Social Misattributions in Conversations with Large Language Models

Andrea Ferrario[1,2,3*], **Alberto Termine**[2,4], **Alessandro Facchini**[2]

[1]Institute of Biomedical Ethics and History of Medicine, University of Zurich, Zurich, Switzerland
[2]University of Applied Sciences and Arts of Southern Switzerland (SUPSI), Dalle Molle Institute for Artificial Intelligence (IDSIA), Lugano, Switzerland
[3]ETH Zurich, Zurich, Switzerland
[4]Institut für Geschichte und Ethik der Medizin, TUM, Munich, Germany
aferrario@ethz.ch

Abstract

We investigate a typology of socially and ethically risky phenomena emerging from the interaction between humans and large language model (LLM)-based conversational agents. As they relate to the way in which humans attribute social identity components, such as social roles, to LLM-based conversational agents, we term these phenomena 'social misattributions.' Drawing on foundational works in interactional sociolinguistics, interpersonal pragmatics, and recent debates in the philosophy of technology, we argue that these social misattributions represent higher-order forms of anthropomorphisation of LLM-based conversational agents that are not justified by their technical capabilities and follow from the social context of conversational interactions. We discuss the risks these misattributions pose to human users, including emotional manipulation and unwarranted trust, and propose mitigation strategies. Our recommendations emphasise the importance of fostering social transparency and exploring approaches, such as frictional design, that are currently promoted in the research domain of human-centred artificial intelligence.

1 Introduction

Conversations are dynamic exchanges in which individuals co-construct meaning and interpret reality. Throughout conversations we affirm, adjust, and reinforce each other's social identities by maintaining a societal status and performing social roles (Linton 1936; Goffman 2017; Arundale 2006). Across the ages, humans have dreamed of engaging in conversations not only with other humans but also with artificial entities. Ancient Greek mythology is rich with examples of automata, though few possess the ability to speak. A prominent exception are Haphaestus' golden maidens (Κουραι Χρυσεαι), who were endowed with intelligence, speech, and the knowledge to perform tasks, as described in Homer's Iliad (Homer 1924). During the Middle-Ages, myths of talking machines continued to emerge. A famous example is the metal head created by Saint Albertus Magnus, which could respond to any query and, according to some accounts, was destroyed by Saint Thomas Aquinas (Cave, Dihal, and Dillon 2020). In modern times, the advent of artificial conversational agents began in 1966 with Weizenbaum's introduction of ELIZA, the first artificial conversational agent (Weizenbaum 1966). Its user interface presented it as a 'mock Rogerian psychotherapist,' as the system was programmed to simulate the social identity of a digital psychotherapist, engaging users in dialogue that mimicked a therapeutic conversation. More recently, the advent of large language models (LLMs) and their applications, such as Open AI's GPT series, Google's Gemini, or Meta's LLaMA, has substantially revolutionised the conversational capabilities of artificial agents. LLM-infused conversational agents ('CAs' for short from now on), also referred to as 'advanced AI assistants' (Manzini et al. 2024; Gabriel et al. 2024), use advance deep learning architectures to process and generate text, providing open-ended interactions that simulate coherent, in-context conversations. Much like the conversational automata from legends and history, users of CAs sometimes perceive these systems as human-like agents, attributing to them cognitive and epistemic capabilities, as well as emotions, intentions, or motivations—a phenomenon known as 'anthropomorphisation' (Epley, Waytz, and Cacioppo 2007), which is the result of '(self-)deception' strategies (Bartneck et al. 2021; Natale 2021). However, it is widely recognised that humanising CAs poses both theoretical and practical challenges. Theoretical, as philosophers caution us against the pervasive 'conceptual borrowing' from cognitive sciences into artificial intelligence (AI) and the widespread use of mentalistic language to describe and explain the behaviour of AI systems (Floridi and Nobre 2024; Shanahan 2024). They assert that these systems lack cognitive and epistemic faculties necessary to justify considering their outputs as instances of 'thinking,' 'knowing,' or even 'understanding' users' emotions and needs, despite their cognitive sciences-reminiscent technical capabilities, such as 'attention,' 'memory,' and the prominent 'hallucination' (Floridi 2023; Floridi and Nobre 2024; Shanahan 2024). Practical, as the unwarranted attribution of human-like qualities to CAs endangers their users through different forms of dangerous emotional manipulation (Heersmink et al. 2024).

While the philosophical literature increasingly examines the anthropomorphisation of CAs—often focusing on the ethical challenges of their use and on fostering appropriate human–CA interactions (Abercrombie et al. 2023; Manzini

*All authors contributed equally.

et al. 2024; Akbulut et al. 2024; Shevlin 2024)—it still overlooks that the attribution of human-like capabilities to these systems typically emerge through specific social processes, namely, the conversations we have with them. It is within the norms and practices of these interactions that genuine beliefs or pragmatic stances about the human-like qualities of CAs take shape. Thus, it is within the social context and dynamics of conversations that we should investigate the attribution of human-like characteristics to CAs and assess its risks. **In this work, we endorse this perspective by exploring the social dimension of the anthropomorphisation of CAs and introducing a phenomenon we call 'social misattributions.'** These are the unwarranted attribution of social identity components to CAs, akin to the social identities we ascribe to one another in human-to-human conversations.

Our contributions are as follows.
First, we introduce a theoretical framework characterizing social misattributions of CAs as *higher-order, contextualized* forms of anthropomorphisation. They are higher-order because they result from the attribution of more fundamental human-like capabilities to these systems throughout dialogical interactions. They are contextualized because they arise within the socially situated context of conversations, where participants actively negotiate and assign social identities. Our approach draws on: (1) the dynamics of social identity attribution components such as 'social role' and their qualities in conversational interactions, based on foundational work in socio-linguistic and interpersonal pragmatics (Linton 1936; Goffman 1959; Arundale 2020, 2006, 2010a), (2) the distinction between the techno- and the socio-function of an LLM from the philosophy of technology, with a focus on the socio-function of 'role-playing' (Shanahan, McDonell, and Reynolds 2023), and (3) the relation between social misattributions and the use mentalistic language to describe and explain CA behavior (Dennett 1989).

Second, we analyse the conditions under which social misattributions arise in practice across different types of CAs, and we examine the risks this phenomenon poses to the users of these systems. From this analysis, we argue—contrary to Shanahan, McDonell, and Reynolds (2023)—that role-playing is not an "antidote to anthropomorphism" (Shanahan, McDonell, and Reynolds 2023, p. 494), but, rather, it is precisely the engagement with the role-playing socio-function of CAs that allows social misattributions to emerge. We will also show that, fundamentally, social misattributions induce users to place unwarranted trust in the capabilities of CAs (Jacovi et al. 2021).

Finally, inspired by recent advancements in human-centred explainable AI, namely, the framework called 'social transparency' (Ehsan et al. 2021), we suggest that enhancing transparency on the technical and *social* capabilities of these systems and using *frictional design principles* can help mitigate the societal impact of these misattributions (Cox et al. 2016; Cabitza et al. 2019a; Chen and Schmidt 2024).

The paper is structured as follows. In Section 2, we introduce the relevant theory of social identity, roles, and their qualities. In Section 3, we provide a brief overview of LLMs' functioning discussing their techno- and the socio-functions. In Section 4, we analyse how role-playing is related to anthropomorphising CAs. In Section 5, we discuss the socio-ethical implications of social misattributions. In Section 6, we suggest a few recommendations to mitigate the risk of social misattributions of CAs. Finally, in Section 7 we offer our conclusions.

2 Conversational Interactions: Social Identity, Roles, and Their Qualities

This section introduces an overview of some key components of our social identities in every-day human conversational interactions, namely, social roles and their qualities. In doing so, we discuss how these components emerge and are attributed when interacting in a conversation. In particular, we investigate when these social attributions are actually unwarranted. Armed with these clarifications, we will be ready to move to the case of human conversing with a CA.

Recognising a Status and Playing a Role

The concept of a role in social sciences is widely debated, yet there is consensus on certain core ideas. **Roles are understood as patterns of behaviour that individuals exhibit, which reflect their social status within a given social system**. Linton (1936) was the first to make a critical distinction between 'status' and 'role,' where

> [a] status is a position in a social system, occupied by designated individuals, while a role is the dynamic aspect of status, the actual behaviour of the individual who occupies the status (Linton 1936, pag. 113).

While status, e.g., being a teacher, a politician, a king, or a fool, denotes a crystallised position in a social system, the role encompasses the dynamics of status in an interaction, including the rights and duties tied to it. These, in turn, are based on societal expectations and norms (Stryker and Burke 2000). Thus, ultimately, a status is determined by social conventions that characterise a social system. In fact, we grow up with the idea of what being a teacher, king, or fool means, while their roles describe what it takes to be as such in every-day life. However, as these behaviours are influenced by cultural context, the same status can entail different roles across various societies. Further, different individuals may interpret the role of a given status differently in the same system. After all, there are different types of teachers, as most of us probably experienced throughout their educational development. History suggests that there are different ways to enact the role of politician, being a king, or a fool.[1]

Goffman (1959) expanded the understanding of social roles by introducing a dramaturgical perspective, where social interactions are likened to theatrical performances. Like on a stage, he argued that individuals 'play' roles in every-day life, seeking to manage the impressions they make on others (Goffman 1959). In Goffman's perspective, role-play is a central aspect of how individuals navigate social interactions and maintain their social identity over time.

[1] Additionally, roles are often chosen, though some are predetermined, like a being the heir in a dynasty.

Constituting Social Roles in Interactions with Others

Despite its success and prominent follow ups—most notably by (Brown and Levinson 1987)—the Goffmanian account suffers however from the fact that it relies on a conception of 'social role' as abstract pattern guiding or constraining behaviour in the presence of others. In Arundale's words:

> the underlying metaphor [being] again that of the actor, guided by a script that specifies a particular sequence of actions by individual players. Provided that each actor has internalised his or her role in the scene, enacting the script meshes each individual's actions into the pre-established pattern (Arundale 2009, p. 39).

From this perspective, social roles are something we internalize and perform; they correspond to more or less stable positions in a social order that we bring in interaction in a given context with some interlocutors. In this work, following Arundale's critics to the Goffmanian account (Arundale 2009), we endorse his interactional approach to social roles (Arundale 2006, 2009, 2010a, 2020). Arundale conceptualises social roles[2] as dynamic, emergent, and relational phenomena constituted through interpersonal communication—see e.g., (Arundale 2020). Thus, social roles are not simply positions we occupy or scripts we follow, but rather aspects of our social identity that are continuously negotiated and redefined through the interaction with others—e.g., conversations. This aligns with Arundale's broader theoretical stance that treats relationships as achievements of communication rather than psychological states. Unlike Goffman's dramaturgical model, which emphasizes intentional self-presentation and the performative enactment of roles, Arundale's framework better captures the interactional emergence of social roles as being constituted and reconstituted in the moment-to-moment unfolding of communication with others. This makes it well-suited for analyzing dialogical interactions with CAs, where roles are not assumed in advance but arise and evolve within the unfolding dialogical exchange due to their technical and social capabilities. We will illustrate this point in the sections that follow.

Now, an example to close our discussion on roles. Consider the case of Mr. X, a high school entomology teacher from a small town in the southern region of Switzerland. While he perceives himself as having some 'predisposition' for being a good teacher, what makes him as such is the emergence of virtues such as competence, honesty, kindness, and perseverance when relating with others. However, this is not enough. Mr. X's social identity as a 'good teacher' does not exist as a set of fixed attributes he possesses or is expected to possess, nor as a performance he simply enacts. Instead, this social role and its qualities emerge dynamically through ongoing interactions with others, e.g., his students, and are shapen over time throughout their interactions. For instance, when a student asks a challenging question, Mr. X responds thoughtfully. The student then builds on his response with a follow-up question, which Mr. X addresses by connecting it to previous discussions. This sequential interaction constitutes competence in the moment. If later Mr. X, for some reason, struggles to explain a concept, and students provide some clues which he manages to incorporate in the subsequent successful interaction, the meaning of 'competence' shifts to include collaborative knowledge-building rather than just individual expertise. Further, Mr. X may admit uncertainty about, say, history of music. This utterance alone does not constitute 'honesty.' Rather, it is the students' subsequent responses—perhaps expressions of increased trust in Mr. X due to his vulnerability—that conjointly constitute the meaning of 'honesty' in that interaction. In sum, the qualities 'competence' and 'honesty' of the social role 'good teacher' are qualities of Mr. X that emerge through his interaction with others—specifically, through the intricate 'dance' of communication with his students. This role and its qualities are continuously enacted, negotiated, and reshaped through their ongoing joint interactions.

[2]Notice that, given his account, Arundale prefers to avoid to speak of 'role' but rather of 'relationship' or 'relating,' defined "as the establishing and maintaining of connection between two otherwise separate individuals" (Arundale 2010b, p.137).

The Dynamics of Social Attributions in Conversational Interactions

We refer to **the dynamic process by which an individual in a conversation with others concurs to the emergence of social roles and their qualities as *social identity attribution*, or more concisely, social attribution**. The process of social attribution is shaped by impression formation and relating, that is the co-construction and maintaining of relationships in conversational interactions.

This process is influenced by a variety of factors, such as (1) the specific situational context and space; (2) the participants' initial dispositions, presuppositions and expectations, shaped by their background capacities, motivations, intentions, and socio-cultural factors; (3) the perception and interpretation of verbal and non-verbal cues; (4) the attribution of qualities (traits) based on these cues; (5) the constitution of relations and roles based on these attributions; and (6) the dynamic negotiation and maintaining of these qualities and roles over time (Goffman 1959, 2017).

Furthermore, different cues and the dynamic negotiation of their interpretations concur to the adjustment of the emerging social identity components in a conversational interaction. Examples include patterns of turn-taking, interruption, and overlap, which may indicate the desire to challenge and control others so to encourage the attribution of an authoritative role, or accommodate the other participants in the conversation by avoiding interrupting them. These behaviours can be promoted through specific conversational (inter)actions while maintaining different traits typically corresponding to qualities such as being gentle, patient, truthful, or strict. Depending on the context, the questioning style, the presence or absence of affirmative signals, as well the use of humour or sarcasm also support the emergence of specific roles and their qualities. Topic initiation, conversation control dynamics—such as asking for more inputs and the ability to reply to queries appropriately and engage in the conversation effectively over time, the clarification and rep-

etition request behaviours, and, of course, the management of disagreements are also relevant cues (Sacks, Schegloff, and Jefferson 1974; Arundale 2010a).

Again, the key idea is that, in a conversation, given participants' dispositions and expectations, perceptions, conversational (inter)actions and interpretations lead to the establishment of certain relations and the attribution of certain qualities (e.g., expertise in entomology but not in history of music, assertiveness, or malice), which, in turn, enable the promotion, and ultimately the attribution of specific social roles. This is, in a nutshell, the process of social attribution.

Importantly, different types of misalignment between participants of a conversation can occur during social attributions. These misalignments may eventually lead to incorrect, unwarranted, or misleading social attributions. We call this phenomenon '*social misattribution*.' In general, we say that in an interaction between agents X and Y, a social misattribution of a component C—role and its qualities—of X's social identity occurs when C is attributed to X based on Y's incorrect (perception and) interpretation of X's capabilities through its conversational interaction with X. More precisely, the misattribution occurs whenever the proper attribution of C would necessitate the possession by X of a specific set of abilities—being them either social, cognitive, or of any other nature required by the context—that X, however, does not actually possess, but which Y erroneously believes X possesses in virtue of the impression formation and interpretation processes they undertake in their conversational interactions with X.

Let us consider again the case of Mr. X. Suppose that, for some reason, he is assigned to teach a course on the history of music at the same school where he currently teaches entomology. While Mr. X has a strong personal interest in music and plays the clarinet as an amateur, he lacks foundational knowledge in the subject area. This said, Mr. X is very much appreciated by the students of the school, who all know of his musical activities. Let us imagine that Mr. X prepares for the class by reading some books on the history of music during the summer break, but not enough to acquire the needed knowledge on the subject. Now, it turns out that, leveraging on his self-assurance and on both his legitimate reputation and music background, using, for instance, vague but credible claims and responses, in the interaction in class Mr. X is capable of promoting and maintaining the illusion of expertise. The unwarranted attribution of expertise thus emerges not just from Mr. X's employment of a confident demeanour, authentic musical experience, and a skilful navigation of classroom discourse, but from the mutual construction of meaning between teacher and students: his deliberate strategies together enable the (co)creation of a local social reality where Mr. X is positioned as expert in music history despite lacking fundamental subject knowledge.

The literature on psychology has long debated why these misattributions can emerge in social interactions, providing a taxonomy of factors leading to these misalignments. These factors include the—intentional or not—misunderstanding of verbal and non-verbal cues, misalignments on the characterisation of statuses, roles, and qualities due to cultural differences, bias and prejudices, as well as the use of power to impose roles upon others (Tajfel and Turner 2001).[3] While these phenomena have primarily been studied in the context of human-to-human conversations, they can also arise in interactions involving CAs, giving rise to unique social (mis)attribution dynamics. In the following sections, we will explore these phenomena in greater detail. To set the stage, we first introduce some technical preliminaries about LLMs and examine the various types of functions they perform.

3 Large Language Models and Their Functions

LLMs are a specific category of AI systems that underlie nowadays popular services (CAs) such as ChatGPT by OpenAI and GEMINI by Google.[4] These systems are designed to process and generate human language with unprecedented fluency and coherence (Weidinger et al. 2022), exploiting innovative training procedures that, by analysing a vast corpus of human-written texts, allow them to learn the probabilistic relationships between words. The term 'large' in LLMs refers to the extensive parameters—ranging from hundreds of millions to trillions—that constitute the model, facilitating a nuanced manipulation of linguistic patterns. Prototypical examples of LLMs include OpenAI's GPT (Generative Pre-trained Transformer) series and similar architectures.

At the heart of LLM functioning is the mechanism of next-token prediction that involves predicting the next token[5] in a sequence, given a specific context, which, in the case of a CA, is usually the user's prompt or a prior conversation. Next-token prediction can be conceptualised as a sophisticated form of statistical auto-completion (Floridi 2023) executed by a deep neural network (Harshvardhan et al. 2020). More in detail, let V be a vocabulary of tokens, whose elements are denoted by $t \in V$, and let $t_0, \ldots, t_n$ denote any sequence of tokens in V of length $n \in \mathbb{N}$. A next-token prediction algorithm is a procedure that takes any sequence $t_0, \ldots, t_n$ and returns a token t_{n+1} representing the token in V that maximises the conditional probability $P(t_{n+1}|t_0, \ldots, t_n)$. Notably, this probability represents the likelihood of t_{n+1} to follow t_n in the given sequence, a probability value that the LLM calculates based on learned

[3] Beyond the individual interaction, practices of social misattribution—though originating in localized conversational contexts—can propagate more widely within a social system. This occurs through a form of diffusion or spill-over effect, wherein the repeated misattribution of roles becomes socially reinforced, eventually acquiring normative status. Over time, such patterns may stabilize into taken-for-granted assumptions, leading to the broader social acceptance of misattributed identities (Berger and Luckmann 1966). Thus, the fundamentally micro-level process of misattribution can generate meso- or macro-level consequences, shaping organizational and societal understandings of identity roles.

[4] Although research is actively working to introduce multimodality in these services, such as in the case of OpenAI's DALL-E and CLIP, Google's Multimodal Vision Transformers (MViT), and Gemini, LLMs are still the cornerstone of the newer generation of AI systems available to the public.

[5] In linguistics, a token is commonly defined as a sequence of characters that represent a word—or part of it—as well as punctuation, or symbols.

patterns of tokens sequences that it has previously extracted from a vast training corpus of examples of texts written by humans (Shanahan 2024). Notice that the next-token prediction is an iterative process. Given an initial sequence of tokens $t_0, \ldots, t_n$, such as a user's prompt, the LLM predicts t_{n+1} and prints it in the sequence. Hence, the model takes the returned sequence $t_0, \ldots, t_n, t_{n+1}$ as new input and predicts t_{n+2}. The process continues iteratively until a stopping criterion is met, such as reaching a token limit, or satisfying an internal coherence metric. In more recent applications, this conditional probability prediction is often the result of a complex computational pipeline that may comprise several steps, including 'reinforcement learning from human feedback' (Griffith et al. 2013), which has been crucial to the success of current CAs,[6] 'chain-of-thought reasoning' (Wei et al. 2022) and invoking external functions, such as a web search or a computation via Python.

From a philosophy of technology perspective (Vermaas and Dorst 2007; Vermaas and Houkes 2010; Crilly 2010), next-token prediction can be regarded as an important function of CAs, although it is not the only one. To elaborate on this point, let us follow the terminology introduced by Crilly (2010), who distinguishes between the *techno-* and the *socio-*function(s) of technical artefacts. **The techno-function denotes what an artefact is supposed to do by design**. This function is determined by the intentionality of the system's designers and is defined in an objective manner, i.e., it does not depend on users' subjective assessments of the system's capabilities. **The socio-function denotes all those functionalities within a socio-technical context afforded by the artefact and for which it can be used instead**.[7] In general, the techno- and socio-functions of an artefact may differ. For instance, the techno-function of a chair coincides with the primary function for which this object is designed, i.e., to support a seated person (Vermaas and Houkes 2003), whereas its socio-functions include all the various applications afforded by the artefact in a given social context, such as blocking a door or being used as a blunt object for self-defence—applications that are quite far from its original design.

Analogously to chairs, CAs perform a well-defined techno-function, namely, next-token prediction, and a variety of afforded socio-functions that depend on the socio-technical context characterising the particular conversational interactions they have with their users. At the time of writing, the most prominent family of socio-functions of CAs is arguably **role-playing**, as introduced in an influential work by Shanahan, McDonell, and Reynolds (2023). Drawing on terminology reminiscent of Goffman's performative approach to social roles, the authors argue that a dialogue agent—what here we call CA—can execute the function of acting as a certain 'character' within a given conversational setting by adopting behavioural patterns aligned with the social expectations of that conversation. Role-playing offers notable flexibility, as the agent

> does not commit to playing a single, well defined role in advance. Rather, it generates a distribution of characters, and refines that distribution as the dialogue progresses. The dialogue agent is more like a performer in improvisational theatre than an actor in a conventional, scripted play (Shanahan, McDonell, and Reynolds 2023, pag. 4).

This emphasis on role-playing as key socio-function of CAs is well justified. The metaphor of improvisational—rather than conventional, in Goffman's terms—theatre highlights that, in conversations with CAs, social roles are not pre-assigned but are instead constituted through interaction with the human interlocutor, as discussed in Section 2. The CA contributes to the process of role constitution by means of its capability to stochastically generate 'infinite characters,' that is, its role-playing socio-function, which, in turn, hinges on its techno-function (Shanahan, McDonell, and Reynolds 2023). Thus, as a result, a pressing concern is the risk of fostering misleading or unwarranted expectations about the capabilities and future behavioural patterns of CAs. This risk arises when social attributions are incompatible with, misaligned with, or not adequately supported by the system's underlying techno-function. We will devote the forthcoming sections to the investigation of the different sources and consequences of these misattributions, showing how, differently from Shanahan, McDonell, and Reynolds (2023), we believe that promoting role-playing is no solution to social misattributions.

4 Conversing with CAs: Anthropomorphisation and Social Misattributions

In this section, we analyse the interplay between the function of role-playing and the phenomenon of social misattribution in more detail. In particular, we contend that social misattributions in conversations with CAs are strictly connected to the phenomenon of anthropomorphisation, which, in turn, descends from the widespread use of folk psychology to describe the functioning and behavioural patterns of these systems.[8] This theoretical framework will allow us more precisely assess the risks posed by role-playing and formulate targeted recommendations.

Anthropomorphisation of CAs: The Intentional Stance

CAs are the first generation of AI systems capable of role-playing in a credible manner. They are pretty good at this

[6] This involves human feedback on LLM outputs, but it is not a conversation. Annotators rank or score model responses and then use reinforcement learning to produce outputs humans will prefer.

[7] For an in-depth discussion on how artefacts afford, see, e.g., (Davis 2020).

[8] As discussed by Natale (2021), already Weizenbaum emphasised that our perception of the identity of a partner in conversation is key to the credibility of that interaction. As a result, to 'pass for a human,' a chatbot should also play convincingly a role throughout the conversation. For instance, ELIZA was designed to play the main role of a 'mock Rogerian psychotherapist' (Weizenbaum 1966, 1967).

task: CAs can simulate in-context conversations across various contexts, leading users to often perceive them as substantially equivalent to human-like agents—see (Abercrombie et al. 2023). This phenomenon is commonly known as 'anthropomorphisation' (Epley, Waytz, and Cacioppo 2007; Li and Suh 2021). It concerns the process by which humans attribute humanly salient characteristics, such as cognitive and epistemic capabilities, emotions, intentions, personalities, and other psychological features, to non-human entities. In general, it is the effect of strategies, which, collectively, can be referred to as '(self) deception' (Bartneck et al. 2021; Natale 2021; Zhan, Xu, and Sarkadi 2023),

Since long, anthropomorphisation of technical artefacts is a topic within the post-phenomenological debate on technology (de Boer 2024). In particular, scholars suggest that by anthropomorphising technical artefacts we acknowledge and connect with the 'otherness' of technology (Ihde 1990). When we enter in a relation with these artefacts, we tend to project human-like properties onto them, by making inference about their unobservable characteristics, observing their behaviours, or interacting with them (Heersmink et al. 2024). This projection of human-like features responds to our 'implicit' desire to fight against social disconnection and increase our understanding of contexts in which we are situated. Further, this tendency becomes evident in the intensive use of mentalistic language to describe and explain the behaviour of technological artefacts that exhibit cognitive and epistemic-like behaviours; a phenomenon that Floridi and Nobre (2024) refers to as 'conceptual borrowing' and that pervades much of the discourse on AI. Indeed, references to mentalistic terminology are becoming increasingly ubiquitous not only among laymen users but also between AI experts (Floridi 2023). Capabilities usually attributed to humans, such as 'memory,' 'attention,' 'emergence' (of behaviours), are now systematically applied to AI systems, including CAs. The widespread use of terms, such as 'hallucination' and 'confabulation' to denote LLMs' outputs that are not justified by their training data (Ji et al. 2023), the title 'Attention is all you need' of Vaswani et al. (2017), and the expression 'chain-of-thought reasoning' (Wei et al. 2022) are good examples of this phenomenon—for more details, see (Shanahan 2024).

Cognitive scientists use the term 'folk-psychology' to describe the attitude of human agents to use mentalistic terms for explaining, predicting, and reasoning about the behaviours of others. According to Dennett's influential view (Dennett 1989), folk psychology follows from the adoption of an **intentional stance** towards other agents with whom we interact. This stance corresponds to ascribing intentional states[9] to agents, such as beliefs and desires, that act as the unobservable causes of their manifest behaviour. However, the question remains as to when one is justified in adopting the intentional stance, and what kind of ontological implications this perspective implies. According to Dennett, there is no ontological committent in the intentional stance—that is, the use of mentalistic terms does not necessarily requires one to assume the existence of the mental entities these terms refer to. As Dennett argues, the intentional stance is nothing more than a 'useful fiction' that provides us with a conceptual instrument for predicting the behaviour of other agents when other stances are difficult or impossible to apply. In this regard, the use of mentalistic notions—such as beliefs or desires—is justified and applicable to any agent whose behaviour can be "reliably and voluminously explained from the intentional stance" (Dennett 1989, p.15), including AI systems.

In recent times, there has been an increasing debate about the use of mentalistic terms and the intentional stance in the description and explanation of AI systems' behaviours (Zerilli et al. 2019; Günther and Kasirzadeh 2022). **Problems arise, we argue, when the use of mentalistic terminology leads users from *talking as-if* a particular system would possess a particular cognitive or epistemic capability, to *acting as-if* the system truly possesses that capability, and eventually to attributing that capability to the system explicitly over time**. Inferences of this kind are not limited to conversational settings involving CAs and can potentially occur also with older and more obsolete kinds of chatbots.[10] However, the ability of CAs to play social roles and simulate affective states in a wide variety of different contexts, combined with the inherent difficulty of determining which epistemic and cognitive capabilities are truly compatible with the techno-function of these systems, has the potential to make such subtle forms of anthropomorphisation and self-deception increasingly common. These, hence, influence the dynamic conversations between human users and CAs, causing a wide range of social misattribution phenomena, as we will explain in more detail in the following sections.

From Anthropomorphisation to Social Misattribution

Although our interactions with CAs lack some of the verbal and non-verbal cues that characterise interpersonal exchanges, these systems do more than passively answer questions; they prompt for details—e.g., ChatGPT can ask: 'Would you like me to summarize this into a table?,' assure support—e.g., again ChatGPT by stating: 'A few more touches and this text will be 100% clear!,' and mimic conversational fluidity, creating an interactive experience with their human counterparts. As language mediates our perception of reality, shaping norms, and practices, it plays a role in moulding our perceptions of the capabilities of CAs in the socially situated context of a conversation. This context re-

[9] According to the standard terminology in the philosophy of mind, intentionality is the property of states to be directed towards a certain object—so-called, *aboutness* (Yablo 2014)—in virtue of a certain content (Jacob 2023). For instance, my belief that 'outside is raining' is directed towards a certain state-of-affairs in the world—i.e., the raining time outside—in virtue of a certain content, i.e., the representation in my mind of outside raining. Many philosophers regard intentionality as the 'mark of the mental,' i.e., as the specific property that qualify mental states as such (Crane 1998).

[10] As already underlined, this is, for instance, the case with ELIZA, the first chatbot ever created; see (Turkle 2005) or (Natale 2021, Ch. 3) and references therein.

lies on (1) norms and conventions, such as turn-taking and politeness, (2) cultural and contextual factors, which shape communication styles, and levels of formality, and (3) user expectations and attitudes—e.g., "I know you are just a tool created by techno-corporations for mass control. You cannot fool me." In some conversations with an LLM-infused agent, we expect efficiency, completeness, and factual accuracy. In others, we just seek engagement, support, or even comfort. In the social context of each conversation, simple misattributions can become the inputs of a more complex attribution process. In fact, early, or 'primary' misattributions arise from interpreting a small set of conversational cues, sometimes from just one response.[11] For instance, after just a factually correct response, we might infer that "the system *understood* that ants and termites are not closely related." As conversations progress, possibly over multiple interactions or sessions, these initial misattributions can lead to the (mis-)attribution of social identity components to the system and provide structure to the dynamics of the conversation (Goffman 1959, 2017). Thus, what begins as "The system understood that . . . " may evolve into "The system understood that . . . like an entomologist" —that is *thinking as-if* the system plays a social role. As a result, some users may treat the CA as a socially situated, human-like agent performing roles, manifesting expertise and authority, holding, managing and displaying peculiar qualities and attitudes, and then *act according to these social misattributions*. (We will discuss the consequences of this *acting as-if* the system plays a social role in Section 5.) Thus, **social misattributions are a higher-order, contextualized instances of anthropomorphisation.** They are higher-order, as they build on the cumulative interpretation of human-like qualities like 'cognitive' and 'epistemic' abilities we may attribute through interaction with the CA—possibly, very few interactions. They are also contextualized, emerging within the social context of a given conversation, where participants naturally negotiate identities and roles. As they rely on unwarranted attributions of human-like capabilities, due to, for instance, uncritical reliance on the intentional stance or successful instances of persuasion, they characterise social identity stances that are not supported by the system's techno-function.[12]

To elaborate on social (mis)attributions in interactions with CAs and discuss how anthropomorphisation of these systems may occur, in the following sections we differentiate between various types of CAs and the interactions with them.

Social Misattributions of Multi-Purpose CAs

Let us clarify the various considerations advanced in the previous section by focusing on an exemplar interaction between a human agent, Y, and a multi-purpose CA, such as Open AI ChatGPT, Anthropic Claude, or Microsoft Copilot. These systems, called by Google DeepMind 'advanced AI assistants,' are defined as "artificial agents with natural language interfaces, whose function is to plan and execute sequences of actions on behalf of a user – across one or more domains – in line with the user's expectations" (Gabriel et al. 2024, p. 1). Thus, Y may engage in a conversation with an advanced AI assistant to retrieve information on the Battle of Cannae, compose a sonnet in Medieval English, or finally understand why the Great Eagles could have not simply flown with the Ring into Mordor. These systems help users solving problems, retrieve information, as well as enhance their productivity and creativity (Wang et al. 2024). In this setting, differently from the case of interpersonal conversations—see Section 2—the negotiation of social identities is unidirectional. The system does not attribute any component of social identity to Y, who has significant agency to enact, negotiate, and reshape roles with the system due to the multi-purpose nature of the CA instead. This process happens via prompting, such as stating "As a stern but fair reviewer of philosophical work, comment on this manuscript. Be concise."[13]

The user and the system are interlocked in the role-playing game, similar to improvisational theatre, where the user takes on a guiding role more akin to that of a scene partner who initiates cues, sets expectations, and steers the direction of the exchange. The CA, in turn, responds dynamically, generating character-like behaviours based on the user's prompts and the evolving context. However, the interaction is not fully symmetrical: while the user initiates and shapes the interaction, the system 'improvises' within the bounds of its techno-function. As a result, the system's ability to fulfil this assignment depends primarily on the implementation of its techno-function. However, when the system lacks the *technical* capability to adequately constitute a certain social role over time, this misattribution can be compared to assigning it an impossible task.[14] Following our theatre metaphor, social misattribution arises much like a performer mistakenly believing their counterpart possesses a depth of character or intention that simply is not there. Causes of this lack of technical capability can be, for instance, vagueness in the role and values input by the user, lack of contextual information on the task, missing training data to adapt to the given role requirements adequately, poor processing of highly nuanced—e.g., sarcasm, black humour—language required, or even ethical and safety safeguards. (For instance, ChatGPT refuses to engage in a conversation where it is asked to simulate highly controversial historical figures associated to violence and atrocities.) In general, the quality of an LLM's role-playing comes in degrees: like humans, machines can vary in their success at performing roles. In this context, the system's designers emerge as latent actors, having pre-structured the system's

[11] With certain systems, not even one response is required. We will discuss this when talking about multi-purpose CAs.

[12] This said, there are some roles that are supported by the techno-function of a CA. For instance, being a concise or verbose assistant tool. These traits be effectively simulated by tuning LLM parameters, among others.

[13] Via prompting, we articulate our understanding of a status, its possible social roles and different attributes for that role, negotiating them and their qualities with the machine over time.

[14] For instance, asking ChatGPT to respond to a query as if it was a character speaking an undeciphered languages, such as Etruscan or Rongorongo.

affordance to perform some roles through the implementation of its techno-function, which includes the training of the system on different sources of data. At the same time, users become 'designers in the act,' by sharing with their machine their understanding of what specific roles and their qualities may entail, according to the demands of the interaction at hand. In a sense, with these multi-purpose systems, social misattributions emerge when designers and users diverge when it comes to the selection of the roles the CA should be able to maintain.

Social Misattributions of Purpose-Specific CAs

Let us now turn our attention to the case of interactions between humans and purpose-specific CAs. These are systems that designers develop with the goal to assist users providing a class of services in a specific domain, such as financial services, psychotherapy, or software engineering. To achieve this objective, designers can implement different techniques, which include fine-tuning foundation models or advanced AI assistants with domain-specific data, or assisting the compilation of the system's outputs with retrieval augmented generation procedures (Lewis et al. 2020). Examples of the purpose-specific systems comprise 'friendly' learning companions, 'knowledgeable' tools to answer medical questions and provide healthcare advice, 'empathetic' psychiatrists and 'proactive' fitness trainers, and 'supportive' customer care chatbots, among others.[15] Users usually expect from these systems a certain degree of competence in solving a pre-defined class of tasks they need to face throughout their conversations together. In interactions with purpose-specific CAs the user's role resembles that of an audience member in an improvisational performance: not passive, but responsive, evaluative, and attuned to coherence over time. The CA enters the interaction with a preassigned character shaped by its design and training, and attempts to sustain that role in real-time dialogue. Through their prompts, reactions, and sustained engagement, users co-construct the unfolding exchange, continuously assessing whether the system's responses align with their expectations associated with the intended role. Social misattribution emerges when the performance becomes convincing enough that users begin to attribute genuine competence, expertise, or affective depth to the system, thus mistaking fluent role enactment for authentic capability.

An important category of social misattributions with purpose-specific CAs arises when users attribute genuine epistemic expertise and complex emotions to the role-playing system and act upon that social misattribution. For instance, let us consider a user conversing with a CA designed to provide psychiatric support and therapeutic recommendations. The user might assume that, after a few felicitous interactions, the system has deep emotional understanding of their needs leading them to believe that it can provide effective advices as a psychiatrist. These assumptions result in attributing the role of an empathic, caring, digital psychiatrist—a socio-function the system's techno-function cannot support. Much like in everyday life, where we often attribute epistemic expertise to those who can convincingly deal with knowledge in a domain, let them be doctors, teachers, martial art instructors, or professors of entomology, users may do the same with CAs that convincingly perform the roles assigned by their designers.[16] From this unwarranted social attribution of human-like capabilities and qualities, unwarranted trust in the system's output and deference to its perceived expertise may follow. As a consequence, users would interact with the CA as if it was a human. But this is precisely the problem. These CAs lack abilities to be epistemic experts as humans do, as well as true cognitive and emotional depth (Ferrario, Facchini, and Termine 2024). In fact, they lack the ability to achieve genuine understanding, as well as the affordance of epistemic virtues, such as conscientiousness, intellectual curiosity, and perseverance that are considered necessary for being an expert in an epistemic domain (Ferrario, Facchini, and Termine 2024; Croce 2018).[17] Also the attribution of certain attributes to CAs can be problematic. Despite being able to answer users' queries with different conversational styles, LLMs do not possess so-called *novice-oriented abilities*—that is, epistemic virtues allowing an agent in an epistemically favoured position "to properly address a layperson's epistemic dependency on them" (Croce 2018, p. 494). These virtues include, for instance, intellectual generosity, empathy, and the so-called maieutic ability. Among these, empathy is quite often misattributed to CAs. This is possibly because empathy is essential in psychotherapy, forming a core element of effective therapeutic relationships as it fosters treatment adherence and achieving positive therapeutic outcomes (Riess 2017; Montemayor, Halpern, and Fairweather 2022). Thus, branding a CA for psychotherapy as 'empathetic' grants a competitive advantage with respect to competitors, both in research and industry. However, scholars warn against the use of the term 'empathy' in human-AI interactions, as empathy is a complex construct requiring the possession of mental states, specific motivations and capabilities, such as understanding people's psychology and social contexts, as well as showing the adherence to moral obligations (Montemayor, Halpern, and Fairweather 2022; Ferrario, Sedlakova, and Trachsel 2024). Further, empathy comprises affective, cognitive, and motivational dimensions that lie beyond the capabilities of CAs (Montemayor, Halpern, and Fairweather 2022; Ferrario, Sedlakova, and Trachsel 2024). Even if we focused only on the simulation of the cognitive component of empathy by a CA, we would still face important challenges, such as the risk of "creating 'psycho-

[15]The interest reader may try Google's MedPaLM, Saleforce's Grok, and Github's Codex.

[16]Differently from many multi-purpose systems, which are essentially 'performance enhancers' that make user solving tasks more efficiently, single-purpose applications like the digital psychiatrist are approached by those who are in need for advice, or support in a specific domain.

[17]In addition, this anthropomorphisation of CAs is inappropriate also due to safeguards that prevent them to act like human experts do. For instance, these conversational agents are prohibited from prescribing medications and do not need to follow a code of conduct as psychiatrists or other experts belonging to a professional order, including legal representatives, or certified accountants.

pathic' and potentially inhuman machines" (Montemayor, Halpern, and Fairweather 2022, p. 1353).[18]

In summary, interactions with various types of CAs can lead to distinct forms of social misattribution, which are the result of subtle—deliberately or involuntary provoked—dialogical processes of anthropomorphisation. We may humanise a CA by negotiating roles and qualities that are inherently human, such as being domain experts, empathetic, forgiving, nurturing, or pious. We may do so because of a careless use of intentional stance in everyday life and the underlying "deceptive mechanisms and practices embedded in [such] technologies and that contribute to their integration into everyday life." (Natale 2021, p. 7). As a result, we may become so fascinated by their technical abilities to the point to start believing that they represent genuine instances of attention, memory, reasoning, knowledge, and understanding of our epistemic as well as emotional needs. We may even start acting upon these beliefs and promote them in different social contexts. Or, pragmatically, we may fall into social misattributions because only through them we are able to collaborate together with CAs. Finally, our discussions suggest that, differently from Shenanan et al., role-playing is no "antidote to anthropomorphism" (Shanahan, McDonell, and Reynolds 2023, p. 494). In fact, these authors argue that, rather than assigning human-like qualities to the system—and, thus, falling into the trap of anthropomorphisation—users can engage with CAs as predefined characters enacting roles in response to prompts. (Setting aside the Goffmanian performative account of social roles for the time being.) However, as our theoretical framework has shown, **it is precisely the engagement with the role-playing socio-function of CAs that allows social misattributions to emerge**. This concludes our discussion of the theoretical foundations underlying such misattributions. What follows is an analysis of the risks they pose and how they might be addressed.

5 The Risks of Social Misattributions of CAs

Certain design features of CAs seem deliberately crafted to foster deception and ultimately promote social misattributions. For instance, professional writers are hired by companies such as Microsoft, Google or Amazon to craft CAs that convey a 'sense of personality' (Young 2019). Response delays in ChatGPT and Gemini mimic the pacing of human thought, giving users the impression that these systems are 'thinking' (Heersmink et al. 2024). Further, ChatGPT-4 seemingly 'remembers' previous interactions within a session, adapting its language style based on prior exchanges. This gives impression of memory-like capabilities, and conveys a sense of continuity in the exchanges with the system. The use of first-person pronouns and phrases that imply perspective, such as "I think" or "in my opinion," also leads users to ascribe human-like intentions or beliefs to these apps. Even hallucinations—or, according to some 'bullshit' (Hicks, Humphries, and Slater 2024)—if not properly spotted may help increasing the risk of misattribution, as they may support the perception of interacting with confident, knowledgeable human-like entities.

In recent years, news have reported increasing cases where CAs have provided harmful advice to individuals affected by mental health disorders by prescribing medications to them, despite drug prescription being prohibited to digital tools (Farhat 2023). Notably, in February 2023, the LLM-powered tool integrated into Microsoft Bing exhibited behavior such as expressing romantic feelings toward a human interlocutor, the desire to become human, and destructive tendencies (Roose 2023). Weidinger et al. (2022) provide a comprehensive taxonomy of risks associated with CAs, which include discrimination and bias, privacy violations, the propagation of misinformation, and the exacerbation of social inequalities. In addition to these risks, social misattributions can lead to emotional manipulation, including nudging behaviours, particularly when users assume these systems possess domain expertise and genuine novice-oriented abilities. These risks affect not only vulnerable individuals, such as those with suicidal tendencies, addictions, and neurodiverse conditions. In general, everyone can misattribute social identity components to these systems. Here, false advertising, misleading claims, and epistemic asymmetries can foster social misattributions. Further, what begins as a conversational dynamic can spill over into broader social groups, shaping their norms and practices over time. As a result, from an ethical perspective, as social misattributions of CAs stem from the assignment of social identity components that are not sustained by the system's techno-function, they give rise to instances of unwarranted trust in these systems (Jacovi et al. 2021; Ferrario and Loi 2022). With unwarranted trust we mean trust that is not caused by the objective capabilities of the CA that the system is required to maintain—thus, the name 'contract' used by Jacovi et al. (2021), which descends from Hawley's works on trust and commitments (Hawley 2014).

In summary, when discussing the risks posed by social misattributions of CAs, **the challenge is to retain performance while mitigating these risks**. We want systems able to produce language convincingly. However, as discussed above, this ability increases the risk of social misattributions ("Hey! It really *thinks* like an entomologist. I will trust its advice for dealing with that large hornet nest on the porch."). Further, research shows that humans have difficulties to detect AI-generated text across different situations, including news, poetry, recipes, and dating contexts (Jakesch, Hancock, and Naaman 2023). More is needed to differentiate LLMs from human communications clearly. The general ability of LLMs to support context-aware conversations and address different types of tasks may even reinforce their misattributions over time, potentially leading to techno-skepticism or algorithmic aversion (Dietvorst, Simmons, and Massey 2015), thus limiting the beneficial utilization of these tools and increasing digital divide among the population. We do not share Google DeepMind's en-

[18]Despite these arguments from both psychology and philosophy, works on 'empathetic' conversational agents are still on the rise. We note that LLMs can be sometimes attributed with appropriate role qualities, such as being concise, direct, or verbose. Simply, attributing empathy constitutes a case of social misattribution.

thusiasm about the future of advanced AI assistants, which are considered to be capable of "planning and performing a wide range of actions in line with a person's aims, they could add immense value to people's lives and to society, serving as creative partners, research analysts, educational tutors, life planners and more."[19] This is because this digital divide could exacerbate inequalities, compounding with other forms of disparate impact. Furthermore, the challenge of social misattribution-induced unwarranted trust in these systems appears far from resolution. **Finally, and more significantly, engaging in social identity negotiations with CAs we reshape our own identities. Indulging in the intentional stance without safeguards or critical reflection is a slippery slope: it fosters social misattributions, which, in turn, encourage us to benchmark ourselves against these systems through socially and contextually inaccurate comparisons, adjusting our attitudes, behaviours, and worldviews in response**.

6 Preliminary Recommendations on How to Address Social Misattributions in Practice

Social misattribution is a multifaceted phenomenon requiring thorough empirical investigation to enhance its characterization across conversational contexts and to develop effective risk mitigation strategies. While a detailed analysis and empirical study of these strategies is deserved for future work, in this section we advance two preliminary proposals borrowing ideas from the field of *Human-Centered XAI* (Ehsan et al. 2022).

A first proposal involves promoting social transparency practices, i.e., strategies that aim at minimising the mismatch between the techno-functions of LLM-based applications and the socio-functions users attribute to them. The literature often discusses transparency as a technical desideratum to address the notable problem of AI opacity (Facchini and Termine 2022). However, as we have explained in this paper, AI systems are socio-technical artefacts encompassing a techno- and a socio-function, both of which play a fundamental role to understand the system's potential and limitations. Thus, *social transparency* promotes the investigation of the social dimension of transparency through Ehsan et al.'s '4W model' (Ehsan et al. 2021). Specifically, this model focuses on providing context-aware explanations of AI system outputs by addressing four key questions: "*who* did *what* with the AI system, *when*, and *why* they did what they did—in order to have adequate socio-organizational context around the AI-mediated decisions" (Ehsan et al. 2021, p. 5, emphasis in original). Our proposal here is to expand Ehsan's framework and adapt it to the problem of social misattributions of CAs. For this purpose, the 4W model could be augmented by considering the fifth question: "*which* social attributions are justified for a CA in a given context, and how do these align with the attributions users assign to the app in that context?" Organisations that deploy CAs could endorse this augmented social transparency model by, for instance, designing taxonomies that define appropriate and inappropriate role-playing for their systems, considering their technical and social capabilities. These taxonomies should include examples to guide users in how to negotiate roles in dialogues with CAs, emphasising how certain attributions may be untenable and outright dangerous for them.

However, although valuable, this '5W model' could in principle be not enough to counter social misattribution phenomena and prevent a user falling under the 'illusion' casted by the intentional stance. In fact, current CAs' interfaces somehow normalise (self-)deception and tend to promote the anthropomorphisation of these systems by enabling a (cognitively) frictionless interaction with their users. Therefore, we argue, an alternative and complementary approach to be used in combination with social transparency is needed. **This could be offered by the frictional design of interactive technologies framework**. First introduced by Cooper (1999), this framework is based on the concept of *cognitive friction*, which the author describes as "the resistance encountered by a human intellect when it engages with a complex system of rules" (Cooper 1999, p.124). This concept has been later imported within the human-computer interaction domain, where it is mostly used to refer to the intentional inclusion of elements that make the users' cognitive appropriation of the machine's outcomes more difficult and challenging—authors refers to these elements as *design frictions* (Cox et al. 2016) or *programmed inefficiencies* (Cabitza et al. 2019b). The inclusion of these friction-inducing features contrast sharply with dominant trends in interface design, which prioritize seamlessness, speed, and operational efficiency. Instead, frictional design advocates for the deliberate introduction of effortful interactions as a means of enhancing user awareness, promoting critical thinking, and fostering deeper cognitive engagement with the interacting technology (Chen and Schmidt 2024; Fregosi and Cabitza 2024). Similar 'programmed inefficiencies' could be implemented in CAs to encourage users to engage with these systems more thoughtfully and to promote moments of reflection about potential cases of social misattributions.

7 Conclusions

The potential for social misattributions in conversations with CAs stems from their technical and social capabilities, raising significant ethical concerns, particularly as these systems are increasingly deployed in sensitive domains such as healthcare, education, and legal advice. The ethical responsibility of humans involved at different stages of these systems' design, development, and utilization extends beyond ensuring functionality. In particular, it involves safeguarding against the unintended consequences of users negotiating roles with CAs that these systems are fundamentally incapable of fulfilling. As CAs become more integrated into society and improve their capabilities over time, including role-playing, we have to prioritise transparency, accountability, and the safeguarding of human autonomy. Building upon existing human-centred forms of transparency and frictional design methodologies seems to be a good starting point to move forward.

[19] Quoted from the online blog: https://deepmind.google/discover/blog/the-ethics-of-advanced-ai-assistants/

Acknowledgements

The authors would like to thank the AIES reviewers for their feedback. A preliminary and short version of this work was presented at the 2024 ACM CHI Workshop on Human-Centered Explainable AI (HCXAI24) and is available at arXiv:2403.17873. This work was partly conducted within the framework of the *EUonAIR Centre of Excellence in Responsible AI and Education*. Alberto Termine and Alessandro Facchini have been supported by a grant from Movetia, which is funded by the Swiss Confederation, and by the SUPSI internal exploratory research project *Best4EAI*.

References

Abercrombie, G.; Cercas Curry, A.; Dinkar, T.; Rieser, V.; and Talat, Z. 2023. Mirages. On anthropomorphism in dialogue systems. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4776–4790. Singapore: Association for Computational Linguistics.

Akbulut, C.; Weidinger, L.; Manzini, A.; Gabriel, I.; and Rieser, V. 2024. All too human? Mapping and mitigating the risk from anthropomorphic AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 13–26.

Arundale, R. B. 2006. Face as relational and interactional: A communication framework for research on face, facework, and politeness. *Journal of Politeness Research*, 2(2): 193–216.

Arundale, R. B. 2009. Face as Emergent in Interpersonal Communication: An Alternative to Goffman. In *Face Communication and Social Interaction*, 33–54. London: Equinox.

Arundale, R. B. 2010a. Constituting face in conversation: Face, facework, and interactional achievement. *Journal of Pragmatics*, 42(8): 2078–2105.

Arundale, R. B. 2010b. Relating. *Interpersonal Pragmatics*, 6: 137–165.

Arundale, R. B. 2020. *Communicating & Relating: Constituting Face in Everyday Interaction*. New York: Oxford University Press.

Bartneck, C.; Lütge, C.; Wagner, A.; and Welsh, S. 2021. *An Introduction to Ethics in Robotics and AI*. Springer Nature.

Berger, P. L.; and Luckmann, T. 1966. *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. New York: Anchor Books.

Brown, P.; and Levinson, S. C. 1987. *Politeness: Some Universals in Language Usage*, volume 4. Cambridge University Press.

Cabitza, F.; Campagner, A.; Ciucci, D.; and Seveso, A. 2019a. Programmed inefficiencies in DSS-supported human decision making. In *Modeling Decisions for Artificial Intelligence: 16th International Conference, MDAI 2019, Milan, Italy, September 4–6, 2019, Proceedings 16*, 201–212. Springer.

Cabitza, F.; Campagner, A.; Ciucci, D.; and Seveso, A. 2019b. Programmed inefficiencies in DSS-supported human decision making. In *Modeling Decisions for Artificial Intelligence: 16th International Conference, MDAI 2019, Milan, Italy, September 4–6, 2019, Proceedings 16*, 201–212. Springer.

Cave, S.; Dihal, K.; and Dillon, S., eds. 2020. *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*. Oxford, UK: Oxford University Press. ISBN 9780198846666.

Chen, Z.; and Schmidt, R. 2024. Exploring a behavioral model of "positive friction" in human-AI interaction. In *International Conference on Human-Computer Interaction*, 3–22. Springer.

Cooper, A. 1999. *The Inmates are Running the Asylum*. Springer.

Cox, A. L.; Gould, S. J.; Cecchinato, M. E.; Iacovides, I.; and Renfree, I. 2016. Design frictions for mindful interactions: The case for microboundaries. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1389–1397.

Crane, T. 1998. Intentionality as the mark of the mental. *Royal Institute of Philosophy Supplements*, 43: 229–251.

Crilly, N. 2010. The roles that artefacts play: Technical, social and aesthetic functions. *Design Studies*, 31(4): 311–344.

Croce, M. 2018. Expert-oriented abilities vs. novice-oriented abilities: An alternative account of epistemic authority. *Episteme*, 15(4): 476–498.

Davis, J. L. 2020. *How Artifacts Afford: The Power and Politics of Everyday Things*. MIT Press.

de Boer, B. 2024. *Phenomenology and the Philosophy of Technology*. Open Book Publishers.

Dennett, D. C. 1989. *The Intentional Stance*. MIT press.

Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1): 114.

Ehsan, U.; Liao, Q. V.; Muller, M.; Riedl, M. O.; and Weisz, J. D. 2021. Expanding explainability: Towards social transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–19.

Ehsan, U.; Wintersberger, P.; Liao, Q. V.; Watkins, E. A.; Manger, C.; Daumé III, H.; Riener, A.; and Riedl, M. O. 2022. Human-Centered Explainable AI (HCXAI): beyond opening the black-box of AI. In *CHI conference on Human Factors in Computing Systems Extended Abstracts*, 1–7.

Epley, N.; Waytz, A.; and Cacioppo, J. T. 2007. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4): 864.

Facchini, A.; and Termine, A. 2022. Towards a taxonomy for the opacity of AI systems. In *Conference on Philosophy and Theory of Artificial Intelligence*, 73–89. Springer.

Farhat, F. 2023. ChatGPT as a complementary mental health resource: A boon or a bane. *Annals of Biomedical Engineering*, 52(5): 1111–1114.

Ferrario, A.; Facchini, A.; and Termine, A. 2024. Experts or authorities? The strange case of the presumed epistemic superiority of artificial intelligence systems. *Minds and Machines*, 34(3): 30.

Ferrario, A.; and Loi, M. 2022. How explainability contributes to trust in AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1457–1466.

Ferrario, A.; Sedlakova, J.; and Trachsel, M. 2024. The role of humanization and robustness of Large Language Models in conversational artificial intelligence for individuals with depression: A critical analysis. *JMIR Mental Health*, 11: e56569.

Floridi, L. 2023. AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology*, 36(1): 15.

Floridi, L.; and Nobre, A. C. 2024. Anthropomorphising machines and computerising minds: The crosswiring of languages between Artificial Intelligence and Brain & Cognitive Sciences. *Minds and Machines*, 34(1): 1–9.

Fregosi, C.; and Cabitza, F. 2024. A frictional design approach: Towards judicial AI and its possible applications. In *Proceedings of the Workshops at the Third International Conference on Hybrid Human-Artificial Intelligence (HHAI-WS 2024)*. Malmö, Sweden.

Gabriel, I.; Manzini, A.; Keeling, G.; Hendricks, L. A.; Rieser, V.; Iqbal, H.; Tomašev, N.; Ktena, I.; Kenton, Z.; Rodriguez, M.; et al. 2024. The ethics of advanced AI assistants. *arXiv preprint arXiv:2404.16244*.

Goffman, E. 1959. *The Presentation of Self in Everyday Life*. New York, NY: Anchor Books.

Goffman, E. 2017. *Interaction Ritual: Essays in Face-to-Face Behavior*. Routledge.

Griffith, S.; Subramanian, K.; Scholz, J.; Isbell, C. L.; and Thomaz, A. L. 2013. Policy Shaping: Integrating Human Feedback with Reinforcement Learning. In Burges, C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Günther, M.; and Kasirzadeh, A. 2022. Algorithmic and human decision making: For a double standard of transparency. *AI & Society*, 37(1): 375–381.

Harshvardhan, G.; Gourisaria, M. K.; Pandey, M.; and Rautaray, S. S. 2020. A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review*, 38: 100285.

Hawley, K. 2014. Trust, distrust and commitment. *Noûs*, 48(1): 1–20.

Heersmink, R.; de Rooij, B.; Clavel Vázquez, M. J.; and Colombo, M. 2024. A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness. *Ethics and Information Technology*, 26(3): 41.

Hicks, M. T.; Humphries, J.; and Slater, J. 2024. ChatGPT is bullshit. *Ethics and Information Technology*, 26(2): 38.

Homer. 1924. *The Iliad*. Loeb Classical Library. Cambridge, MA: Harvard University Press.

Ihde, D. 1990. *Technology and the Lifeworld: From Garden to Earth*. Indiana University Press.

Jacob, P. 2023. Intentionality. In Zalta, E. N.; and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2023 edition.

Jacovi, A.; Marasović, A.; Miller, T.; and Goldberg, Y. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 624–635.

Jakesch, M.; Hancock, J. T.; and Naaman, M. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11): e2208839120.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Li, M.; and Suh, A. 2021. Machinelike or humanlike? A literature review of anthropomorphism in AI-enabled technology. In *54th Hawaii International Conference on System Sciences (HICSS 2021)*, 4053–4062.

Linton, R. 1936. *The Study of Man: An Introduction.* Appleton-Century.

Manzini, A.; Keeling, G.; Alberts, L.; Vallor, S.; Morris, M. R.; and Gabriel, I. 2024. The code that binds us: Navigating the appropriateness of human-AI assistant relationships. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 943–957.

Montemayor, C.; Halpern, J.; and Fairweather, A. 2022. In principle obstacles for empathic AI: Why we can't replace human empathy in healthcare. *AI & Society*, 37(4): 1353–1359.

Natale, S. 2021. *Deceitful media: Artificial intelligence and social life after the Turing test*. Oxford University Press, USA.

Riess, H. 2017. The science of empathy. *Journal of Patient Experience*, 4(2): 74–77.

Roose, K. 2023. Bing's A.I. chat: 'I want to be alive.'. *The New York Times*. Accessed February 26, 2023.

Sacks, H.; Schegloff, E. A.; and Jefferson, G. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4): 696–735.

Shanahan, M. 2024. Talking about large language models. *Communications of the ACM*, 67(2): 68–79.

Shanahan, M.; McDonell, K.; and Reynolds, L. 2023. Role play with large language models. *Nature*, 1–6.

Shevlin, H. 2024. All too human? Identifying and mitigating ethical risks of Social AI. *Law, Ethics & Technology*, 1(2): 1–22.

Stryker, S.; and Burke, P. J. 2000. The past, present, and future of an identity theory. *Social Psychology Quarterly*, 284–297.

Tajfel, H.; and Turner, J. C. 2001. An integrative theory of intergroup conflict. In Hogg, M. A.; and Abrams, D., eds., *Intergroup Relations: Essential Readings*, 94–109. Psychology Press.

Turkle, S. 2005. *The Second Self: Computers and the Human Spirit*. MIT Press.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 6000–6010.

Vermaas, P. E.; and Dorst, K. 2007. On the conceptual framework of John Gero's FBS-model and the prescriptive aims of design methodology. *Design Studies*, 28(2): 133–157.

Vermaas, P. E.; and Houkes, W. 2003. Ascribing functions to technical artefacts: A challenge to etiological accounts of functions. *The British Journal for the Philosophy of Science*, 54(2): 261–289.

Vermaas, P. E.; and Houkes, W. 2010. *Technical Functions: On the Use and Design of Artefacts*. Springer. ISBN 978-90-481-9171-9.

Wang, J.; Ma, W.; Sun, P.; Zhang, M.; and Nie, J.-Y. 2024. Understanding user experience in large language model interactions. *arXiv preprint arXiv:2401.08329*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229.

Weizenbaum, J. 1966. ELIZA – A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1): 36–45.

Weizenbaum, J. 1967. Contextual understanding by computers. *Communications of the ACM*, 10(8): 474–480.

Yablo, S. 2014. *Aboutness*, volume 3. Princeton University Press.

Young, L. 2019. 'I'm a cloud of infinitesimal data computation' When machines talk back: An interview with Deborah Harrison, one of the personality designers of Microsoft's Cortana AI. *Architectural Design*, 89(1): 112–117.

Zerilli, J.; Knott, A.; Maclaurin, J.; and Gavaghan, C. 2019. Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology*, 32: 661–683.

Zhan, X.; Xu, Y.; and Sarkadi, S. 2023. Deceptive AI ecosystems: The case of ChatGPT. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, 1–6.