, selector of constraints, reviewer of patches, guarantor of maintenance and responsible for the decision trace. Here too, as in the robot that speaks to the visitor, accountability is located in the sociotechnical network that makes the model operable, controllable and contestable.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/soc16080250/s1>, File S1: Development diary: the entries discussed in the article (English translation and Italian originals); Video S1: Short demonstration of the AI4Alfieri project.

Author Contributions: Conceptualisation, N.A. and S.B.; methodology, N.A. and S.B.; software and investigation, N.A.; validation, N.A. and S.B.; writing—original draft preparation, N.A. and S.B.; writing—review and editing, N.A. and S.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Fondazione Centro di Studi Alfieriani in Asti, which funded the generative-AI subscriptions and the API usage employed in the project; no specific grant number is associated with this support. The article processing charge was funded in equal parts by the Fondazione Centro di Studi Alfieriani in Asti and by the Department of Philosophy and Education Sciences of the University of Turin.

Institutional Review Board Statement: Not applicable. The study reported in this article does not involve experiments with human participants or animals; the planned deployment with museum visitors and school groups is indicated as future work and will be subject to the relevant institutional review procedures.

Informed Consent Statement: Not applicable.

Data Availability Statement: The three development-diary entries discussed in the article are provided in Supplementary File S1 in English translation together with the Italian originals. The remaining materials analysed in the reflexive case study, namely the full development diary, code artefacts, prompts, policy files, architectural diagrams, logs and laboratory dry-run outputs, are produced within the AI4Alfieri project at the Luciano Gallino Laboratory of the University of Turin. Access conditions and possible public release of these remaining materials are to be determined by the authors and by the institutions involved, subject to institutional, privacy and intellectual property constraints.

Acknowledgments: The authors wish to thank the Fondazione Centro di Studi Alfieriani in Asti, in particular Giulia Carluccio and Carla Forno, for the cultural partnership and for providing the certified documentary corpus that grounds the system, and the Liceo Classico “Alfieri” in Asti for their support and willingness to collaborate in the planned experimental phase with school groups. They are grateful to Renato Grimaldi, of the Luciano Gallino Laboratory and of the Fondazione Centro di Studi Alfieriani, for coordinating the work and for his scientific supervision. They also wish to thank the Collège des Bernardins and Graziano Lingua, of the Department of Philosophy and Education Sciences of the University of Turin, for their collaboration on this project. The implementation of the project, including the use of the Pepper and NAO social robots, was carried out at the Luciano Gallino Laboratory for Behaviour Simulation and Educational Robotics of the Department of Philosophy and Education Sciences of the University of Turin, directed by Paola Borgna. Parts of the software design of the PepperAgent system and parts of the preliminary drafting of this manuscript were conducted in dialogue with generative AI tools and coding agents, employed both for code production and review and for translation, linguistic revision, restructuring and $\text{\LaTeX}$ formatting support. The authors maintained full scientific, theoretical, methodological and editorial responsibility for the final content, including the selection of sources, the interpretation of results and the formulation of arguments.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
AI4Alfieri	Artificial intelligence for Alfieri
Gallino Lab	Luciano Gallino Laboratory for Behaviour Simulation and Educational Robotics
HRI	Human–robot interaction
LLM	Large language model
NAOqi	SoftBank Robotics software framework for NAO and Pepper
RAG	Retrieval-augmented generation
RAISA	Robots, artificial intelligence and service automation
STS	Science and technology studies
TCP	Transmission Control Protocol
TTS	Text-to-speech

References

1. Beer, D. *Metric Power*; Palgrave Macmillan: London, UK, 2016.
2. Couldry, N.; Mejias, U.A. *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*; Stanford University Press: Redwood City, CA, USA, 2019.
3. Couldry, N.; Hepp, A. *The Mediated Construction of Reality*; Polity Press: Cambridge, UK, 2017.

4. van Dijck, J. Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveill. Soc.* **2014**, *12*, 197–208.
5. Recuero-Virto, N.; Blasco López, M.F. Robots, artificial intelligence, and service automation to the core: Remastering experiences at museums. In *Robots, Artificial Intelligence, and Service Automation in Travel, Tourism and Hospitality*; Ivanov, S., Webster, C., Eds.; Emerald: Bingley, UK, 2019; pp. 239–253.
6. Reeves, B.; Nass, C. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*; Cambridge University Press: Cambridge, UK, 1996.
7. Maniscalco, U.; Minutolo, A.; Stornio, P.; Esposito, M. Towards a more anthropomorphic interaction with robots in museum settings: An experimental study. *Robot. Auton. Syst.* **2024**, *171*, 104561. <https://doi.org/10.1016/j.robot.2023.104561>.
8. Foka, A.; Griffin, G. AI, cultural heritage, and bias: Some key queries that arise from the use of GenAI. *Heritage* **2024**, *7*, 6125–6136. <https://doi.org/10.3390/heritage7110287>.
9. Mihailova, M. To dally with Dalí: Deepfake (inter)faces in the art museum. *Convergence* **2021**, *27*, 882–898.
10. Natale, S.; Surace, B.; Mensa, E.; Befera, L. ChatGPT for cultural heritage and the customization of generative AI: A talkthrough analysis of the Luigi Einaudi chatbot. *New Media Soc.* **2025**, advance online publication. <https://doi.org/10.1177/14614448251384258>.
11. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; Hajishirzi, H. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv* **2024**, arXiv:2310.11511.
12. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv* **2023**, arXiv:2312.10997.
13. Zuboff, S. *In the Age of the Smart Machine: The Future of Work and Power*; Basic Books: New York, NY, USA, 1988.
14. Winner, L. Do artifacts have politics? *Daedalus* **1980**, *109*, 121–136.
15. Gallino, L. *Informatica e Qualità del Lavoro*; Einaudi: Turin, Italy, 1983.
16. Rudaz, D. Social robots as designed artifacts: The impact of programming tools on “human-robot interaction”. *AI & Soc.* **2026**, *41*, 3095–3119. <https://doi.org/10.1007/s00146-025-02743-7>.
17. Mayer-Schönberger, V.; Cukier, K. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*; Houghton Mifflin Harcourt: Boston, MA, USA, 2013.
18. Kitchin, R. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*; Sage: London, UK, 2014.
19. Gillespie, T. The relevance of algorithms. In *Media Technologies: Essays on Communication, Materiality, and Society*; Gillespie, T., Boczkowski, P., Foot, K., Eds.; MIT Press: Cambridge, MA, USA, 2014; pp. 167–194.
20. Striplas, T. Algorithmic culture. *Eur. J. Cult. Stud.* **2015**, *18*, 395–412.
21. Seaver, N. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data Soc.* **2017**, *4*, 1–12. <https://doi.org/10.1177/2053951717738104>.
22. Lupton, D. *Digital Sociology*; Routledge: London, UK, 2014.
23. Marres, N. *Digital Sociology: The Reinvention of Social Research*; Polity: Cambridge, UK, 2017.
24. Rogers, R. *Digital Methods*; MIT Press: Cambridge, MA, USA, 2013.
25. Fricker, M. *Epistemic Injustice: Power and the Ethics of Knowing*; Oxford University Press: Oxford, UK, 2007. <https://doi.org/10.1093/acprof:oso/9780198237907.001.0001>.
26. Lee, J.D.; See, K.A. Trust in automation: Designing for appropriate reliance. *Hum. Factors* **2004**, *46*, 50–80. https://journals.sagepub.com/doi/abs/10.1518/hfes.46.1.50_30392.
27. Floridi, L.; Cowls, J. A unified framework of five principles for AI in society. *Harv. Data Sci. Rev.* **2019**, *1*, 1–14. <https://doi.org/10.1162/99608f92.8cd550d1>.
28. Machidon, O.M.; Duguleană, M.; Carrozzino, M. Virtual humans in cultural heritage ICT applications: A review. *J. Cult. Herit.* **2018**, *33*, 249–260.
29. Sylaiou, S.; Fidas, C. Virtual humans in museums and cultural heritage sites. *Appl. Sci.* **2022**, *12*, 9913. <https://doi.org/10.3390/app12199913>.
30. Yan, P.; Hashim, M.E.A.H. Virtual avatars in digital heritage: A systematic review of design evolution and user experience. *Int. J. Creat. Multimed.* **2026**, *7*, 213–227. <https://doi.org/10.33093/ijcm.2026.7.1.11>.
31. Swartout, W.; Traum, D.; Artstein, R.; Noren, D.; Debevec, P.; Bronnenkant, K.; Williams, J.; Leuski, A.; Narayanan, S.; Piepol, D.; et al. Ada and Grace: Toward realistic and engaging virtual museum guides. In *Intelligent Virtual Agents, IVA 2010*; LNCS 6356; Springer: Berlin/Heidelberg, Germany, 2010; pp. 286–300.
32. Bickmore, T.W.; Vardoulakis, L.M.P.; Schulman, D. Tinker: A relational agent museum guide. *Auton. Agents-Multi-Agent Syst.* **2013**, *27*, 254–276.
33. Wang, C.; Adzharuddin, N.A. Synthetic authenticity and audience trust in AI-generated intangible cultural heritage: A qualitative multimodal study of Chinese digital heritage platforms. *e-J. Media Soc.* **2025**, *8*, 1–10. <https://journal.uitm.edu.my/ojs/index.php/e-JOMS/article/view/6894>.

34. DaCosta, B. Speaking with the past: Constructing AI-generated historical characters for cultural heritage and learning. *Heritage* **2025**, *8*, 387.
35. Geninatti Cossatin, A.; Mauro, N.; Ferrero, F.; Ardissono, L. Tell Me More: Integrating LLMs in a cultural heritage website for advanced information exploration support. *Inf. Technol. Tour.* **2025**, *27*, 385–416. <https://doi.org/10.1007/s40558-025-00312-8>.
36. Breazeal, C.; Dautenhahn, K.; Kanda, T. Social robotics. In *Springer Handbook of Robotics*, 2nd ed.; Siciliano, B., Khatib, O., Eds.; Springer: Cham, Switzerland, 2016; pp. 1935–1971.
37. Gasteiger, N.; Hellou, M.; Ahn, H.S. Deploying social robots in museum settings: A quasi-systematic review exploring purpose and acceptability. *Int. J. Adv. Robot. Syst.* **2021**, *18*, <https://doi.org/10.1177/17298814211066740>.
38. Hellou, M.; Gasteiger, N.; Lim, J.Y.; Jang, M.; Ahn, H.S. Technical methods for social robots in museum settings: An overview of the literature. *Int. J. Soc. Robot.* **2022**, *14*, 1767–1786.
39. Burgard, W.; Cremers, A.B.; Fox, D.; Hähnel, D.; Lakemeyer, G.; Schulz, D.; Steiner, W.; Thrun, S. Experiences with an interactive museum tour-guide robot. *Artif. Intell.* **1999**, *114*, 3–55.
40. Thrun, S.; Bennewitz, M.; Burgard, W.; Cremers, A.B.; Dellaert, F.; Fox, D.; Hähnel, D.; Rosenberg, C.; Roy, N.; Schulte, J.; Schulz, D. MINERVA: A second-generation museum tour-guide robot. Proceedings of the 1999 IEEE International Conference on Robotics and Automation, Detroit, MI, USA, 10–15 May 1999; Volume 3, pp. 1999–2005. <https://doi.org/10.1109/ROBOT.1999.770401>.
41. Del Duchetto, F.; Baxter, P.; Hanheide, M. Lindsey the tour guide robot: Usage patterns in a museum long-term deployment. In Proceedings of the 28th IEEE International Conference on Robot and Human Interactive Communication, New Delhi, India, 14–18 October 2019; pp. 1–8. <https://doi.org/10.1109/RO-MAN46459.2019.8956329>.
42. Smithsonian Organization and Audience Research. *An Evaluation of Pilot Pepper Robot Program*; Smithsonian Institution: Washington, DC, USA, 2018. Available online: https://web.archive.org/web/20250429005020/https://soar.si.edu/sites/default/files/reports/pilot_pepper_robot_program_evaluation_190320.pdf (accessed on 31 July 2026).
43. Castellano, G.; De Carolis, B.; Macchiarulo, N.; Vessio, G. Pepper4Museum: Towards a human-like museum guide. In Proceedings of the AVI²CH Workshop on Advanced Visual Interfaces and Interactions in Cultural Heritage, Island of Ischia, Italy, 29 September 2020; pp. 1–5.
44. Germak, C.; Lupetti, M.L.; Giuliano, L.; Kaouk Ng, M.E. Robots and cultural heritage: New museum experiences. *J. Sci. Technol. Arts* **2015**, *7*, 47–57. <https://doi.org/10.7559/citarj.v7i2.158>.
45. Vidovićová, L.; Menšíková, T. Materiality, corporeality, and relationality in older human-robot interaction, OHRI. *Societies* **2023**, *13*, 15.
46. Rojas, L.; Romero, J.A.; Manrique, R. Improving autonomy and natural interaction of Pepper robot via large language models. *SN Comput. Sci.* **2026**, *7*, 328. <https://doi.org/10.1007/s42979-026-04892-y>.
47. Garelo, L.; Cocchella, F.; Sciutti, A.; Catalano, M.G.; Rea, F. Next-gen museum guides: Autonomous navigation and visitor interaction with an agentic robot. *arXiv* **2025**, arXiv:2507.12273.
48. Garelo, L.; Cocchella, F.; Catalano, M.G.; Sciutti, A.; Rea, F. One robot, many minds: Factors shaping visitors' evaluation of an autonomous museum robot guide. In Proceedings of the 13th International Conference on Human-Agent Interaction, HAI '25, Yokohama, Japan, 10–13 November 2025; Association for Computing Machinery: New York, NY, USA, 2026; pp. 67–75. <https://doi.org/10.1145/3765766.3765777>.
49. Iio, T.; Satake, S.; Kanda, T.; Hayashi, K.; Ferreri, F.; Hagita, N. Human-like guide robot that proactively explains exhibits. *Int. J. Soc. Robot.* **2020**, *12*, 549–566.
50. Şahin, İ.; Cakmakoglu Arici, N.; Koç, D.E. Tour guiding technologies: A bibliometric analysis, mapping trends and future research agenda. *J. Hosp. Tour. Technol.* **2026**, advance online publication. <https://doi.org/10.1108/JHTT-12-2024-0867>.
51. Rosa, S.; Randazzo, M.; Landini, E.; Bernagozzi, S.; Sacco, G.; Piccinino, M.; Natale, L. Tour guide robot: A 5G-enabled robot museum guide. *Front. Robot. AI* **2024**, *10*, 1323675. <https://doi.org/10.3389/frobt.2023.1323675>.
52. Kahneman, D. *Thinking, Fast and Slow*; Farrar, Straus and Giroux: New York, NY, USA, 2011.
53. Vanderelst, D.; Winfield, A. An architecture for ethical robots inspired by the simulation theory of cognition. *Cogn. Syst. Res.* **2018**, *48*, 56–66.
54. Yang, Z.; Raman, S.S.; Shah, A.; Tellex, S. Plug in the safety chip: Enforcing constraints for LLM-driven robot agents. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024; pp. 14435–14442. <https://doi.org/10.1109/ICRA57147.2024.10611447>.
55. Pesjak, D.; Žabkar, J. Robot planning via LLM proposals and symbolic verification. *Mach. Learn. Knowl. Extr.* **2026**, *8*, 22. <https://doi.org/10.3390/make8010022>.
56. Durante, Z.; Gong, R.; Sarkar, B.; Wake, N.; Taori, R.; Tang, P.; Lakshminanth, S.; Schulman, K.; Milstein, A.; Vo, H.; et al. An interactive agent foundation model. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 11–12 June 2025; pp. 3652–3662. <https://doi.org/10.1109/CVPRW67362.2025.00350>.

57. Latour, B. *Science in Action: How to Follow Scientists and Engineers through Society*; Harvard University Press: Cambridge, MA, USA, 1987.
58. Akrich, M. The de-scription of technical objects. In *Shaping Technology/Building Society: Studies in Sociotechnical Change*; Bijker, W.E., Law, J., Eds.; MIT Press: Cambridge, MA, USA, 1992; pp. 205–224.
59. Suchman, L. *Human-Machine Reconfigurations: Plans and Situated Actions*; Cambridge University Press: Cambridge, UK, 2007.
60. Anderson, L. Analytic autoethnography. *J. Contemp. Ethnogr.* **2006**, *35*, 373–395. <https://doi.org/10.1177/0891241605280449>.
61. Timmermans, S.; Tavory, I. Theory construction in qualitative research: From grounded theory to abductive analysis. *Sociol. Theory* **2012**, *30*, 167–186. <https://doi.org/10.1177/0735275112457914>.
62. Haraway, D. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Fem. Stud.* **1988**, *14*, 575–599.
63. Flyvbjerg, B. Five misunderstandings about case-study research. *Qual. Inq.* **2006**, *12*, 219–245.
64. Goffman, E. The interaction order: American Sociological Association, 1982 presidential address. *Am. Sociol. Rev.* **1983**, *48*, 1–17. <https://doi.org/10.2307/2095141>.
65. Marge, M.; Bonial, C.; Byrne, B.; Cassidy, T.; Evans, A.W.; Hill, S.G.; Voss, C. Spoken language interaction with robots: Recommendations for future research. *Comput. Speech Lang.* **2022**, *71*, 101255. <https://doi.org/10.1016/j.csl.2021.101255>.
66. Skantze, G.; Irfan, B. Applying general turn-taking models to conversational human-robot interaction. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, Melbourne, Australia, 4–6 March 2025; pp. 859–868. <https://doi.org/10.1109/HRI61500.2025.10973958>.
67. Hu, Z.; Zhu, Q.; Zhao, R.; Liang, D.; Yan, H.; He, Y.; Gui, L. Beyond RAG for agent memory: Retrieval by decoupling and aggregation. *arXiv* **2026**, arXiv:2602.02007.
68. Ichter, B.; Brohan, A.; Chebotar, Y.; Finn, C.; Hausman, K.; Herzog, A.; Ho, D.; Ibarz, J.; Irpan, A.; Jang, E.; et al. Do as I can, not as I say: Grounding language in robotic affordances. In Proceedings of the 6th Conference on Robot Learning (CoRL 2022), Auckland, New Zealand, 14–18 December 2022; Liu, K., Kulic, D., Ichnowski, J., Eds.; Proceedings of Machine Learning Research (PMLR): Online, 2023; Volume 205, pp. 287–318. <https://proceedings.mlr.press/v205/ichter23a.html>.
69. Gallino, L. *L'attore Sociale: Biologia, Cultura e Intelligenza Artificiale*; Einaudi: Turin, Italy, 1987.
70. Grimaldi, R. (Ed.) *La società dei Robot*; Mondadori Università: Milan, Italy, 2022.
71. Latour, B. *Reassembling the Social: An Introduction to Actor-Network-Theory*; Oxford University Press: Oxford, UK, 2005.
72. Bourdieu, P. *The Field of Cultural Production*; Columbia University Press: New York, NY, USA, 1993.
73. Bourdieu, P. *Language and Symbolic Power*; Harvard University Press: Cambridge, MA, USA, 1991.
74. Katzenbach, C.; Ulbricht, L. Algorithmic governance. *Internet Policy Rev.* **2019**, *8*, 1–18. <https://doi.org/10.14763/2019.4.1424>.
75. Ananny, M.; Crawford, K. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc.* **2018**, *20*, 973–989.
76. Foster, M.E.; Alami, R.; Gestranis, O.; Lemon, O.; Niemelä, M.; Odobez, J.-M.; Pandey, A.K. The MuMMER project: Engaging human-robot interaction in real-world public spaces. In *Social Robotics*; Springer: Cham, Switzerland, 2016; pp. 753–763. https://doi.org/10.1007/978-3-319-47437-3_74.
77. Niemelä, M.; Heikkilä, P.; Lammi, H.; Oksman, V. A social robot in a shopping mall: Studies on acceptance and stakeholder expectations. In *Social Robots: Technological, Societal and Ethical Aspects of Human-Robot Interaction*; Korn, O., Ed.; Springer: Cham, Switzerland, 2019; pp. 119–144. https://doi.org/10.1007/978-3-030-17107-0_7.
78. Cantucci, F.; Marini, M.; Falcone, R. Effects of robot's adaptive autonomy on users' experience in a museum scenario. In Proceedings of the WOA 2024: 25th Workshop "From Objects to Agents", Bard, Italy, 8–10 July 2024; Alderighi, M., Baldoni, M., Baroglio, C., Micalizio, R., Tedeschi, S., Eds.; pp. 5–19.
79. Doğan, F.I.; Torre, I.; Leite, I. Asking follow-up clarifications to resolve ambiguities in human-robot conversation. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, Sapporo, Japan, 7–10 March 2022; pp. 461–469. <https://doi.org/10.1109/HRI53351.2022.9889368>.
80. European Parliament and Council. Regulation (EU) 2016/679, General Data Protection Regulation. 2016. Available online: <https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng> (accessed on 17 June 2026).
81. European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence, Artificial Intelligence Act. 2024. Available online: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng> (accessed on 17 June 2026).
82. ISO 31101:2023; Robotics, Application Services Provided by Service Robots, Safety Management Systems Requirements. International Organization for Standardization: Geneva, Switzerland, 2023.
83. Pasquale, F. *New Laws of Robotics: Defending Human Expertise in the Age of AI*; Harvard University Press: Cambridge, MA, USA, 2020.
84. Crawford, K. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*; Yale University Press: New Haven, CT, USA, 2021.

85. van Dijck, J.; Poell, T.; de Waal, M. *The Platform Society: Public Values in a Connective World*; Oxford University Press: Oxford, UK, 2018.
86. Chen, V.; Talwalkar, A.; Brennan, R.; Neubig, G. Code with me or for me? How increasing AI automation transforms developer workflows. *arXiv* **2025**, arXiv:2507.08149.
87. Watanabe, M.; Li, H.; Kashiwa, Y.; Reid, B.; Iida, H.; Hassan, A.E. On the use of agentic coding: An empirical study of pull requests on GitHub. *ACM Trans. Softw. Eng. Methodol.* **2026**. <https://doi.org/10.1145/3798166>.
88. Punziano, G. Adaptive epistemology: Embracing generative AI as a paradigm shift in social science. *Societies* **2025**, *15*, 205. <https://doi.org/10.3390/soc15070205>.
89. Hassan, A.E.; Li, H.; Lin, D.; Adams, B.; Chen, T.-H.; Kashiwa, Y.; Qiu, D. Agentic software engineering: Foundational pillars and a research roadmap. *arXiv* **2025**, arXiv:2509.06216.
90. Galster, M.; Mohsenimofidi, S.; Lulla, J.L.; Abubakar, M.A.; Treude, C.; Baltes, S. Configuring agentic AI coding tools: An exploratory study. *arXiv* **2026**, arXiv:2602.14690.
91. Gloaguen, T.; Mündler, N.; Müller, M.; Raychev, V.; Vechev, M. Evaluating AGENTS.md: Are repository-level context files helpful for coding agents? *arXiv* **2026**, arXiv:2602.11988.
92. Zhou, X.; Chen, V.; Wang, Z.Z.; Neubig, G.; Sap, M.; Wang, X. TOM-SWE: User mental modeling for software engineering agents. *arXiv* **2025**, arXiv:2510.21903.
93. Li, H.; Zhang, H.; Hassan, A.E. AIDev: Studying AI coding agents on GitHub. *arXiv* **2026**, arXiv:2602.09185.
94. Agarwal, S.; He, H.; Vasilescu, B. AI IDEs or autonomous agents? Measuring the impact of coding agents on software development. *arXiv* **2026**, arXiv:2601.13597.
95. Horikawa, K.; Li, H.; Kashiwa, Y.; Adams, B.; Iida, H.; Hassan, A.E. Agentic refactoring: An empirical study of AI coding agents. *arXiv* **2025**, arXiv:2511.04824.
96. Sawada, S.; Shirai, T.; Kashiwa, Y.; Yamaguchi, K.; Iwata, H.; Iida, H. To what extent does agent-generated code require maintenance? An empirical study. *arXiv* **2026**, arXiv:2605.06464.
97. Lenzi, F.R.; Delli Paoli, A.; Catone, M.C. What counts as “people” in digital social research? Subject rethinking and its ethical consequences. *Societies* **2025**, *15*, 329. <https://doi.org/10.3390/soc15120329>.
98. Trezza, D.; De Luca Picione, G.L.; Sergianni, C. AI response quality in public services: Temperature settings and contextual factors. *Societies* **2025**, *15*, 127. <https://doi.org/10.3390/soc15050127>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.