universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Faking It Right: Human-Likeness in Pragmatic Authenticity

verfasst von | submitted by

Barbara Geld univ. bacc. psych.

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 013
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Joint-Degree Middle European interdisc. master prog. in Cognitive Science
Betreut von | Supervisor:	Dr. Astrid Weiss

Za dedu Stjepana.

Abstract

This thesis introduces Pragmatic Authenticity, a framework for designing and evaluating interactions between humans and artificial agents, emphasising subjective experience over inherent agent traits. It advocates moving away from ontological authenticity, which replicates human qualities like consciousness, due to ethical and practical limitations. The framework encompasses four factors: trustworthiness (institutional and individual), intelligence, and human-likeness. Trustworthiness involves constructs such as privacy, transparency, and ethics, while intelligence focuses on both the agent's measurable capabilities and complexity, and user perceptions. Human-likeness, shaped by anthropomorphism, social presence, and perceived empathy, examines the effect of an agent's ability to resemble human traits, on human perception, and its relation to overall authenticity perception. An empirical study tested the relationship between human-likeness and authenticity using ChatGPT-4o in advice-seeking scenarios. Participants' perceptions were assessed through established tools, including the Godspeed and Networked Minds Questionnaires, and a novel authenticity scale. Results showed positive correlations between human-likeness facets and perceived authenticity, supporting their unique contributions to agent design. Furthermore, the study confirmed no significant multicollinearity among these facets, emphasising the independent role of each in shaping human-likeness. The study also highlighted the influence of interaction experience over preconceived attitudes, underscoring the importance of user-centred evaluation.

Abstrakt

Diese Masterarbeit stellt Pragmatische Authentizität vor, ein Framework zur Gestaltung und Bewertung von Interaktionen zwischen Menschen und künstlichen Agenten, das subjektive Erfahrungen gegenüber inhärenten Eigenschaften von Agenten hervorhebt. Es plädiert für eine Abkehr von der ontologischen Authentizität, die menschliche Eigenschaften wie Bewusstsein repliziert, aufgrund ethischer und praktischer Einschränkungen. Das Framework umfasst vier Faktoren: Vertrauenswürdigkeit (institutionell und individuell), Intelligenz und Menschenähnlichkeit. Vertrauenswürdigkeit beinhaltet Konstrukte wie Datenschutz, Transparenz und Ethik, während Intelligenz sich sowohl auf die messbaren Fähigkeiten und Komplexität des Agenten als auch auf die Wahrnehmung der Nutzer konzentriert. Menschenähnlichkeit, geprägt durch Anthropomorphismus, soziale Präsenz und wahrgenommene Empathie, untersucht die Auswirkung der Fähigkeit eines Agenten, menschliche Eigenschaften nachzuahmen, auf die menschliche Wahrnehmung und deren Beziehung zur umfassenden Wahrnehmung ihrer Authentizität. Eine empirische Studie testete die Beziehung zwischen Menschenähnlichkeit und Authentizität unter Verwendung von ChatGPT-4o in Beratungsszenarien. Die Wahrnehmungen der Teilnehmer wurden durch etablierte Instrumente wie die Godspeed- und Networked Minds-Fragebögen sowie eine neue Authentizitätsskala bewertet. Die Ergebnisse zeigen positive Korrelationen zwischen Aspekten der Menschenähnlichkeit und wahrgenommener Authentizität, was deren einzigartige Beiträge zum Agentendesign unterstützt. Darüber hinaus bestätigte die Studie keine signifikante Multikollinearität zwischen diesen Aspekten, was die unabhängige Rolle jedes einzelnen bei der Gestaltung der Menschenähnlichkeit betont. Die Studie hob auch den Einfluss der Interaktionserfahrung gegenüber vorgefassten Einstellungen hervor, was die Bedeutung der nutzerzentrierten Bewertung unterstreicht.

Table of Contents

1. Introduction

Artificial agents have captured our imagination for decades, appearing in stories, movies, and increasingly in real life. Especially in science fiction, they are often presented not just as tools designed to serve us, but as companions, partners, and even friends. Consider Doraemon, the robotic cat from the future, not only a problem solver, but also a confidant to Nobita, their accident-prone human friend. Or Baymax, the soft, inflatable healthcare robot from *Big Hero 6*, which, quite literally, embodies warmth and care, acting as a committed protector and healer. TARS, the adaptive robot from *Interstellar*, may be clearly mechanical in form, but its dry humour and unwavering loyalty forge a bond with the crew that feels deeply human. These agents, fictional or otherwise, are more than the sum of their mechanical or coded parts. They inspire feelings of trust, gratitude, and even affection. As machines become companions, it is apparent that the lines blur between utility and partnership. What begins as simple functionality evolves into something more layered. As explored by [43], we don't necessarily need Baymax to genuinely care about our well-being, or TARS to actually grasp the nuances of loyalty and humour. What matters might be that they act in ways that make us feel understood, supported, and connected. The interplay between these perceptions and the qualities of the agents themselves forms the heart of the exploration into *pragmatic authenticity*.

1.1 About Authenticity

Pragmatic authenticity focuses on the perceived authenticity of an artificially intelligent agent based on its behaviour and interaction with users. The pragmatic approach accepts that AI may not possess genuine emotions or consciousness, but assesses whether its behavioural responses are appropriate and believable within the context of the interaction [104]. In contrast, what is labelled *ontological authenticity* concerns the "realness" of an AI's emotions or intelligence, as suggested by [18, 61] in their discussions rooted in phenomenology, focused on deception, and popularly advocated for by [23]. It questions whether AI can genuinely possess the qualities it simulates, and often leads to discussions about the distinction between "reality and appearance" in AI [18, 61]. There are different perspectives on how to achieve authenticity in AI, with some focusing on functional approaches and others aiming for ontological replication. The contextual black box approach, aligned with pragmatic authenticity, suggests that an AI's inner workings are less important than its observable behaviour. If the AI's responses are appropriate and

consistent within the specific context, its authenticity can be accepted without needing to understand, or relate to, its internal processes [128]. However, this raises questions about whether such a system can only showcase what its designers intend, potentially limiting the scope of its perceived authenticity [99]. Second point being raised in line with authenticity is that transparency and explainability are crucial for achieving authenticity, particularly in the realm of trustworthy AI [99, 61]. This view suggests that users need to understand how an AI reaches its conclusions and what principles guide its behaviour to be perceived as authentic. This aligns with the idea that a transparent purpose and the ability to explain its decision-making process contribute to an AI's value of perceived authenticity [99]. Authors also suggest that mimicking human-like behaviour, including emotional responses and social cues, is essential for establishing authenticity, some of which would involve incorporating anthropomorphic design elements [99] and programming agents to exhibit seemingly socially intelligent behaviour [18, 61]. These approaches, however, raise the question of whether humanness is the desired standard for authenticity within AI, and whether this reflects an egocentric bias [61]. Lastly, the idea of an artificial agent as having an "inner self" or "true character" is central to discussions about ontological authenticity [18], which is central to the question of whether AI can possess genuine intentions, and even more complex, philosophical questions about the nature of human consciousness and whether it can be replicated in artificial systems [18, 23]. Discussing such ideas leads to the two perspectives offering polarly opposite opinions: Constructivism and Computationalism. Constructivism emphasises the role of experience and social interaction in shaping intelligence, suggesting that replicating human-like qualities might require AI to undergo similar developmental processes, as suggested by the work of [18, 99]. Computationalism, on the other hand, states that the mind can be understood as a computational system, implying that replicating its functions in AI might be achievable through sufficiently complex algorithms, as can be inferred by the work done by [23]. However, even if we did manage to be able to integrate consciousness and other cognitive traits in agents, there is a question of: *Why?* and *Should we?* As already suggested, the use of human-like qualities as a benchmark for AI's authenticity potentially reflects an anthropocentric bias [18] and urges discussions about whether alternative forms of intelligence or authenticity might exist to navigate, or moderate, the human-agent interaction.

1.2 Research Motivation

Firstly, choosing to explore pragmatic authenticity over ontological authenticity was a practical motivation, grounded in technical and ethical reasons. Rather than getting caught up in philosophical debates about true agency or consciousness in machines, a pragmatic approach focuses on measurable aspects of authenticity in interactions. Although useful, the current focus on ontological authenticity would be doing a disservice to optimising human-agent interaction (HAI), with the (limited) tools at our disposal. The pragmatic authenticity approach is more attainable and relevant in the current technological landscape, enabling the development of design principles that can be empirically tested and refined, aiming to design systems that are perceived as more genuine and trustworthy by human users. [Bickmore] define "relational agents" as "computational artifacts designed to build long-term, social-emotional relationships with their users", aligning with the qualities emerging from pragmatism-based interactions. The authors acknowledge that the notion of relationship itself is multifaceted and lacks an agreed-upon definition, which reinforces the previously stated need for a framework focused on measurable aspects of authenticity, as opposed to delving into the philosophical complexities of *true* agency or consciousness, which are still debated. Secondly, despite the increasing social nature of human-technology interactions and the rapid growth of AI and its applications, design guidelines seem to be dispersed across academic communities and confined to separate disciplines. This lack of a clear framework is hindering the progress relating to authentic interactions, especially knowing how successful, accepted and beneficial the standards and guidelines, within human-computer interaction (HCI) and human-robot-interaction (HRI) communities, can be. [63] highlights this lack of a unified set of design guidelines for HAI, despite the increasing presence of AI across contexts, and states how it is not uncommon that researchers identify usability issues without explicitly noting the proposed design principles to improve these issues in design, which further reinforces the challenges for designers. This supports the motivation for developing a guideline-like framework for designing and evaluating authentic interactions. Furthermore, researchers have expressed how different fields contribute to the concept of pragmatic authenticity, with [81] underscoring the importance of intersubjectivity in human-agent interaction, highlighting the need to evoke "natural reactions" from users, grounding their work in psychology and sociology of human-human interactions. They advocate for conceptualising agent design in "behavioural realism", as they believe that the human perception of this "naturalness" is

what gives rise to intuitive human-agent interactions. [104] discusses the ethical implications of robotic deception, acknowledging arguments for and against its use. The discussion draws on philosophy, ethics, and social psychology to explore the complexities of deception and its potential impact on human-robot relationships. [108] emphasises the importance of trust in human-robot interaction and the potential for deception to negatively affect it. They also highlight the role of human designers in shaping robot behaviour, and therefore deception, emphasising the responsibility which falls upon humans if, or when, using robots as "tools of deception". This work builds upon ethics, HRI, and social psychology. From a perspective of a genuinely human-centred, safe, responsible AI design, [16] underscores the potential of international human rights standards as a basis for ethical AI development, which would ensure legal and ethical implementation beyond the regulations assigned by technology companies or the profiting industry. This corpus alone provides compelling evidence for the need for a unified framework for designing authentic interactions, focusing on pragmatism, intuition, and safety in interaction. Mentioned research demonstrates how various fields are necessary for their unique contributions to the implementation of what will be defined as factors and facets of pragmatic authenticity in human-agent interaction.

As for the empirical aspect of this thesis, one of the proposed factors from the pragmatic authenticity framework, human-likeness, will be validated. Firstly, this could serve as empirical validation for human-likeness as a measure of pragmatic authenticity, if the agent perceived as human-like, gives rise to an authentic interaction. Secondly, this study could serve as a validation measure for the design principles of pragmatic authenticity, in case a specific measurable aspect (anthropomorphism, likeability, social presence) leads to the perception of the agent being perceived as more authentic.

1.3 Concepts and Terminology

The key concepts which this thesis was built upon have been defined below, in order in which they will be brought up and discussed in the following sections.

Pragmatic Authenticity: Unlike ontological authenticity, which strives to replicate human qualities in AI, pragmatic authenticity focuses on the user's subjective experience. It aims to make AI interactions feel genuine, intuitive, and natural for the user, regardless of whether the AI possesses human-like consciousness or not. It is a newly proposed concept, however it has been previously discussed by [61]. According to discussions and

definitions from philosophy, offered in [87] and [34], this could also be equated to their "Iconic authenticity", which describes whether an (a human perception of an) object is aligned with the human's *expectations* of how said object should *appear* (example given are uniforms used in reenactments of a war battle, which might be perceived as authentic based on the information people have about uniforms and clothes from that era). This is, as [87] reports, also labelled as "authentic reproduction".

Ontological Authenticity: This concept aims to create AI that genuinely replicates human qualities like consciousness, empathy, and intelligence. However, it raises ethical and philosophical concerns about the feasibility and desirability of such replication. As previously discussed by [18]. According to philosophical categorisations proposed in [87] and [34], this could also be equated to the "Indexical authenticity", as it concerns whether something is an original or a replica, in the sense of physical originality of an object at hand (example given was a *Picasso* painting, and determining whether it was indeed painted by the renowned artist, Pablo Picasso).

Entanglement Theory: A paradigm shift in human-computer interaction that recognises the increasingly intertwined relationship between humans and technology. It challenges traditional design approaches that view technology merely as a tool and instead advocates for focusing on the co-shaping between humans and agents and emphasises the role of ethics in technology. As proposed by [52].

4E Cognition (Enaction): A framework in cognitive science that proposes cognition is not confined to the brain but is embodied, embedded, enacted, and extended through interactions with the body, environment, and tools. This framework challenges traditional, brain-centric views of cognition. As explored by [20].

Responsible AI (RAI): An approach to AI development and deployment that prioritises ethical considerations, transparency, accountability, and fairness, ensuring that AI systems are aligned with human values and societal well-being. As discussed by [103].

Explainable AI (XAI): A subfield of AI focused on making systems' decision-making processes understandable to humans, promoting transparency and trust. As extensively discussed in [57].

Human-Likeness (HL): The degree to which an artificial agent exhibits characteristics, behaviours, or qualities that are perceived as human-like. This concept encompasses various dimensions, including perceived traits of anthropomorphism, and social presence. [93, 75]

Social Presence (SP): "The degree of initial awareness, allocated attention", and for mutual understanding and behavioural interdependence between two interlocutors, in an interaction [78]. In AI, presence can be enhanced through design choices that promote a sense of interaction, engagement, and emotional responsiveness.

Anthropomorphism (ANT): The attribution of human-like characteristics or behaviours to non-human entities, including artificial intelligence agents [93]. Anthropomorphism can be used to enhance user engagement, but excessive or misleading anthropomorphism can create unwanted or unsettling effects [42].

Empathy: As defined by [51] in 1975, from a standpoint of a clinician, empathy "means entering the private perceptual world of the other and becoming thoroughly at home in it. (...) moving about in it delicately, without making judgments, sensing meanings of which he/she is scarcely aware (...). It includes communicating your sensings of his/her world as you look with fresh and unfrightened eyes at elements of which the individual is fearful." I leave you with this definition to express just how human, personal and feeling-based, genuine empathy is. On the other hand, [113] offer a *contemporary* definition of empathy, defining its affective and cognitive components, where the *affective* approach "defines empathy as an observer's emotional response to the affective state of another", comparing it to sympathy. In their work, *cognitive* empathy can be seen compared to using the theory of mind, or "setting aside one's own current perspective, attributing a mental state to the other person, and then inferring the likely content of their mental state."

Perceived Empathy: An agent's ability to understand and respond to human emotions in a way that is *perceived* as caring, supportive, and authentic. *Empathy* in AI can be conveyed through language, tone of voice, and responsiveness to emotional cues, as modeled by [3], and this perception of *artificial* empathy emerges through interaction, as discussed by [43].

Uncanny Valley: A phenomenon in which an artificial agent's near-human appearance or behaviour evokes feelings of unease or discomfort in humans. The uncanny valley suggests there is a point where increasing human-likeness can lead to negative rather than positive responses. As defined by [121].

Individual Trustworthiness: A key component of the pragmatic authenticity framework, encompassing an AI agent's reliability, moral integrity, and accountability. It focuses on whether users perceive the agent as acting in a dependable, ethical, and transparent manner. As previously discussed by [61] in terms of general trustworthiness.

Intelligence: In the context of the pragmatic authenticity framework, intelligence refers to both the measurable capabilities of an artificial agent (objective intelligence) and how users perceive those capabilities (perceived intelligence). Perceived intelligence can be influenced by factors such as an agent's perceived agency, autonomy, and theory of mind. Objective intelligence defined as the complexity of the intelligent agents is a general definition, adopted also by [Russell & Norvig]

Theory of Mind (ToM): The ability to attribute mental states, such as beliefs, desires, and intentions, to oneself and others. In AI, ToM refers to an agent's ability to understand and predict human behaviour based on an understanding of mental states. As previously discussed by [99].

Interaction-Centred Design: A design approach that prioritises the quality and authenticity of the interaction between humans and intelligent agents. This approach focuses on creating engaging, meaningful, and trustworthy interactions, moving beyond superficial human-likeness and prioritising user experience and ethical considerations. As highlighted by [65]

1.4 Overview of Thesis Structure

This first section has set the stage by defining authenticity and introducing key concepts and terminology relevant to human-agent interaction. It also establishes the motivation for this research, emphasising the need for a unified framework to guide the design of authentic HAI.

The second section explores existing theoretical frameworks that inform the understanding of HAI, such as human-robot interaction, socially interactive agents, and interaction design theory. It also introduces entanglement theory as a paradigm shift in human-computer interaction, highlighting its relevance to understanding the intertwined relationships between humans and technology.

The third section examines the existing literature on responsible AI, social and cognitive robotics, and human-likeness in design. It explores concepts such as trustworthiness, explainability, agency, social presence, and anthropomorphism, highlighting their significance in shaping user perceptions of artificial social agents. This section was introduced as a separate unit to the previous one, as the second section explores a broad context for developing the framework of pragmatic authenticity, as well as the empirical

study, while the third section targets more specific concepts relevant for the understanding of the novel framework.

The fourth section presents the core theoretical framework of the thesis, explaining the rationale for prioritising pragmatic authenticity over ontological authenticity. It defines the three factors of pragmatic authenticity: individual trustworthiness, intelligence, and human-likeness. Each factor is further broken down into facets, providing a comprehensive model for understanding and evaluating interactions.

The fifth section describes the empirical study conducted to investigate the relationship between human-likeness and pragmatic authenticity. The study examines how social presence, anthropomorphism, and perceived empathy influence user perceptions of authenticity, also exploring the role of expertise and attitudes toward AI as potential moderating factors. It introduces a novel semantic differential scale to measure the authenticity of artificial agents and evaluates its reliability.

Sections six and seven describe the methodology of the empirical study, including data collection procedures, participant demographics, and the statistical analyses employed.

Section eight discusses the study's findings in relation to the research questions and hypotheses, exploring the implications of the results for the design and development of artificial agents with which an authentic interaction may emerge. It also discusses limitations of the study, integrating criteria of rigour and ethical reflection in evaluating the research process.

Section nine summarises the key findings of the thesis, highlighting the contributions of the pragmatic authenticity framework and the empirical study. It reiterates the limitations of the research, identifies areas for future work, and emphasises the importance of interaction-centred design in creating authentic and engaging human-agent interactions.

The remaining sections (References and Appendices) provide supplementary information, including citations for all sources used in the thesis and additional materials related to the empirical study.

This introductory section establishes the main focus of the thesis, which is the exploration of the concept of pragmatic authenticity in the context of human-agent interaction. It distinguishes pragmatic (functional) authenticity, centred on human perception, from ontological (theoretical) authenticity, which discusses whether AI can possess genuine human-like qualities, and the ways they might be implemented. The practical and ethical reasons for prioritising pragmatic authenticity were also explained, arguing that a

pragmatic approach might be a more relevant undertaking regarding the impact on humans, as well as it being a possibly more attainable goal. The motivation for this research was also highlighted, stemming from both rapid growth in technological developments, and human intertwinedness with technology, necessitating guidelines for trustworthy, comfortable, and authentic interaction with artificial companions that (will) surround us. Finally, the section outlines the structure of the thesis, indicating the following work will both propose a novel theoretical framework, as well as present findings of an empirical study which followed one aspect of said framework. The subsequent two sections introduce the existing frameworks and theories, which will be useful for contextualising the novel framework and methodologies used in the study, and also a state-of-the-art literature review is presented to present some specific concepts which are relevant for the development of the framework.

2. Established Frameworks and Theories

Having established the topic of interest as *pragmatic authenticity* and explored its distinction from ontological authenticity in the introductory section, we now turn to seminal frameworks and theories within which the novel framework of pragmatic authenticity was developed. This section first follows the historical evolution of human-robot interaction as a field, highlighting its arc from task-oriented machines to socially embedded agents. Introduced in this section are also the standards which guided the approach to research and design practices, in the fields of human-computer, human-robot, and human-agent interaction. The discussion will then explore the emergence of human-agent interaction, emphasising its focus on shared agency and collaboration, introducing the concept of different agent embodiments. Next, socially interactive disembodied agents will be in focus, specifically the challenges and opportunities presented by their disembodiment. Explored will also be ethical considerations surrounding their design. Lastly, this section introduces the interaction design theory, and its contribution to interdisciplinary research and practical developments of interactive systems. However, limitations are discussed, and an even more suitable approach is proposed within the entanglement theory, and its intertwinement with the 4E cognition framework.

2.1 Human-Robot Interaction

The history of human-robot interaction (HRI) shows a long-standing interest in how artificial beings could interact with humans. Myths, stories, and philosophical ideas from

many cultures, including ancient Greece and China, explored the idea of artificial beings with human-like qualities. A prominent example of this fascination is Leonardo da Vinci's 1495 sketch of a mechanical man, which blends imagination with mechanical ideas [69]. For HRI to become a scientific field of study, however, important technological advances had to take place. One early step was the creation of teleoperated robots, like the "Electric Dog" developed at the Naval Research Laboratory in 1923, which allowed humans to control machines remotely [69]. Another key development was the improvement of remotely piloted vehicles, showing that humans could guide machines over distances, and in the 1960s, the development of "Shakey", referred to as "one of the first modern robots", marked a major milestone. Shakey was able to navigate and interact with its environment without constant human direction, demonstrating the possibility of robot autonomy [69]. By the 1980s, robots were designed to respond more effectively to their surroundings, making interactions with humans more practical and dynamic, and later, reactive systems were combined with higher-level reasoning, allowing for a more *natural* interaction. HRI became its own distinct field in the mid-1990s and early 2000s, through technological progress combined with the efforts of researchers from different disciplines. There was an obvious need for interdisciplinary work, as addressing the challenges of human-robot interaction required input from areas such as robotics, cognitive science, and psychology. The combined expertise allowed researchers to tackle the many aspects of HRI, from the technical to the behavioural. Specific fields like search and rescue, assistive robotics, and space exploration contributed to the development of HRI as they presented real-world challenges that demanded innovative solutions. Robots designed for these environments had to work effectively with humans, whether navigating difficult physical spaces or communicating in high-stakes situations. These real-life contexts pushed the development of new tools and approaches, shaping how robots could interact with people in meaningful ways [69].

Behind the scenes of all the research and developments, an important practical aspect was the evolution of the International Organization for Standardization (ISO) standards, which reflected a progression from a process-focused approach to a more holistic perspective that encompasses the entire design cycle. Published in 1998, the ISO 9241-11 provided a specific definition of usability, emphasising effectiveness, efficiency, and satisfaction within a defined context of use [119]. This standard set the groundwork for assessing and measuring the usability of interactive systems, establishing benchmarks for

evaluating the quality of user interactions. These principles formed the basis for a shared understanding of usability, primarily benefiting the designers, allowing them to create systems that align with user expectations. Building on the practical ideas of the previous standard, the ISO 13407, introduced in 1999, expanded its principles to a human-centred understanding of design for interactive systems. This standard emphasised user involvement, iterative design, and multidisciplinary collaboration, offering guidelines on incorporating usability principles into design processes [119]. It highlighted the necessity of including, and actively engaging users throughout the design stages to refine the final products. Moreover, the inclusion of diverse perspectives throughout the process ensured that the multifaceted nature of HCI was considered thoroughly and holistically [119].

However, while ISO 9241-11 provided a detailed usability definition, and ISO 13407 offered an expansion for the design processes, ISO 9241-210 incorporated these elements into a broader framework that considered the context of use, user requirements, and continuous evaluation throughout a product's *life cycle* [84]. A distinctive feature of ISO 9241-210 is its emphasis on producing design solutions that draw upon HCI knowledge and subsequently rigorously evaluating these solutions to be in compliance with usability requirements. This iterative process of designing, testing and refining is foundational in developing final products that are hallmarked by usability.

Despite its comprehensive approach, ISO 9241-210 has not gone without critique. Some argue that its anthropocentric focus is a limitation, particularly in the context of the current climate crisis, explaining that prioritising human needs can lead to designs that negatively impact the environment and other stakeholders [115], suggesting that a shift toward eco-centric might be necessary. Moreover, while ISO 9241-210 acknowledges sustainability, it has been criticised for emphasising the positive aspects of economic and social dimensions while providing less attention to environmental considerations [115]. And while above mentioned critiques raise important points about the potential need for a different approach to design, I believe it to be worth noting that anthropocentric design does not inherently (need to) exclude considerations of environmental and societal impact. A more nuanced approach might involve integrating environmental sustainability into the core of human-centred design, ensuring that designing for humans inevitably includes addressing the broader consequences of human activities, also put into context by [115].

2.2. Human-Agent Interaction

The field of human-agent interaction arose from the convergence of human-computer interaction (HCI) and automation. Initially, the focus of automation was simply to replace human tasks with machines to improve efficiency, speed, and reduce costs [70]. Early goals were defining the strengths and weaknesses of both humans and machines in order to define appropriate tasks for humans and machines, respectively [70]. However, the increasing complexity of systems and the development of more sophisticated agents led to a realisation that interactions between humans and autonomous agents required a different approach, because the straightforward replacement of human tasks with machines is not a possibility. The new understanding was that effective automation means changing the nature of the task itself, which required new ways of thinking about how tasks can be shared and coordinated between humans and machines [41]. This was supported by the concept of "man-computer symbiosis" [90], where human capabilities are augmented by computational systems. Following the new paradigm, [9] proposed an "un-Fitts list" highlighting how human and machine capabilities can be mutually enhanced through interaction and collaboration, presenting, of course, an opposition to the original list, describing tasks allocated to each and the other, alone.

This realisation that interdependence, rather than simple task replacement, is central to effective HAI led to the development of the field as a distinct area of research. The focus of HAI is on designing agents that can work collaboratively with humans, understand human goals and contribute to the achievement of shared ones, and adapt to changing situations [70]. While some of these goals can be found as benchmarks in HRI research, and sometimes, HCI, the focus of HAI is indeed distinct, as it requires moving beyond traditional HCI by prioritizing shared agency, coordination, mutual understanding, and beyond HRI, by incorporating agents of various embodiments. The proposed categorisation of embodiments divides agents into physical, virtual or disembodied [50], with the physical agents being co-situated in the user's environment, virtual ones being an instigator, drawing the user into the agent's (virtual) environment, and disembodied agents having no physical design features resembling an artificial agent, but only operating on an interface [50].

The concept of *physical* embodiment, as "physical instantiation of an agent in its environment" [50], became relevant in HAI as researchers began exploring the impact of an agent's form on its interactions with humans. This question stemmed from the

observation that, while embodiment is crucial for robots that physically manipulate their environment, many socially interactive agents don't require physical interaction to fulfil their tasks [27]. Relevant to the conversation is also the 'embodiment hypothesis', which states that physical embodiment measurably affects the performance and perception of social interactions [117], further highlighting the importance of physical embodiment, and empirically backed by a plethora of research [46, 22]. However, there are physically disembodied agents, such as text interfaces, animations, or high-fidelity virtual characters [50], which are becoming all the more present today, in the daily lives of humans, in the form of voice user interfaces, whether in smart assistants in homes, in-car systems, or within the widespread use of ChatGPT. While not necessitating focus on physical design to evoke acceptance, disembodied agents have a highlighted feature of *social* embodiment. While [107] and [50] offer a definition of social embodiment that is tied to the "states of the body" and its movement, social embodiment may also be achieved through embedding social design cues modified and tailored for disembodied agents, such as engagement, tonality, expression, and more.

2.3 (Socially Interactive) Disembodied Agents

The development of robots and agents capable of meaningful social interactions has been a central focus for researchers such as [124], [26, 60, 109] and [10] for over two decades. Their work aims at creating embodied agents that could effectively integrate into human social environments. However, the emergence and exponential growth of disembodied agents, particularly conversational agents and voice assistants, has significantly impacted the trajectory of human-agent *integration*. This rapid rise of disembodied agents can be attributed to two significant factors. Firstly, advances in artificial intelligence and natural language processing have enabled conversational agents like *ChatGPT* to engage in interactions that closely mimic human conversation. Secondly, technological advancements have made disembodied agents particularly accessible and readily available [71]. Voice assistants such as Alexa and Siri have become integral to many people's daily routines, underscoring how these technologies have seamlessly (or at least *rapidly*) integrated into modern life.

However, designing disembodied agents presents a unique set of challenges distinct from those associated with embodied systems. While ethical considerations such as transparency, reliability, and explainability are significant hurdles for developers in both

scenarios [63, 71], the design of disembodied agents should work towards addressing the absence of cues commonly present in physical interactions and communication. Without the ability to rely on gestures, facial expressions, or body language, the role of language [63] in interactive disembodied agents is greatly highlighted.

As these agents get better equipped with the exhibition and recognition of social cues, questions about their evolving social and emotional roles arise. Research has demonstrated that robots and agents can evoke social and emotional responses in humans, resembling the reactions commonly elicited in human-human interactions [66]. Hence a potential bidirectional effect to follow, in examining if the way individuals engage with voice assistants could have a role in shaping the expectations for human interactions. Aside from the social, equally critical are the ethical implications surrounding the integration of disembodied agents, in ensuring that these systems respect human values and autonomy. These design hurdles will involve addressing concerns about bias, manipulation, and potential privacy violations [63]. The ability of these agents to tailor their responses to the user preference, or engage with users in ways that might exploit vulnerabilities, could pose risks that demand careful oversight from their developers.

2.4 Interaction Design Theory and Entanglement Theory

Interaction Design Theory is a widely used framework for researching and developing effective interactive systems, with its primary foci being usability and user engagement. The theory is underpinned by several paradigms, such as cognitive psychology, ergonomics and human factors, and social sciences. The cognitive psychology paradigm provides insights into how humans perceive and process information. Concepts such as information processing theory, mental models, and situated action help designers predict how users will approach an interface and interact with its elements. By designing interfaces that align with users' mental models, designers can reduce cognitive load and improve usability. Similarly, theories of situated action, which emphasise the context-dependent nature of human behaviour, remind designers that perceptions of interactions are often shaped by the user's immediate environment and specific tasks [77]. Ergonomics and human factors add another layer to interaction design by emphasising the physical and cognitive limitations of users, underscoring the importance of comfortable and accessible systems. In line with that, concepts such as Fitts's Law, which *connects the time required to move to a target area, to the distance and size of the target*,

play a significant role in designing interactive elements. Applying such principles not only contributes to accessibility, but also works towards reducing the likelihood of user errors, minimising frustration [77]. Complementing the views of the situated action theories, the social sciences expand the scope of interaction design by acknowledging the broader social and cultural contexts in which technology is used, proposed in the Activity Theory. This perspective encourages designers to consider how their systems will be perceived and utilised within contexts, rather than a *research vacuum* [77, 15].

Entanglement Theory represents a significant paradigm shift in HCI, reconsidering the relationship between humans and technology, in light of their increasing interconnectedness.

While interaction design theory remains a widely employed model for designing artificial artifacts, the conceptual framework of entanglement theory offers a more suitable foundation for addressing the complexities of contemporary, as well as future technological systems, particularly regarding their embeddedness in social contexts, as explored through the framework of pragmatic authenticity.

The emergence of entanglement theory, as articulated by [52], stems from the issues that had been unaddressed by prior theories, such as ontological uncertainties that have become apparent due to developments in fields such as virtual reality and artificial intelligence, as well as ethical challenges which seem to have an ever-growing impact on our lives, both as individuals and as a society. Entanglement theory adopts a perspective based on the Actor-Network Theory, which posits that the social and the technical (*human and non-human*) are inseparably intertwined within "socio-material networks" [52]. Vital to entanglement theory is a focus on post-phenomenology, shifting the design focus from mere interactions to *relationships*, recognising the ways in which technology *shapes*, and *is* shaped, by human engagement with the world. Object-Oriented Ontology (OOO) [115, 52] contributes to this framework by advocating for a flat ontology that decentres human perspectives, wanting to balance out the agency of humans with that of the material world, going so far to emphasise the "autonomy of objects", highlighting that objects should be considered entities regardless of the interactions they are, or are not in. Lastly, agential realism complements the *entangled* perspective (even though not fully aligning with OOO) by proposing a performative ontology in which entities emerge through intra-actions rather than existing as pre-defined representations. This emphasis on emergence

puts in focus an interactive aspect which resonates with the entanglement theory and reinforces the dynamic nature of human-technology relations.

The role of entanglement theory in relation to pragmatic authenticity is further highlighted by its alignment with Enaction (i.e. 4E Cognition), a paradigm in cognitive science that challenges traditional, brain-centric views of cognition. The four Es, embodied, embedded, enacted, and extended, provide a comprehensive model for understanding cognition as an active, situated, and distributed process. Embodiment emphasises that cognition is deeply tied to the physical body and its interactions with the environment, while embeddedness situates cognition within specific contexts that cannot be isolated. Enactment highlights the role of action and interaction in the emergence of cognitive processes, and extension expands the boundaries of cognition to include tools, artifacts, and social systems [20]. Entanglement theory and 4E Cognition both recognise the distributed nature of agency, where human and non-human actors contribute to cognitive processes. Moreover, they share an understanding that knowledge is enacted through ongoing interactions, suggesting that design approaches should prioritise adaptability and learning in human-agent systems.

3. Literature Review

After establishing the broad landscape of frameworks and a theoretical foundation for pragmatic authenticity in section 2, it becomes crucial to specifically discuss the fields and constructs directly related to the key factors of the pragmatic authenticity framework. The pragmatic authenticity framework encompasses four main factors: institutional and individual trustworthiness, intelligence, and human-likeness, and this literature review will explore each of these factors in relation to relevant areas of research, and other related constructs. This section is structured to address each factor in relation to a designated area of research, meaning each subsection will be exploring the relationships between a factor and its related field of study, and arising findings. The subsection on Responsible AI will explore both institutional and individual trustworthiness, by examining how ethical principles, transparency, and accountability are embedded into artificial systems, and how they might affect human perception of trust. The intelligence factor will be discussed in the subsection on Social and Cognitive Robotics, drawing on concepts like the Theory of Mind and social intelligence, and the human-likeness factor will be examined within the

last subsection on design, which considers the impact of anthropomorphism, social presence, and social design cues on user perception. By exploring each of these fields, this section aims to contextualise the factors of the pragmatic authenticity framework and provide the understanding for their relevance within it.

3.1 Responsible AI: FATE and Trust

Responsible AI is rooted in prioritising human well-being and societal values, by aiming to create systems that are both understandable and beneficial to users and society [103]. It also incorporates principles of the human-centred approach [103], crucial for fostering user confidence and ensuring that agents are operating within ethical guidelines. Trustworthiness is essential to AI acceptance, requiring the system to perform efficiently and accurately [25]. This involves the system consistently operating dependably, guided (or rather *constrained*) by rigorous design and deployment strategies that aim to minimise harm [44, 103]. Moreover, explainability is vital to portray the decision-making process happening within the agent, allowing for humans to validate system actions [57], while transparency is considered when an agent in itself is understandable to users [57]. Ethical considerations, including fairness, accountability, privacy, security, and sustainability, remain at the forefront of RAI [103]. The intended purpose of these principles is to protect user privacy, secure data [82, 57], and, when grand societal impact is considered, promote environmentally responsible AI development practices [103]. Numerous authors have proposed guidelines to ensure safe, (responsible, explainable) use of AI in practical fields, such as "The Ten Commandments of Ethical Medical AI," offering sector-specific guidelines for healthcare applications, outlined by [120]. The European Commission has also worked towards establishing guidelines in support of fundamental rights and ethical values, when integrating AI into various fields. The EC's guidelines advocate for transparency, risk management, and sustainability, reflecting a broader commitment to ethical standards across all AI applications [103].

Trustworthiness is a concept in AI which is integral to RAI and directly affects user acceptance and engagement with the technology [25]. Key factors contributing to AI trustworthiness include expected competence, performance, and consistency [103], which ensure that the system reliably performs its tasks with minimal errors. Transparency and explainability support trustworthiness by offering insight into how AI makes decisions, providing users with an understanding of the data, processes, and biases that influence

agent's decision-making [56]. Fairness and justice are essential for ensuring equity when handling data, controlling for biases that could lead to discriminatory outcomes [57], with privacy and security employed through robust data protection, privacy awareness from the developers' side, and development of technologies that could be used for privacy enhancement [57].

Regarding *Fairness, Accountability, Transparency, and Ethics* or *FATE*, the two principles being highlighted in fostering trust within HAI and HRI, are transparency and explainability. Transparency is essential for trust-building in AI, as it gives users a clear view of decision-making processes. Transparency also aids bias mitigation by enabling users to examine AI decision-making, promoting fairness and reinforcing the technology's ethical integrity [57]. In HRI, transparency can be enhanced through nonverbal cues like a robot's gaze, which signals intent, and user interaction, allowing experimentation with AI responses to foster better understanding of system capabilities. Moreover, handing control over AI settings to the user, and enabling them to familiarise themself with system limitations, further supports clarity and trust [39]. Multimodal explainability, which integrates verbal and nonverbal cues, further enriches user understanding in HRI [6]. Together, explainability and transparency reduce ambiguity by helping users anticipate agent's behaviour and validate its fairness [105, 103]. However, these principles also pose challenges, such as balancing accuracy with interpretability, and tailoring explanation complexity to suit the user's level of expertise, and knowledge of complex artificial systems [103].

Emerging challenges in RAI highlight the importance of authenticity, particularly as generative AI advances raise concerns about deception and the potential for both the production and mishandling of misinformation. Defining, assessing, and embedding authenticity into AI design will be essential to address ethical implications of AI-generated content and maintain public trust [105].

3.2 Social and Cognitive Robotics: Fundamental Constructs

The constructs of social and cognitive robotics provide a framework for understanding artificial agents' perceived intelligence and the dynamics of HAI. Central to the conversation are agency and autonomy, two interlinked but distinct concepts. Agency is broadly defined as the capacity of an entity to take meaningful action, influenced by sociocultural and contextual factors [89]. It describes an ability to affect change in the

environment, which becomes an interesting task when applied to designing agents that lack consciousness or genuine intentionality. Autonomy, on the other hand, emphasises self-direction and independence from external control, a crucial factor in robotic systems designed to operate with minimal human intervention [20]. In other words, autonomy in robotics is often measured by a system's self-sufficiency or "self-determination" [20], and the ability to manage uncertainty. Despite advancements in machine learning, which allow agents to adapt and exhibit complex behaviours, the debate still continues regarding whether true autonomy or agency can exist without subjective experience or embodiment [2].

In discussions of agency in artificial agents, an agent's behaviour resembling the *Theory of Mind (ToM)*, the perceived ability to understand others' mental states, proves critical. Exhibiting ToM enhances the user's perception of an agent's authenticity and credibility by facilitating interactions that mimic human social cognition [66]. If embedded, it enables agents to respond in ways that align with human expectations of mutual recognition and shared understanding, a concept known as intersubjectivity [81]. When users observe agents acting as if they understand emotions, preferences, or intentions, it fosters a belief in the agent's authenticity and *apparent agency*, which is a context of users attributing intentionality and decision-making capabilities to an agent, even if its actions are a result of pre-programmed algorithms rather than genuine agency [72]. The interaction context plays a significant role in determining whether users perceive an agent as possessing these qualities, which [72] discusses in relation to ethics or *appropriateness*. For instance, [72] proposes that "robotic agency" should be attributed in relation to "human agency", meaning that ascribing *one*, inevitably affects the *other*, and technological limitations of robots should be situationally considered.

3.3 Human-Likeness and Social Presence in Design

Embodiment and social presence are fundamental to HRI, with embodied agents directly enhancing perceptions of presence, through embedded expressive cues that convey internal states and intentions in a human-like manner [107, 50]. Social presence, as the perceived "social" engagement with an entity, is vital for users to experience interactions with robots as meaningful and socially rich, and is often accomplished through embedding human-like or anthropomorphic traits [126, 67]. It has also been shown that effective or appropriate embodiment can amplify social presence, allowing robots to be

perceived as more than mere tools, fostering a sense of being with a social entity and enhancing the quality of the interaction [50, 22]. In HRI, embodiment acts as a driver of social presence by allowing robots to engage users multimodally, thereby increasing the likelihood of humans perceiving the robot as a social actor rather than a pre-programmed machine [50].

Intertwined with social embodiment and social presence are the concepts of anthropomorphism and social design cues. Anthropomorphism, the attribution of human characteristics to non-human entities, is a broader construct which affects both the intentional and unintentional design elements that shape user perceptions of robots as human-like [93]. Social design cues, on the other hand, are deliberate design features that evoke specific social responses from users, grounded in social norms and expectations in order to create more natural interactions [11]. The distinction lies in the intentionality: while anthropomorphism may emerge naturally due to users' inclination for personification (often, however, triggered by social cues), social design cues are purposefully integrated to facilitate positive and "credible" [11] human-robot interaction. These cues are often focused on behaviours such as polite language, gestures, and expressions to prompting familiarity, enhancing the perception of social presence [11].

Successfully operationalising anthropomorphism and social cues has been a central design feature of developing trust-evoking agents in HRI and HAI thus far, and I believe it to be a crucial aspect of achieving authenticity, as well. Anthropomorphic elements such as human-like faces, voices, gestures, and even names [126] can foster strong positive social responses. By aligning the agent's communication features (e.g. voice, mimicking thought processes) with its appearance and purpose, designers can enhance its perceived human-likeness and assigned capabilities [93]. On an embodied level, physical features such as a humanoid design as well as expressive behaviours like gaze and nodding make the robot appear more alive and capable of complex social interactions, elevating the user's experience [50]. Complex social design cues such as empathy, politeness, and personalisation are used to communicate warmth and consideration, fostering *authenticity* and encouraging users to engage with robots in ways that are both natural and socially satisfying [37, 123].

However, the purpose of the agent significantly affects how human-like it should be for users to perceive it as authentic. For example, a virtual therapist, whose goal is to provide emotional support (i.e. *someone who has niche expertise in the human, physical world*), might benefit from more traits resembling human-likeness, and focusing on a single task

[102], and ideally adopting a human-like voice [7]. On the other hand, a customer service bot might be expected to prioritise efficiency over social presence [93], where users might align with such a design solution because that agent serves as a "system substitute", rather than a "human substitute", replacing an expert [131]. It is also essential to contextualise the role of perceived empathy, rather, how equipping agents with empathetic behaviours might not be as obvious, beneficial, or straightforward of a design approach to optimising trust or authenticity. As suggested by [51], empathy is defined as an embodied, and a deeply human trait, so embedding it in agents in an overexaggerated way, which does not match user's perception of the agent's capabilities, or its purpose, might lead to unacceptance and discomfort. The issues of discomfort can also be discussed through the familiar label of the "uncanny valley" effect, a phenomenon where a near-human appearance evokes unease and discomfort. This effect, described by [121], occurs when there has been an unsuccessful balancing of human-like traits in design, but the agent is indeed distinguishably artificial, leading to a dip in likeability before reaching the goal of lifelike authenticity [96, 68]. Striking a balance is essential; using selective anthropomorphic cues, appropriate to the context, agent, the task it was developed to perform, as well as the characteristics of the user, seem to be the complex, multilayered approach to design which could lead to the dynamic development of trusting, authentic interactions.

4. Theoretical Framework of Pragmatic Authenticity

4.1 Reasons for *Pragmatic* Authenticity

This subsection presents three detailed reasons as to why I believe future research and industry design practices should be focusing on attaining, measuring and developing for pragmatic authenticity. To my knowledge, the first mentioning of *pragmatic authenticity* as a concept, within the context of human-agent interaction, was in [61], when they described the differences between ontological authenticity [18], and pragmatic or "personal" authenticity. Moreover, the approach to researching and defining individual factors which *pragmatic authenticity* is based upon have also been discussed in prior studies [19, 125, 43, 58]. However, proposed below is a framework whose novelty lies in cohesion and inclusion of various relevant fields, and which places the multifaceted, subjective experience of human users at the forefront of designing and assessing

interactions with artificial social agents. The opposing research and design approach to utility or pragmatism would be developing for ontology, in which case approaches prioritise the replication of, and reliance on human cognitive processes [49, 98, 79]. This distinction is relevant because the aims of research within ontological development, and a functional, pragmatic one, differ significantly. For example, work that could be characterised as ontological, has the aim to "based on the architecture and physiology of the mammalian brain (...) computationally model, at different levels of abstraction, different brain structures and their interactions" [98]. Similar studies have set the goal even higher, aiming to *faithfully* model human emotional processes, and using those models for expanding on theoretical discourse on emotion [35]. On the other hand, the pragmatic authenticity framework emphasises the relevance of human perception and quality emergence which arise from human-agent interactions. This perspective aligns with a broader understanding of authenticity as a construct deeply rooted in the user's ever-changing perception rather than the solely intrinsic properties of the artificial agent. By focusing on pragmatic authenticity, the goal shifts toward understanding what humans need, experientially, from these interactions to perceive them as authentic.

Firstly, this is particularly important given that the pursuit of *ontological* authenticity—replicating human intelligence, empathy, or consciousness—raises significant uncertainties. For instance, the *uncanny valley effect* demonstrates the irregularities in human perception when faced with *overly* human-like agents that do not align with expected interaction patterns. This effect is importantly highlighted because of the observation that humans naturally project familiar patterns from human-human interactions onto their relations with artificial agents [95]. While the projection of patterns is occurring, the valence of these perceptions, whether they are positive or negative, varies greatly based on the context [106], and the process is much more complex than mere transference. This puts an emphasis on the importance of understanding and designing for the nuanced and personalised nature of these experiences. In this light, simply *replicating* human attributes is neither a prerequisite nor a guarantee for achieving authentic interactions. Instead, pragmatic authenticity advocates for creating interactions that users perceive as genuine and meaningful, regardless of whether the agent possesses replicated human processes since such characteristics can benefit, but also hinder the perception of authenticity.

Secondly, the limitations of developing ontological authenticity provide compelling arguments for the shift away from it. The endeavour to replicate human cognitive

processes is hindered by the incomplete understanding of the mechanisms underlying human consciousness and the theory of mind. Without a comprehensive understanding of human cognition, even if such is trivialized to its one aspect of neuronal activity within the brain, attempts to create truly authentic replicas are currently fundamentally flawed, and even, naïve. In their work on computationally modelling emotions, [49] mention how their model cannot yet operationalise what is, in humans, known as the role of the *subconscious* on emotion development and regulation. By labelling this construct as *mood*, rather than subconscious, authors introduce an already existing concept from psychological terminology to stand in for an arguably much more complex, and far less empirically studied one. Furthermore, in trying to provide reasons as to why embodied cognitive-affective architectures should be studied, [97] gives insight into just how qualitatively intertwined, and emergent cognitive processes are, when questioning the validity of the brain's functional specialisation. While the general message of [97] work could be viewed as beneficial for its practical development of agents which are more grounded in interaction and their environment, it is also clearly stated that "complex cognitive-emotional behaviour has their basis in dynamic coalitions of networks of brain areas, none of which should be conceptualized as either affective or cognitive." While [97] does propose a network model based on those findings, it is yet to address the potential of the emergent quality present in humans, similar to the subconscious element not being operationalised in the EMA developed by [49]. To specify, drawing from the 4E cognitive science framework, it is posited that cognition extends beyond the brain into the body and the environment. Based on that, it becomes evident that modelling such complex, dynamic systems within artificial agents poses significant challenges. Even if the fundamental principles of cognition were fully understood, replicating them within the confines of artificial systems would remain a significant creative and engineering hurdle.

Lastly, the pursuit of ontological authenticity raises important ethical concerns. The desire to replicate human attributes in artificial agents may stem from an egocentric drive to mirror human superiority rather than considering alternative designs that are context-specific, safe, and personalised. Such replication efforts risk overlooking the unique capabilities of artificial agents that could be developed to serve specific human needs in innovative ways. By moving away from the replication paradigm, pragmatic authenticity offers a more responsible approach, focusing on the subjective experience of users and

the specific contexts in which social interactions occur. This shift necessitates a focus on practical considerations for evaluating and enhancing authenticity in HAI.

The development of this framework is rooted in my positionality within philosophical, technical, and ethical considerations. Philosophically, the feasibility and desirability of replicating human consciousness and cognitive processes are questioned, given the complexities and uncertainties involved. Technically, I acknowledge the limitations of current technology in achieving ontological authenticity, particularly considering the embodied nature of cognition. Ethically, the potential replication of human biases and the implications of creating human-like systems without addressing their broader impact are significant concerns. This positionality supports the shift toward pragmatic authenticity as a more interaction-centred and human-focused approach, underscoring the human experience, while addressing the limitations of current technologies and ethical considerations. The dynamic and bidirectional nature of human-agent interactions further emphasises the importance of this framework. Authenticity is not a static property of the agent (nor the human and its perception) but rather a co-constructed outcome of the interaction between the human and the agent [94]. This perspective offers a more interactionist approach to assessing authenticity, recognising that it is shaped through constant exchanges between users and agents [94, 1]. Authenticity, in this sense, becomes a "negotiation process" [94] involving developers, consumers, and researchers. Grounding the framework in established research areas such as responsible and explainable AI, and 4E cognition ensures its relevance and applicability across domains. The integration of different perspectives reinforces the need for "epistemological diversity" in understanding the authenticity in HAI [65].

4.2 Components and Formative Structure

The framework I propose for pragmatic authenticity first emerged from a definition of HAI, and the various ways humans interact with agents; the interactions include cars, digital avatars, voice interfaces, and more. Each of these agents serves a unique purpose, and the type of interaction or relationship humans build with them, will depend not only on the agent's function, but also on the individual user, and the specific context. While eliciting trust or introducing anthropomorphic qualities have been proven to be important design aspects in HAI, I argue that, on their own, neither can be the sole factor required

to build meaningful relationships with agents. I also believe there is much interconnectedness to be explored, even within the two aforementioned factors, which calls for operationalisation of these constructs and their further studying. However, guided by the notion that interaction dynamics play a role in what defines an effective, successful, comfortable, or a likeable (yet to be defined as *authentic*) interaction, my research scope undoubtedly included trust as a valuable measure, but also went beyond it, introducing institutional and individual trustworthiness, human-likeness, and intelligence in the proposed framework.

Rather than focusing on trust itself and influenced by [125] on distinguishing between trust and trustworthiness, I wanted to highlight that the perception of trustworthiness does not always lead to a feeling of trust, or, paradoxically, humans might trust entities that are not trustworthy. In the context of HAI, this raises the question about how trustworthiness is communicated and, in turn, understood. [125] also suggests that designing for trust(worthiness) should go beyond agent-centric design and include broader considerations, such as policies and regulatory frameworks, which would implicitly regulate agent's abilities and possible (negative) actions. These ideas influenced my inclusion of institutional trustworthiness as a key factor, emphasising that systemic measures can indirectly influence how humans perceive an agent's trustworthiness.

For individual trustworthiness, my approach is inspired by the field of responsible and explainable AI, however more specifically within HRI, Professor Alessandra Sciutti's insight into transparency and explainability [6] played a pivotal role in my development of all the detailed facets. In [5], Sciutti pointed out that transparency, as traditionally implemented, often fails to meet the needs of lay users, and must be designed in a way that evokes understanding and engagement in all users. This laid the groundwork for my reading into and structuring of the individual trustworthiness factor.

The inclusion of human-likeness stems from a long line of anthropocentric research and findings suggesting human-likeness is a positive design trait, while also insinuating there is a narrow path between comfort and eeriness, regarding the uncanny valley effect. This phenomenon was the reason I wanted to further conceptualise the human-likeness factor, wanting to define various aspects that could be contributing to the positive connotations.

Lastly, intelligence in the framework was guided by the field of philosophy of mind, and its concepts of the theory of mind and agency, which I suppose influence the human perception of the agent's capabilities or *intelligence*, which can be contrasted by its actual capabilities and complexity.

In this subsection, the four factors, alongside their facets, are presented in greater depth.

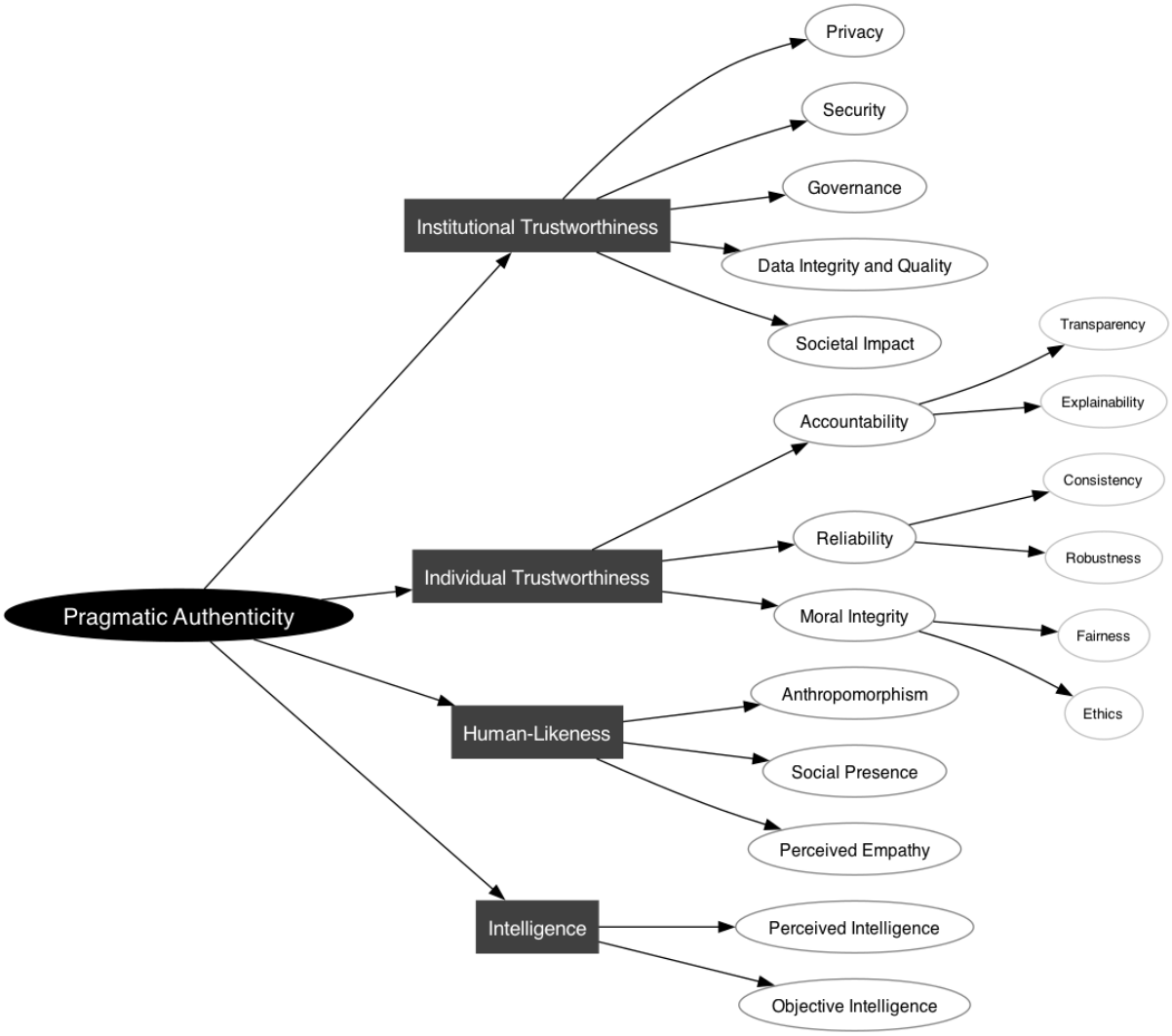

Figure A. Model of The Pragmatic Authenticity Theoretical Framework.

Institutional trustworthiness is a foundational element in the proposed framework of pragmatic authenticity, shaping how artificial systems are perceived and accepted in human-agent interactions. This factor incorporates critical dimensions such as Privacy, Security, Governance, Data Integrity, Quality, and Societal Impact, all of which are interconnected in influencing user trust. These facets could serve as *credibility markers* of institutions managing the systems we daily interact with, in turn contributing to our perception of authentic HAI.

Privacy and security emerge as fundamental concepts in establishing institutional trustworthiness. These elements are particularly significant as they reinforce user

confidence regarding the proper handling of personal and sensitive data. [130] highlights that when discussing agents in conversational settings, trustworthiness is predominantly considered in utilitarian terms, emphasising security, privacy, and transparency over emotional trust. This perspective suggests a shift in the conceptualisation of trust in human-agent interactions, where the focus is on tangible measures such as data protection and ethical data usage. Transparency in how data is processed and the mechanisms through which algorithmic biases are accounted for, plays a pivotal role in fostering accountability, especially when related to the role anthropomorphic perception can play in assigning accountability, as explored by [59]. Ensuring that users can accurately comprehend the capabilities, functioning and decision-making processes of AI systems is central to delivering institutional trustworthiness.

The *Governance* facet further reinforces the trustworthiness of institutions by highlighting the need for ensuring responsible development that is controlled by policies beyond individual, organisational, company-proposed principles. [94] examine the challenges of the process of authenticating and accepting AI, pointing out that negotiations surrounding authenticity are deeply influenced by "political and infrastructural forces", describing how our perceptions of what is intelligent or human-like inherently is influenced by what we consume in the world around us. The emphasis on pragmatic authenticity does require a design approach that prioritises accountability and functionality within the domain, but also implementation of controlled ethical alignment in the use and development of technology, within a top-down approach in hierarchy.

Data integrity and quality also constitutes an essential component of institutional trustworthiness. The lack of transparency regarding data usage can create scepticism about the reliability or fairness of agents. Moreover, [94] underscores the importance of demonstrating a commitment to accuracy and unbiased data, highlighting the plethora of systems being trained on (inevitably) biased data, by those who (inevitably) will instil some biases during the design process.

The *Societal impact* of AI extends the scope of institutional trustworthiness by exposing the need for an active examination of how AI defines our society. [94] discuss the overlap in the tasks given to artificial agents, with the questions about the future of human work. This perspective sheds light on consequences of AI deployment, as the pursuit of overly predictable AI systems puts in question "the indeterminacy characterising human behaviour", viewed as a fundamental aspect of human agency [94].

Inherently connected to institutional, the second factor explored in the framework is that of *Individual Trustworthiness*, serving as a critical component in fostering trustworthy interactions directly between a human and its artificial interlocutor. Within this framework, individual trustworthiness is defined through three interrelated dimensions: Reliability, Moral Integrity, and Accountability. These are further defined as consisting of Robustness and Consistency (Reliability), Fairness and Ethics (Moral Integrity), and Transparency and Explainability (Accountability).

Research highlights the importance of reliability as a key determinant of trust, particularly in contexts involving human-robot interaction. [44] defines performance trust as a measure of whether an agent's actions meet user expectations, directly linking the agent's ability to perform reliably with the user's confidence in the interaction. This emphasis on consistent performance underscores the role of reliability in shaping perceptions of trustworthiness. From a methodological standpoint, [91] points to the critical need for reliable measurement instruments in assessing human perceptions of robots, noting that inconsistent evaluation methods can influence how trustworthiness is perceived. This reinforces the notion that reliability extends beyond an agent's capabilities and extends onto the tools and methodologies used for agent capability evaluation. As [48, 74, 53] explore in their work on error scenarios and unreliability in HRI, failures or inconsistencies in performance can significantly undermine trust and intelligence perceptions, and even minor deviations from expected performance can lower user confidence. The research suggests that agent resilience in challenging scenarios is not only an expectation but a prerequisite for sustaining perceptions of reliability, intelligence, and trust.

Moral integrity addresses the ethical dimensions of an agent's behaviour and interactions. [44] observe that much of the research on trust in human-robot interaction has focused on performance-related factors, often overlooking the ethical considerations that are vital for fostering trust. This omission underscores the importance of integrating moral integrity into the framework, where fairness and ethics serve as guiding principles. [131] draw a distinction between "goodwill-based" and "qualification-based" trustworthiness, arguing that competence alone is insufficient for establishing trust, and that the *perceived intentions* of an agent play a crucial role, influencing how users evaluate its morality. The framework aligns with these observations by incorporating fairness and ethics as subfacets of moral integrity, alongside factors of individual trustworthiness (i.e.

transparency) emphasising the need for agents to act in ways that are aligned with human values. [44] further explore the moral dimensions of trust in their framework, questioning whether agents exploit human vulnerabilities due to a lack of moral integrity. This ethical lens reinforces the expectation that artificial agents should not only perform competently but also follow the principles of fair and ethical conduct.

Accountability focuses on the importance of transparency and explainability in fostering trust in artificial agents. The complexities of accountability are exposed in probabilistic automation systems, as [59] notes, where users may perceive these systems as autonomous agents. This perception raises concerns about users assigning *responsibility* or *blame*, especially when users attribute "intelligence" to these entities. Employing mechanisms to ensure accountability are essential, to allow users to hold systems responsible for their actions and decisions. Transparency, as [104] argues, is particularly critical in addressing ethical challenges, such as robotic deception. This perspective underscores the necessity of making AI systems comprehensible to users, ensuring that their functionality is both transparent and explainable. Furthermore, [44] identify transparency and explainability as key dimensions of moral trust in robots, reinforcing the idea that users need to understand the rationale behind an agent's actions (transparency), and to be able to provide justifications for those actions (explainability), to assess its trustworthiness. Within the PA framework, accountability is described through both transparency and explainability; where the operationalised definitions are adopted from [44], highlighting the importance of users knowing the *how* and *why* of agent's processing and behaviour.

The third factor influencing pragmatic authenticity of interactions is *Intelligence*. Intelligence in this context is conceptualised through two primary facets: Objective Intelligence, which encompasses the measurable capabilities of the system, and Perceived Intelligence, which reflects how users interpret and evaluate an agent's cognitive abilities. *Objective intelligence* are the inherent capabilities and complexity of an artificial agent. It is often evaluated through its ability to perform tasks traditionally associated with human intelligence, such as reasoning, problem-solving, learning, and adaptive action within an environment. [73] define intelligence as "the ability to accomplish complex goals, learn, reason, and adaptively perform effective actions within an environment," encapsulating its foundational attributes. This conceptualisation aligns with the broader understanding of artificial intelligence as systems designed to replicate certain human

cognitive functions. Further, [94] extend this notion by discussing artificial general intelligence (AGI), which represents the aspiration to develop AI systems capable of human-like cognitive functioning across a variety of domains. However, within the framework of pragmatic authenticity, objective intelligence is not defined by the pursuit of AGI, rather, it stands for the mere description of functional complexity and learning capabilities of agents, as articulated by [73], and contextualised by [132] and foundationally framed by [Russell & Norvig] when creating a general categorisation of artificial agents.

Perceived intelligence, in contrast, focuses on how users interpret and assess an agent's cognitive abilities during interactions. This facet is deeply influenced by constructs such as agency, apparent agency, and Theory of Mind (ToM), which shape the subjective experience of interacting with artificial agents. The distinction between agency and apparent agency is particularly significant. Agency, as traditionally understood, implies intentionality and the capacity for autonomous action driven by goals and desires. However, in the context of artificial agents, *true agency* is a grey area, since agency is often discussed in terms of "human" and "machine agency" [2]. Machine agency, emerging from human attribution and mere inference, is a mirror to *apparent agency*, which refers to the perception of intentionality and decision-making, even in the absence of genuine autonomy [65]. Whatever the definition, it has been established that agents that exhibit behaviours suggestive of intentionality (i.e. anthropomorphic behaviours) are often perceived as more intelligent, regardless of their underlying algorithms [126]. This focus on *perception* is indeed integral to the PA framework, as it prioritises user experience over the objective inner workings of an agent.

Theory of Mind (ToM) further enriches the concept of perceived intelligence by introducing the ability to attribute mental states—such as beliefs, desires, and intentions—to others. ToM is a cornerstone of social interaction, enabling humans to predict and interpret the behaviour of others [85, 66]. As such, agents equipped with ToM-like capabilities might foster trust and credibility by aligning their behaviour with human expectations. Moreover, ToM (or the perception of one possessing it) facilitates intersubjectivity, or the shared understanding that emerges during communication, which is critical for effective collaboration, as it contributes to the perception of the agent as a genuine partner in interaction [33, 81].

The potential benefits of developing ToM-like abilities in artificial agents could include improved human-agent collaboration and more socially adept interactions, however,

ethical concerns are apparent with the development of strikingly anthropomorphic agents, particularly around the potential for manipulation and deception, as well as the blurring of boundaries between humans and machines [96, 104]. From a constructive standpoint, the concept of "Theory of an AI's Mind" (ToAIM), introduced by [85], offers an alternative perspective by focusing on a human's ability to predict and understand an agent's behaviour. ToAIM emphasises tasks such as failure prediction and knowledge prediction, suggesting that perceived intelligence is shaped not only by the agent's capabilities but also by the user's understanding of its limitations and strengths (i.e. user's "intuition" about agent's capabilities). This aligns with the framework's emphasis on transparency and trust, as users are more likely to engage comfortably with agents whose workings and intentions are comprehensible.

The interaction between objective and perceived intelligence further highlights the dynamic nature of intelligence in artificial agents, as [85] found that increasing users' understanding of an AI model's internal mechanisms enhanced their ability to predict its behaviour, demonstrating that knowledge of objective capabilities can influence perceptions of intelligence.

The exploration of *human-likeness* comes as the fourth and last factor of this novel framework. Within pragmatic authenticity, human-likeness is examined through three facets: Anthropomorphism, Social Presence, and Perceived Empathy. Each of these facets has played a pivotal role in foundational as well as state-of-the-art HRI and HAI research, therefore my aim for this framework is to contextualise these concepts within a broader perspective.

Anthropomorphism, or the attribution of human characteristics, behaviours, or motivations to non-human entities, is a central component of human-likeness [95]. In the context of artificial agents, anthropomorphism often manifests through deliberate design choices that evoke human forms, traits, or social roles. For instance, assigning a conversational agent a distinct, human-like voice can significantly enhance its perceived humanness [7]. This design approach uses familiar social patterns or occurrences, enabling users to relate to the agent as they would to another person. An example is a voice assistant mimicking human-human communication style, which fosters a perception of the agent as human-like, and more likeable [86]. While this approach can enhance user engagement and acceptance, excessive anthropomorphism, particularly within physical appearance, can lead to unrealistic expectations, and also the uncanny

valley effect. This phenomenon, where agents that are almost but not quite human-like evoke feelings of unease or discomfort, highlights the balance required in anthropomorphic design [68, 121].

Social presence represents another essential facet of human-likeness, referring to "the extent to which other beings in the world appear to exist and react" to us [24]. This dimension is particularly relevant for designing disembodied agents such as voice assistants, where physical cues are absent, and social presence is employed through various design strategies based on anthropomorphism, including the use of common linguistic factors, emotional responsiveness, and the ability to engage in meaningful conversations [131, 93].

Perceived empathy, the third facet of human-likeness, describes the human's perception of an agent's ability to understand and respond to human emotions, creating a sense of shared understanding and connection. The perception of *empathy* is grounded in the quality of communication, meaning it can be conveyed through both verbal and non-verbal cues, such as validative behaviour and responsiveness, or showcasing interest and concern [67]. *Empathetic* behaviour does not indicate an agent's ability to empathise, but it might contribute to the experience of the agent being perceived as empathetic, based on the behavioural markers that were portrayed.

To conclude, the proposed framework highlights the interconnectedness of institutional and individual trustworthiness, intelligence, and human-likeness, for which the goal is to explore the potential relationship toward the latent factor of pragmatic authenticity. This proposed interrelation emphasises the potential of the framework to provide a comprehensive guide to research and design in human-agent interaction, however it is not without its challenges. To have this framework be relevant, it will take extensive validation for both the relevance of the proposed factors within the framework, and its interconnected, possibly bidirectional, nature. Moreover, even if validated, such a framework will remain open to further refinement, particularly because of the quickly evolving technological environment where new factors continually emerge.

5. Empirical Study of Human-Likeness in Pragmatic Authenticity

The purpose of this study is to explore the relationship between human-likeness and pragmatic authenticity in interactions with artificial conversational agents as a first partial empirical proof-of-concept validation of the proposed framework. Human-likeness, operationalised through the constructs of social presence, anthropomorphism, and perceived empathy, plays a crucial role in how users perceive and interact with technology – computers and artificial agents, both embodied and disembodied. The concept of pragmatic authenticity aligns with the aim of [18] examination of deceptive robots, where it is highlighted that the focus of proposed queries and definitions might have to be on the coherent and trustworthy "qualities that emerge from these interactions", regardless of whether the agent possesses inherent human traits.

While previous studies in HRI and HAI have focused on the design and implementation of human-like features [126, 93], and their role in sustaining trust between users and agents [17], there is still much to understand about whether, or how, human-likeness fits into the framework of building authentic interactions. This study is considered a pilot study, a first step towards empirically testing the proposed theoretical framework of pragmatic authenticity. The study builds on recent advancements in human-robot and human-agent interaction, and cognitive science, drawing on ideas of interaction-centred design [65] and social grounding. It also addresses the lack of agreed-upon terminology, by redefining some and introducing other concepts and novel methodologies. By investigating the interplay between human-likeness and authenticity, the research aims to quantitatively and qualitatively describe this relationship, as well as to establish factors that contribute to users' perceptions of *authentic* interactions with artificial agents.

This empirical study was designed to investigate eight primary research questions and hypotheses, with a focus on the relationships between human-likeness, authenticity, expertise, and attitudes toward artificial intelligence. The hypotheses are based on existing literature, which suggests that concepts of social presence, anthropomorphism, and likeability might all be related to human perception of, what is in this research called, artificial agents' *human-likeness* [28]. Previous research also suggests possible differences in perceptions of artificial agents based on characteristics of the user, such as previous experience with artificial intelligence, personal values and beliefs, and expertise [129]. Following those ideas, in this study, we also explore whether expertise and

attitudes toward technology may moderate relationships between the measured concepts of human-likeness and authenticity. Hypotheses regarding the constructs of authenticity, although with valid grounding in existing literature, were made based on the proposed novel theoretical framework, and while directional, could for that reason be viewed as exploratory research questions. This study will also introduce a novel semantic differential scale to measure the authenticity of artificial agents and evaluate its reliability.

5.1 Research Questions and Hypotheses

This subsection presents the research questions (RQ) and hypotheses (H) that guide the empirical study. Each research question addresses a specific aspect; the relationship of factors within human-likeness, the relationship between human-likeness and authenticity, and the role of expertise and attitudes toward artificial intelligence in authenticity and human-likeness perception.

RQ1: What is the relationship between social presence, anthropomorphism, and perceived empathy in artificial intelligence interactions?

H1: Social presence, anthropomorphism, and perceived empathy will show high positive correlations with one another.

This question investigates whether the three key components of human-likeness are interrelated. Drawing from studies on anthropomorphism and social presence, relating to trust, it is hypothesised that these factors will be positively correlated [131]. Studies in HRI suggest that controlled empathy and other anthropomorphic behaviours are valuable for creating engaging, natural interactions [83].

RQ2: Is there multicollinearity between social presence, anthropomorphism, and perceived empathy?

H2: There will be a significant effect of multicollinearity between social presence, anthropomorphism, and perceived empathy.

Multicollinearity refers to a situation where two or more independent variables are highly correlated, which can cause them to explain the same portion of variance in the dependent variable. In other words, if present, multicollinearity can also be interpreted as a presence of a latent factor among the three constructs. Previous research on anthropomorphism and

social presence suggests these constructs may correlate significantly [126], and the presence of multicollinearity would suggest that they are not independent dimensions.

RQ3: Is the novel semantic differential scale for authenticity reliable and internally consistent?
H3: The semantic differential scale for authenticity will demonstrate high internal consistency and reliability, as measured by Cronbach's alpha.
This hypothesis tests the reliability of a newly developed tool for measuring authenticity in AI interactions. Based on psychometric principles, the scale is expected to show strong internal consistency, supporting its use in future studies on pragmatic authenticity.

RQ4: What is the relationship between social presence, anthropomorphism, and perceived empathy (aggregated as human-likeness), and perceived authenticity of artificial conversational agents?
H4: There will be a high positive correlation between the aggregated human-likeness score and perceived authenticity of artificial conversational agents.
While acknowledging the presence of the uncanny valley effect, and the possibility of this hypothesis being rejected with valid reasoning [96, 83, 68], this hypothesis explores the direct relationship between human-likeness (as a composite of social presence, anthropomorphism, and perceived empathy) and perceived authenticity. Considering the hypothesis, previous work suggests that, although context-specific, greater human-likeness enhances users' trust and authenticity judgments of artificial agents [126, 114, 102].

RQ5a: Can expertise serve as a grouping variable in understanding human-likeness perceptions in artificial conversational agents?
H5a: Expertise will serve as a significant grouping variable, with experts (high expertise) showing significantly different perceptions of human-likeness compared to students (low expertise).
RQ5b: Can expertise serve as a grouping variable in understanding authenticity perceptions in artificial conversational agents?
H5b: Expertise will serve as a significant grouping variable, with experts (high expertise) showing significantly different perceptions of authenticity compared to students (low expertise).

These hypotheses examine the role of expertise in shaping user perceptions of human-likeness and authenticity. Because studies in HRI, have found inconclusive (all positive, negative, as well as insignificant) correlations between familiarity or expertise, and perception of interaction [64, 30, 32, 129], a significant difference is hypothesised, in that participants with higher expertise will have different views of their interaction with artificial social agents, than those of lower expertise, however no direction is assigned.

(Post-hoc) RQ6: Do expertise levels moderate the relationship between human-likeness and authenticity?
(Post-hoc) H6: Expertise will moderate the relationship between human-likeness and authenticity.
This post-hoc hypothesis extends the investigation into whether expertise acts as a moderator, influencing the strength or direction of the relationship between human-likeness and authenticity.

RQ7a: What is the relationship between attitudes toward artificial intelligence and perceived human-likeness?
H7a: There will be a high positive correlation between (positive) attitudes toward artificial intelligence and perceptions of human-likeness.
RQ7b: What is the relationship between attitudes toward artificial intelligence and perceived authenticity?
H7b: There will be a high positive correlation between (positive) attitudes toward artificial intelligence and perceptions of authenticity.
These questions and hypotheses address how participants' attitudes toward technology and AI influence their perceptions of human-likeness and authenticity. Prior research indicates that users' perceptions and attitudes towards AI, its *intention* and purpose is correlated with their trust in agents [100].

RQ8: Do attitudes toward artificial intelligence moderate the relationship between human-likeness and authenticity?
H8: Attitudes toward artificial intelligence will moderate the relationship between human-likeness and authenticity.

This final research question investigates whether participants' attitudes towards AI strengthen or weaken the relationship between human-likeness and authenticity, potentially shaping the overall interaction experience.

Despite having prior research available within related fields, some hypotheses are posited as non-directional to allow for a more exploratory approach. This acknowledges that while relationships between variables are likely, the specific nature or direction of these relationships may vary depending on contextual factors. However, in cases where existing literature and theory provide a stronger foundation, or a research question is approached with a desired outcome in mind, directional hypotheses are used to test specific, predicted outcomes, ensuring the analysis is both comprehensive and flexible in its approach.

5.2 Experimental Design

The study used a series of variables to assess the relationship between human-likeness and authenticity, as well as the role of expertise and attitudes in moderating these perceptions.

Independent variables were *expertise* (students vs. experts) and *attitudes*, as key grouping variables used to assess differences in perception. Other independent variables included anthropomorphism, social presence, and perceived empathy, all contributing to the concept of human-likeness. The main dependent variables in the study were human-likeness (an aggregated measure of the independent variables) and authenticity of the artificial conversational agent.

The study employed a within-subject design, where each participant experienced all three experimental conditions (social vignettes). This design was appropriate for the study because it allowed participants to engage with the agent in three affectively different contexts, and based on their overall experience, rate the interaction. By exposing participants to all three contexts, the goal was to better equip them for evaluating the agent, showcasing its potential adaptability and versatility in responses; it aimed to prevent any bias that might arise from evaluating the agent on a *single* interaction. Every participant was presented with all three conditions, but the order of presentation was randomised. The randomisation process was implemented automatically through the used tool "LimeSurvey", reducing any potential bias caused by the order in which the conditions were presented.

5.3 Participants

The study involved two primary participant groups: students and experts, selected intentionally to explore whether expertise would influence perceptions of human-likeness and authenticity in interactions with artificial conversational agents. Students were recruited via emailing lists from the University of Zagreb and the University of Vienna, targeting cognitive science students. Experts were recruited primarily through targeted emails sent to colleagues within the Vienna University of Technology (specifically, the HCI Research Group) and the University of Zagreb (the faculty of the Applied Cognitive Science programme). Both groups represented an accessible population, while also having the knowledge in areas relevant to HCI and HAI. The distinction between students and experts would allow for comparisons across different levels of experience and familiarity with digital technologies and artificial intelligence, forming the basis for key research questions. Given that this study is considered a pilot study, the intention was not to have a large sample size, but to collect enough data to conduct appropriate analyses, with the specific aim being to collect data from 30 participants to meet the requirements for parametric statistics. However, due to technical malfunctions and failed attention checks, the final number of participants did not meet initial expectations. Final sample size (n(students)=10, n(experts)=13), while slightly smaller than anticipated, was sufficient for a pilot and allowed for the use of non-parametric tests to conduct meaningful exploratory analyses.

There were no age requirements for participants, however, all participants were required to have a certain academic background. The students in this study were enrolled in master's programs in interdisciplinary fields, such as cognitive science, or its sub-disciplines, and related digital and information sciences. These fields were selected to ensure participants had sufficient familiarity with the topics of human-human and human-agent interaction, as well as artificial intelligence alone. The experts were professionals working within academia, with expertise in the same areas mentioned above, with a focus on human-computer interaction, robotics, AI and machine learning, and cognitive linguistics. These criteria ensured that the expert group had substantial knowledge and experience in fields relating to communication and interaction, whether in the human-human or the human-technology context.

5.4 Methods

The LimeSurvey platform was employed for its simplicity for the user, variety of design and presentation options for the designer, and prior familiarity with the interface. It was used to present instructions and collect participant responses, as well as to present the experimental contexts, *social vignettes*.

Social vignettes were a key methodological tool in this study, used to simulate real-world contexts for interactions with the agent. A vignette is a brief, descriptive scenario that participants read and imagine themselves as part of [122]. However, as [127] point out, a definition of a vignette should be grounded in its purpose, rather than its content, seeing how vignettes can be focused on fictional, as well as every-day scenarios. [122] highlights how vignettes are effective tools for gathering responses in social and behavioural research, particularly when direct observation of behaviours is difficult or impractical. They are widely used in research to create standardised experiences. In HCI, vignettes allow researchers to investigate how users engage with technology in varied contexts, without needing to develop complex or resource-intensive physical environments, or empirical studies of a grand-scale.

In the case of [13], digital vignettes were used to study the negativity on sensitive cultural topics in the digital realm, further developing the vignette approach, using complementary stimuli. What [80] deem a vignette in their study, is an *approach* to research, an *experiential* vignette, characterised by less rigour and extensive protocols of study procedures, which benefit targeting specific results, proposed through research questions. While this is not the semantic definition of vignette used in this study, a shared goal or a board purpose is reflected in both studies; using a methodological approach to elicit specific responses in what would otherwise be a wide interaction space. Vignettes have also been employed in human-robot interaction studies to explore user perceptions of robot behaviour in a medical setting, measuring perceived emotional intelligence, trustworthiness, and acceptability, as well as the perceptions of the patient who is being taken care of by the robot, as shown by [47]. Vignettes in HRI/HAI are particularly suited to studies like this one, where a framework is in its developmental stages. The social vignettes used in this study were designed to encourage advice-seeking interactions between participants and the agent. *Advice-seeking vignettes* are a specific type of vignette that place participants in situations where they are encouraged to seek guidance or solutions to problems. This method of simulating human-agent interaction is particularly useful in examining how people perceive agents in social contexts.

Advice-seeking vignettes are commonly used in social sciences to study help-seeking behaviours and how people interact in contexts requiring assistance or problem-solving. This context was chosen to closely mimic real-world scenarios in which individuals seek advice from other humans. In HAI research, using advice-seeking vignettes allows researchers to assess how well artificial agents perform in roles typically reserved for humans, such as providing guidance or emotional support. The advice-seeking vignettes used in this study focused on three core areas of human life: *health, social relationships,* and *career*. These vignettes were generated using ChatGPT to simulate realistic and relatable problems, and the specific topics were chosen because they reflect common issues that people seek advice on, enhancing the realism of the interaction with the conversational agent. Research in social psychology, more specifically interaction and communication, highlights the importance of creating (and maintaining) relationships under the conditions of anxiety and stress, where the social aspect of creating *affiliations* is hallmarked by striving for, approval, friendship, and, most importantly in the context of this research: *social support* [14]. By asking for advice, participants are encouraged to engage with the agent on a deeper level, simulating a more authentic social interaction.
As the *artificial disembodied conversational social agent* for this study, ChatGPT-4o was chosen. This agent was chosen for its advanced natural language processing capabilities, allowing for nuanced interactions. ChatGPT-4o was used in its *Voice Conversation Mode*, so participants were having a spoken conversation as a mode of interaction. This was a crucial aspect, as conversation social design cues (tonality, emotional expression, speech fidelity) [11, 7] highlight the study's focus on artificial conversational agents as *social*, rather than solely mechanical agents. Since conversational agents are increasingly prevalent in real-world applications (e.g., customer service, personal assistance), the idea was to understand these interaction characteristics in simulated environments.

5.5 Materials

The instruments selected for this study are established, validated tools commonly used in research on HCI, HAI, and HRI. These tools were chosen based on their reliability in measuring key factors relevant to this study. The semantic differential scale of authenticity is a novel instrument of measuring authenticity which was necessary to implement in order to define and approach authenticity, as defined by the framework.

The two Likert-scale-based questionnaires distributed prior to experimental conditions were questionnaires regarding participants' attitudes toward artificial intelligence. The Attitudes Toward AI (ATI) scale, developed by [8], is a validated questionnaire widely used in HCI and AI-related research, designed to assess participants' general attitudes towards artificial intelligence and readiness to accept technology. The scale includes items related to perceived usefulness, fear, and trust in AI, and is critical in establishing participants' baseline attitudes towards AI before interacting with the agent.

Another scale relating to attitudes is the GAAIS (General Attitudes Toward AI) scale, developed by [76] GAAIS extends the measurement of AI-related attitudes to include broader societal implications by including items that assess participants' views on the ethical, social, and economic impact of AI. The inclusion of GAAIS is relevant as it examines participants' perceptions of AI's role in society which transcends their individual interaction but could very well influence it.

The questionnaires which followed the experimental conditions were measuring anthropomorphism, likeability, and social presence. The Godspeed Questionnaire [91], specifically the Anthropomorphism and Likeability scales, is a widely used measurement tool in human-robot and human-agent interaction research. The Anthropomorphism scale assesses how human-like the agent appears to the participant, while the Likeability scale measures the participant's emotional response to the agent. Anthropomorphism is crucial to this study as it directly addresses the degree to which participants perceive the agent as human-like. The Likeability scale is used to operationalise the construct of *perceived empathy* within human-likeness, with the idea that assigning someone "high likeability" could be reflected in our perception of them to also behave *empathetically*. These ideas stem from seminal psychological and sociological theories, such as [4] Halo Effect, which discusses a positive bias in which our positive impressions of someone influence us to assign other positive qualities to them, without any proof of these traits. It can also be connected to the *Similarity-Attraction Hypothesis* [112], where it is noted that we are more likely to feel empathy towards those whom we like. Lastly, Likeability was chosen to represent perceived empathy due to its focus on affective traits within the Godspeed questionnaire in which it was assessed, such as friendliness, kindness, and approachability, all of which manifest through behaviour. The Godspeed descriptors align with comforting and supporting behaviours that are also foundational to perceived empathy, specifically in advice-giving contexts, such as our study. These scales are

among the most validated tools in HRI and HCI research, developed to assess human perceptions of robots and human-agent interaction [91].

The *Networked Minds Measure*, or the *Social Presence Questionnaire* (SPQ), created by [78], measures the degree to which participants feel a sense of social presence when interacting with an artificial agent. Social presence refers to the feeling that the agent is "there" with the participant, contributing to a sense of co-presence or mutual awareness. This is particularly important when evaluating the pragmatic authenticity of disembodied agents, as agents that are perceived to be more socially present are often judged as more trustworthy [38]. The SPQ was adapted for use with disembodied agents in this study. By focusing on the items that demonstrated the highest internal consistency, the adapted SPQ allows for concise yet precise measurement of the social presence factor.

The novel instrument created for this study is the semantic differential scale of authenticity which consists of five items, each designed to capture different dimensions of perceived authenticity in artificial agents. This type of scale, same as Godspeed scales, uses bipolar adjectives to measure attitudes or perceptions along a spectrum. The five items included in the scale are Truthful – Deceptive, Consistent – Unreliable, Coherent – Confused, Authentic – Inauthentic, and Authorized – Unauthorized. These items were chosen based on existing literature that links used terminology to the broader construct of authenticity in human-agent interactions, ranging from topics in RAI, social and cognitive robotics, as well as anthropomorphism and social presence in design (Section 3). For example, truthfulness and coherence are traits related to fostering trust and transparency, while reliability helps establish a sense of stability and confidence in the interaction. The concept of authorisation addresses the agent's perceived legitimacy, which is (possibly) fundamental in determining how users evaluate the agent's authenticity. The choice of these terms is based in the concept of *lexical familiarity*, which refers to how frequently certain words are used and understood within a culture or a language. In the context of psychometry, specifically semantic differential scales, participants might not fully grasp the nuances of "authenticity," however, they are more likely to understand related, more familiar terms such as truthfulness or reliability. This approach allows for more accurate and meaningful responses, as participants are better equipped to relate to familiar, everyday language. Some of the items used (truthfulness, consistency, coherence) were used as a direct reference to [110] and his description of the subordinate concepts of authenticity.

The expertise scale used in the study was adapted from the *computer expertise* scale by [8], In this case, the scale was tailored to measure interaction expertise, focusing on three domains: human social interaction, human-agent interaction, and advanced technology and AI. Each domain was assessed using a six-point Likert scale, ranging from 1 ("No experience") to 6 ("Recognized expert"). Participants provided self-assessments in these areas, which were cross-checked with their cohort grouping (students or experts) to ensure that their expertise aligned with the study's research questions regarding how expertise influences perceptions of human-likeness and authenticity.

6. Data Collection

6.1 Study Procedure

The study took place over the course of two weeks, and in two cities, Zagreb and Vienna, conducted at the offices of the University of Zagreb and Vienna University of Technology. Upon arrival, participants were welcomed into the testing room and given an introduction to the study. The introduction provided participants with the study's motivations, aims, and goals, while avoiding any mention of expected outcomes or hypotheses to prevent bias.

After the introduction, participants were provided with a MacBook Pro, which was used for the experiment. Each participant was presented with a digital consent form that reiterated the details of the study, as well as their rights and data privacy information. Once participants signed the consent form, they were introduced to ChatGPT 4.0, the artificial disembodied conversational agent used in the study. Notably, none of the participants had prior experience with ChatGPT's voice assistant feature, so the experimenter provided a demonstration on how to use the feature. The example interaction involved asking ChatGPT for activity recommendations, such as "It is a sunny day in Vienna, what should I do in the fourth district?" After the demonstration, participants were invited to conduct a trial interaction with ChatGPT to ensure they were comfortable using the voice assistant feature. Once participants successfully completed the demo, they were asked to open LimeSurvey, the platform used for data collection. Participants were instructed to input a Participant ID, which they were allowed to choose freely but were advised not to use their real names for privacy reasons. All participants followed this recommendation. After inputting their Participant ID, participants were

presented with an official introduction to the study on LimeSurvey. At this point, the experimenter stepped back, instructing participants to approach them if they had any questions during the study. Participants were also informed that they could stop the experiment at any time if they felt uncomfortable or needed a break. The experimenter then left the participants to complete the study independently.

The LimeSurvey was divided into four sections: pre-questionnaires, main study, post-questionnaires, and demographics. The first section involved two pre-questionnaires designed to capture the participants' pre-existing beliefs and attitudes toward artificial intelligence: the GAAIS questionnaire and the ATI scale. Both questionnaires were used in their original form, including instructions and items, and measured attitudes on a Likert scale, 5-point (GAAIS), and 6-point (ATI), respectively.

Participants then progressed to the main part of the study, which involved interacting with ChatGPT based on three social vignettes. Before starting this section, participants were provided with an introduction explaining what vignettes are and how to interact with ChatGPT. After reading each vignette, participants interacted with ChatGPT to seek advice, simulating real-life advice-seeking scenarios. After every interaction, participants were asked to type a one-sentence summary of the advice they received to ensure that the topic of conversation was consistent with the vignette. Participants were given no specific time limits or rules on how to conduct the interaction beyond being told to continue the conversation until they felt satisfied with the advice they received, or until they felt the interaction was not developing further. The available voice *personality* of the agent was the only one available at the time of study conductance, but as per ChatGPT, *Version 1.2024.282*, it is the voice known as *Juniper*. Juniper is designed to be "friendly and engaging", characterised by a lively, androgynous voice.

Once all three interactions were complete, participants moved on to the post-questionnaires, which measured human-likeness and authenticity. The human-likeness constructs were assessed using the validated Social Presence Questionnaire (SPQ), the Godspeed Anthropomorphism scale, and the Likeability scale. The authenticity of the agent was measured using a novel semantic differential scale created specifically for this study.

Finally, participants completed a demographics questionnaire, which included basic information such as age, gender, and academic background, as well as a self-assessment of expertise in fields relevant to the study.

The study took approximately 45 minutes to complete, after which participants were given the opportunity to ask any questions or provide feedback on the content and execution of the study. Participants were also given a debrief which included an overview of expected results and explanation of hypotheses. Following the end of the study, participants were emailed the signed copy of their consent form. The chat history of interactions was deleted from ChatGPT to ensure participant privacy and no interference for future participants interacting with the agent.

To ensure transparency and facilitate further research, all novel materials used in this study, including the authenticity questionnaire and social vignettes, are provided in the appendices.

7. Data Analysis

7.1 Data Preparation and Cleaning

In preparing the data for analysis, several key steps were taken to ensure accuracy and completeness. Initial data cleaning involved evaluating participant responses for missing data, attention check compliance, and accurate scoring based on established Likert and semantic differential scales.

From an initial participant pool of 26, three participants were excluded. Two participants were removed due to significant missing data, as their responses were only partially recorded in the LimeSurvey platform, resulting in over 50% missing responses. Given the small sample size, their inclusion would have risked introducing bias or error in subsequent analyses. Additionally, one participant failed the attention check embedded in the *GAAIS*, leading to their exclusion. This check was crucial in assessing data reliability, and exclusion criteria were set to maintain high data quality. All measures used in this study were Likert-type scales or semantic differential scales, requiring specific coding and, in some cases, reverse scoring for negatively worded items.

The *ATI* scale employed a 6-point Likert format (1 = completely disagree to 6 = completely agree). Items 3, 6, and 8 were negatively worded, so these responses were reverse-coded to ensure consistency in scoring. After transformation, a mean score across all nine items was computed. This scale was analysed for mean, standard deviation, and Cronbach's alpha reliability (to be reported in relevant subsections).

The *GAAIS* consisted of positive and negative, 5-point Likert subscales, with items reverse-coded accordingly. Positive responses ranged from 1 (strongly disagree) to 5 (strongly agree). Negative items required reverse scoring, with higher scores indicating positive attitudes toward AI. Scores were analysed separately for the positive and negative subscales, without creating an aggregate scale mean, following the scale guidelines.

The *Godspeed* scales were 5-point semantic differential scales ranging from negative to positive descriptors (e.g., 1 = inhuman to 5 = human). No transformations were required, as the scale directly aligned with the intended scoring, with higher scores indicating higher anthropomorphism or likeability.

The *Social Presence Questionnaire* used a 5-point Likert scale (1 = totally disagree to 5 = totally agree) without any reverse scoring. As with the Godspeed scales, higher scores reflected a more positive perception of social presence, allowing for direct analysis.

Following initial preparation in Excel, the data was imported into SPSS and Python for analysis and visualisation. SPSS was used for statistical analysis and reliability checks, while Python was employed to create standard and advanced data visualisation.

7.2 Demographics

The study's final sample had a total of n = 23 individuals. The gender distribution included 52.2% man and 43.5% woman, and 4.3% non-binary, with participants' ages ranging 19-63 years and an average age of m = 33.61 (SD =11.085). Among the participants, 56.5% were identified as experts in their respective fields, while the remaining 43.5% were non-experts (students). This categorisation enabled targeted comparisons of expertise levels in the subsequent analyses.

7.3 Authenticity Scale Reliability Analysis

In alignment with RQ3, a reliability analysis was conducted of the novel semantic differential scale measuring authenticity perceptions. The initial analysis, which included all five items (Truthfulness, Consistency, Coherency, Authenticity, and Authorisation), yielded a Cronbach's alpha (α) value of 0.376. This low reliability coefficient indicated that the scale, in its initial form, did not consistently measure the construct of authenticity across items, so to further examine the scale's structure, item-total statistics were reviewed. The analysis revealed particularly low inter-item correlations for the item *Authorized/Unauthorized*, and based on this observation, *Authorized/Unauthorized* was

excluded from the scale to improve internal consistency. Following the removal of *Authorized/Unauthorized*, the scale was recalculated with the remaining four items, resulting in a revised alpha value of $\alpha = 0.614$. While this still indicates moderate reliability, it is within an acceptable range for exploratory research [101]. Although with potential for refinement, the revised alpha did support the use of the scale as a measure of authenticity, and therefore confirmed hypothesis H3.

Reliability Statistics

Cronbach's Alpha	N of Items
.614	4

Table 1. Reliability statistics of the Authenticity scale after removing the item of Authorisation.

As mentioned previously, the item-total statistics for the four-item scale showed moderate inter-item correlations, indicating that the items capture distinct yet complementary facets of authenticity. Although some inter-item correlations remained low, this is not uncommon in exploratory research. This outcome provides a basis for refinement in future research.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
AUTH-1 [Truth|Decept]	10.78	4.360	.273	.626
AUTH-2 [Consist|Unreli]	10.74	3.474	.427	.519
AUTH-3 [Coher|Confus]	10.13	4.119	.484	.501
AUTH-4 [Auth|Inauth]	11.26	3.383	.429	.519

Table 2. Item-Total Statistics of the four items constructing the final Authenticity scale.

7.4 Human-Likeness Facets and Multicollinearity Analysis

The human-likeness facets; social presence (SP), anthropomorphism (ANT), and likeability (LIK), were assessed for correlation and multicollinearity to evaluate their

relationship, and the potential presence of an underlying latent factor, addressing RQ1 and RQ2, respectively.

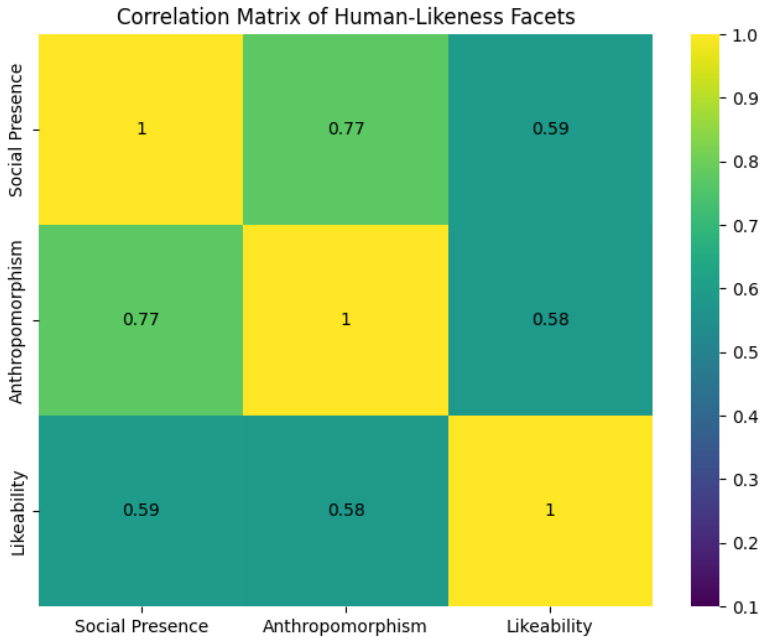

Figure B. Heatmap of Correlation Matrix for Human-Likeness facets where colour intensity indicates the strength and direction of correlations.

The correlation matrix for the human-likeness facets (Figure B) provides insight into the interrelationships among SP, ANT, and LIK. The Spearman correlation values between these facets are as follows: SP and ANT ($r = 0.77$), SP and LIK ($r = 0.59$), and ANT and LIK ($r = 0.58$). These moderate-to-strong correlations suggest a meaningful degree of overlap among the facets, indicating they are interrelated components within the human-likeness construct. This result aligns with hypothesis H1, which anticipated high positive correlations among the facets.

The scatterplots presented (Figures C, D, E) further display relationships between the three human-likeness facets. Figure C reveals a strong, positive linear relationship between SP and ANT. The trend line closely aligns with the data points, indicating a stronger and more consistent association compared to the relationship between other pairs of facets, as seen below (Figure D and Figure E). This linearity suggests that as perceptions of social presence increase, anthropomorphic qualities in the AI are perceived more consistently, supporting the hypothesis of interconnectedness among these variables.

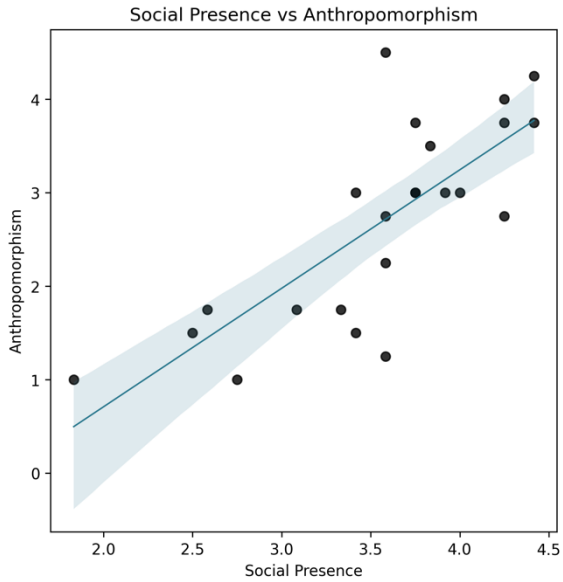

Figure C. Scatter plot of Social Presence and Anthropomorphism, with a trendline.

Figure D, displaying the variables of SP and LIK also shows a positive correlation, though with more variation around the trend line. The spread of data points suggests that while higher social presence generally corresponds with increased likeability, individual responses vary more compared to the SP-ANT relationship. This pattern indicates that factors other than social presence alone may influence the likeability of artificial agents.

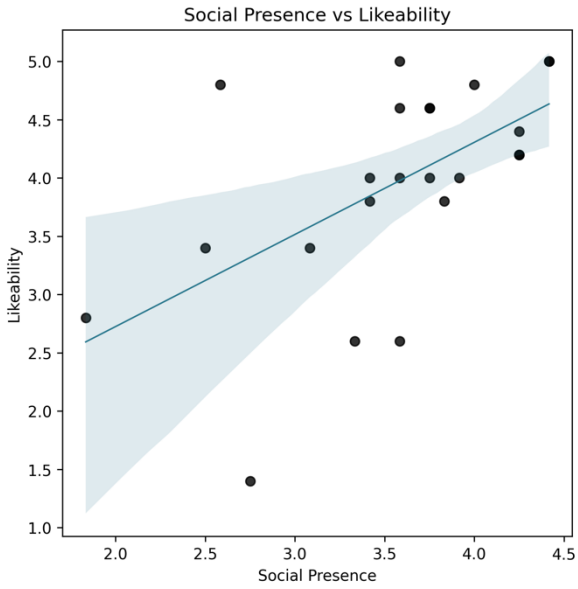

Figure D. Scatter plot of Social Presence and Likeability, with a trendline.

Within Figure E, we can see that the relationship between ANT and LIK follows a positive trend similar to the previous two plots, though with a moderate amount of variability. While higher anthropomorphism is generally associated with higher likeability, the dispersion of points around the trend line indicates that factors of individual differences may play a role in how anthropomorphic qualities impact likeability.

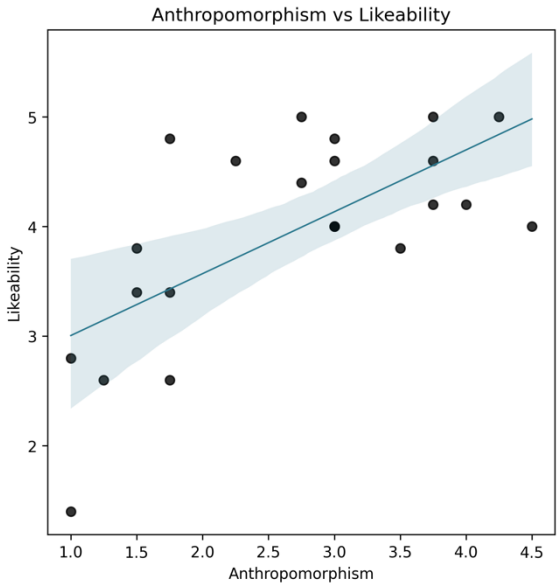

Figure E. Scatter plot of Anthropomorphism and Likeability, with a trendline.

The patterns shown in these scatterplots further justify examining each facet individually while recognising their collective contribution to the human-likeness construct, therefore a multicollinearity analysis was performed. When cross-tested for multicollinearity, the human-likeness facets showed results below a threshold, supporting the idea that each facet contributes uniquely to the overall concept of human-likeness, without redundancy. Using Variance Inflation Factor (VIF), Tolerance, and Condition Index, it was found that the results indicated no significant multicollinearity, with VIF values (ranging from 1.46 to 2.48), tolerance values ranging from 0.40 to 0.68, and the Condition Index having a maximum value of 18.92. The lack of multicollinearity is explored further by visualisation, using Partial dependence plots (PDP), which show non-linear surface patterns that visualise the unique impacts of each facet on the perceived human-likeness of artificial agents.

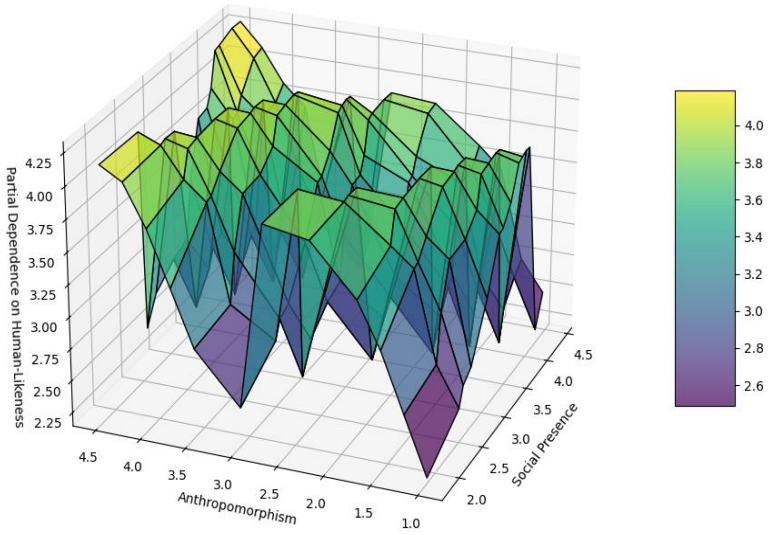

Figure F. 3D Partial Dependence Plot for Anthropomorphism and Social Presence on Human-Likeness

Figure F (ANT and SP) reveals a highly variable surface with numerous peaks and troughs, indicating a fluctuating relationship between these two facets in contributing to human-likeness. Rather than a smooth gradient, this plot shows that as social presence and anthropomorphism increase, their combined influence on human-likeness is not uniform, instead marked by significant ups and downs. This suggests a complex interplay where shifts in either facet can lead to varying levels of perceived human-likeness, highlighting the distinct and dynamic nature of each dimension's contribution.

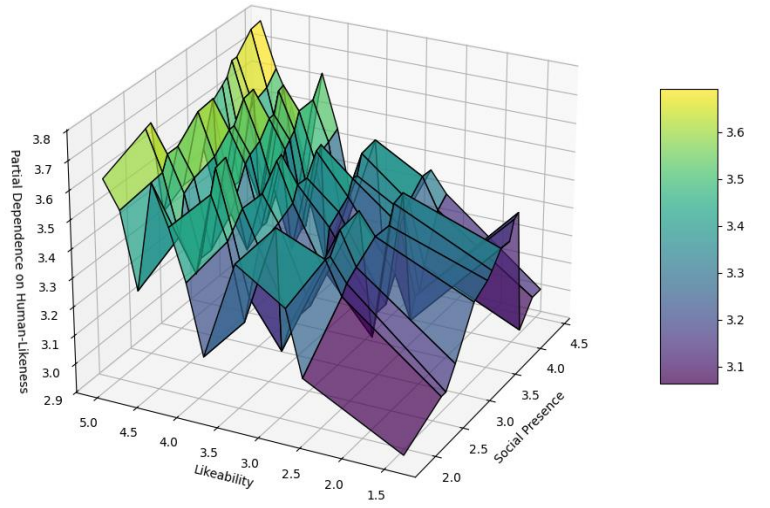

Figure G. 3D Partial Dependence Plot for Likeability and Social Presence on Human-Likeness

Figure G (LIK and SP) reveals more variability in its surface. While higher values of social presence and likeability are generally associated with increased human-likeness scores, the fluctuations in the mid-ranges indicate a nuanced relationship. This variation implies that while both facets positively influence human-likeness, they do so with slightly different strengths across various levels, which reflects their unique contributions to the construct of human-likeness.

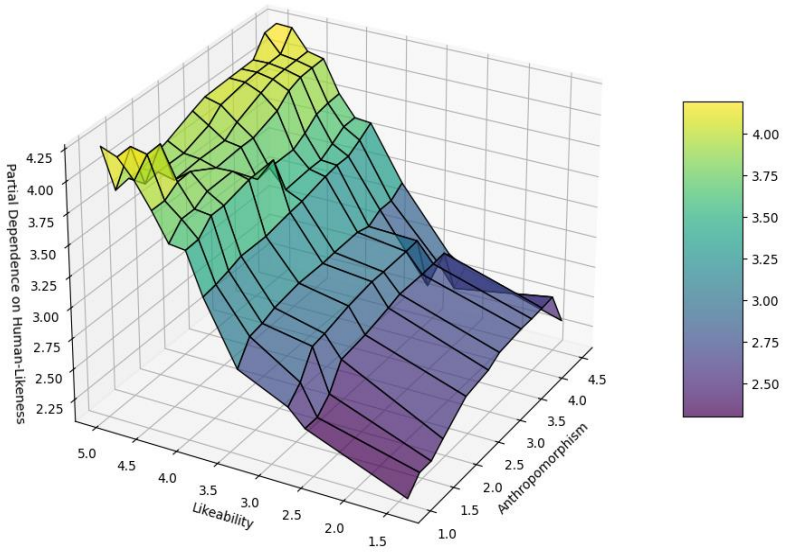

Figure H. 3D Partial Dependence Plot for Anthropomorphism and Likeability on Human-Likeness

Figure H (ANT and LIK) shows a consistent, positive relationship between these two facets and human-likeness, with a smooth incline throughout the surface. This suggests that anthropomorphism and likeability may work in a complementary manner, offering a stable influence on human-likeness without the variability observed in the other plots. The stability here indicates that these two facets interact predictably, supporting the notion that they consistently enhance the perception of human-likeness when combined.

By incorporating PDPs, this study demonstrates a nuanced view of the relationships within human-likeness facets, while validating the inclusion of all three related, yet not overlapping facets, into the framework.

To ensure transparency and facilitate further research, the codes used for creating PDPs, generated by [31], are provided in the appendices.

7.5 Authenticity and Human-Likeness Correlation

The descriptive statistics for the human-likeness measure indicated the mean score of m = 3.40 with a standard deviation of 0.78. For the authenticity measure, the mean was m = 3.58, and the standard deviation of 0.61.

To address RQ4, Spearman's rho correlation analysis was conducted between the aggregated Human-Likeness (HL) score and perceived Authenticity (AUT). The correlation yielded a positive and significant relationship ($\rho = 0.673$, $p < .01$), supporting H4. As suggested by the correlation analysis, the scatter plot (Figure I) representing HL and AUT reveals a positive relationship, indicating that as the perceived human-likeness of an artificial agent increases, its perceived authenticity also tends to rise, as embodied by the upward trend in the line of best fit. The distribution of data points around the trend line suggests a moderate correlation; while higher human-likeness is generally associated with higher authenticity, some variability remains, indicating that unspecified factors might also play a role in shaping authenticity perceptions.

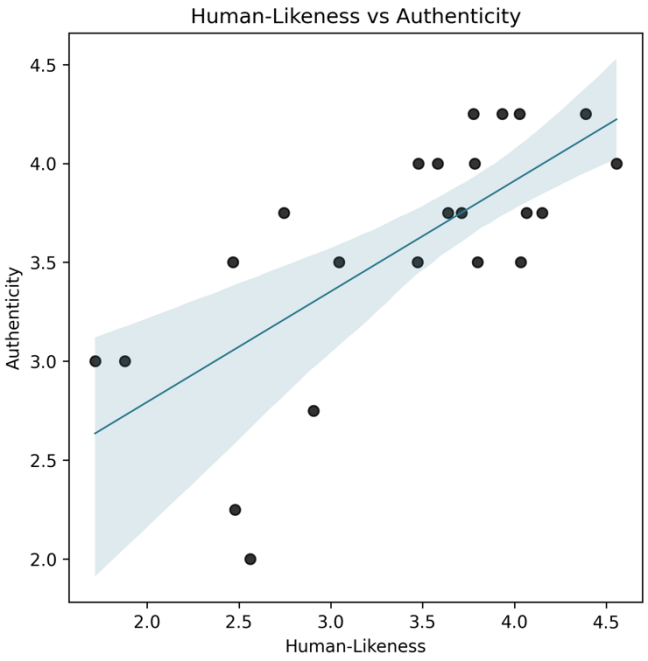

Figure I. Scatter plot of Human-Likeness and Authenticity, with a trendline

The additional correlations between AUT and individual HL facets indicated moderate positive relationships: AUT and SPQ ($\rho = 0.540$), AUT and ANT ($\rho = 0.651$), and AUT

and LIK ($\rho = 0.698$), all at $p < 0.01$. Although on the level of individual variables, these correlations support the idea that individual facets of human-likeness relate positively to authenticity.

7.6 Expertise and Attitudes

Addressing RQ5a and RQ5b, a t-test was performed to examine the influence of expertise on perceptions of human-likeness and authenticity. The results of Levene's Test for Equality of Variances indicated $p > 0.05$, allowing us to assume equal variances. However, the t-test results were not significant for human-likeness ($t(21) = 0.870$, $p = 0.394$) nor authenticity perceptions ($t(21) = 0.843$, $p = 0.409$), suggesting no significant difference between experts and non-experts, and thus refuting both H5a and H5b.

Given these findings, a moderation analysis was pursued post-hoc to investigate if expertise would moderate the relationship between human-likeness and authenticity, as proposed in RQ6. While the moderation model was significant ($R^2 = 0.54$, $F(3, 19) = 7.398$, $p < .002$), the *interaction term* did not yield significance ($B = 0.297$, $SE = 0.26$, $t = 1.12$, $p = .276$), indicating that expertise did not significantly alter the strength or direction of the relationship between HL and AUT. Naturally, these findings refuted the hypothesis H6, however the potential effect of a small sample size on the results is discussed in *Limitations.*

For the *GAAIS,* the descriptive statistics and reliability were assessed separately for the positive and negative subscales. The positive subscale of the GAAIS yielded a mean of 3.70, a standard deviation of 0.55, and a Cronbach's alpha of 0.791, indicating moderate to strong internal consistency. Similarly, the negative subscale had a mean of 3.13, a standard deviation of 0.60, and a Cronbach's alpha of 0.673, which suggests moderate reliability. For the *ATI scale*, the analysis revealed a mean score of 4.18, a standard deviation of 0.99, and a Cronbach's alpha of 0.876, reflecting strong reliability. These results support the internal consistency of the ATI scale and its utility in capturing participants' attitudes toward innovation.

Examining RQ7a, RQ7b, and RQ8, correlation and moderation analyses were conducted to explore attitudes' role in perceptions of human-likeness and authenticity. Spearman's rho correlations between attitudes toward AI and HL showed no significance (ATI: $\rho = 0.012$, $p = 0.957$; GAAIS-p: $\rho = 0.198$, $p = 0.366$; GAAIS-n: $\rho = 0.116$, $p = 0.599$), as

did the correlations between attitudes and AUT (ATI: ρ = -0.106, p = 0.631; GAAIS-p: ρ = -0.038, p = 0.864 ; GAAIS-n: ρ = 0.075, p = 0.733). Subsequently, moderation analysis also showed that attitudes did not moderate the relationship between human-likeness and authenticity, as reflected by an insignificant interaction effect with none of the three constructs measuring attitudes; ATI (B = 0.15, SE = 0.099, t = 1.54, p = .141), GAAIS positive (B = 0.01, SE = 0.22, t = 0.06, p = .956), GAAIS negative (B = 0.25, SE = 0.26, t = 0.96, p = .351). Following the results of correlation and moderation analyses, hypotheses H7a, H7b, and lastly H8 were all rejected.

8. Discussion

Informed by my background in psychology and cognitive science and building on the ideas from the entanglement theory and the 4E cognition approach, this discussion is guided by the assumption of an emerging hybrid future where humans and agents will be increasingly entangled. The work mentioned throughout this study, and the thesis as a whole, suggest that authentic interactions do not necessarily require replicating human-like features in artificial agents. Instead, the thesis points out a priority of designing interactions that foster a sense of interdependence and shared agency between humans and agents, recognising that technology is an actor in shaping our experiences.

The implications which will follow also draw upon broader ideas of theories in social psychology, such as theory of similarity and attraction [112], attachment theory [116], and the theory of cognitive dissonance [12], to understand the social and emotional dimensions of human interaction, and how these might translate to human-agent interaction. These theories, mentioned in earlier sections, provide a deeper understanding of how users might perceive and relate to artificial agents, particularly in terms of similarity, attachment, and potential conflicts between their beliefs and experiences with AI. For example, the strong correlation between social presence and anthropomorphism observed in the study aligns with [112] and his theory, suggesting that users might be more likely to perceive agents as authentic if they find them similar to themselves.

8.1 Resolving Research Questions and Hypotheses

The results of the study provided confirmation for Hypothesis 1 (H1), which posited that social presence (SP), anthropomorphism (ANT), and perceived empathy, represented by likeability (LIK), would show high positive correlations with one another. Data analysis revealed strong positive correlations between SP and ANT, alongside moderate

correlations between SP-LIK and ANT-LIK, indicating that these factors are significantly interconnected in influencing perceptions of human-likeness in artificial conversational agents. The strong correlation between SP and ANT suggests that when users experience a sense of social presence with an agent, they are more likely to attribute human-like characteristics to it [29]. Additionally, the moderate correlation between SP-LIK and ANT-LIK implies that social presence and anthropomorphism may contribute to the perceived likeability of an agent, potentially leading users to view the agent as empathetic. However, likeability was used as a proxy for empathy in this study, although the limitations arising from such an approach are implied in existing operationalisations and findings that followed [45]. Therefore, while plausible, interpreting likeability scores as indicators of perceived empathy requires further acknowledgment of their differences. Empathy encompasses the understanding and sharing of another's feelings, involving both cognitive awareness and affective responsiveness [51, 3]. Perceived empathy, however, focuses on how empathetic the human perceives an agent to be, rather than focusing on the actual empathetic understanding or capacity, which humans can assess through interaction, based on agent's behavioural manifestations. Likeability, however, is built on the perception of appropriateness, and contained to the realm of favourability and pleasantness. In the context of artificial agents, the literature addresses these concepts separately, stressing the complexity of the construct of empathy. For instance, [83] work on social intelligence in chatbots underscores empathy's behavioural aspect without necessarily equating it with general likeability. Meta-analysis from [126] on anthropomorphism further describes likeability as the "extent to which a robot gives positive first impressions". This research does suggest that while both likeability and perceived empathy are grounded in positively valanced observable behaviours, as well as give some support to the idea of an overlap between the two constructs, there seems to be an apparent distinction between the two concepts which does not warrant using one two operationalise another. Measuring likeability as a way of describing perceived empathy can lead to oversimplification of a construct which is much more formative and complex. Putting aside the aforementioned relationship, the findings also align with considerable research highlighting the significance of social presence and anthropomorphism in shaping user perceptions of artificial agents. Ideas such as the Computers-Are-Social-Actors (CASA) paradigm [36] posit that individuals apply human social norms when interacting with computer agents, emphasising the critical role of social presence and anthropomorphism in influencing user perceptions. In HRI, social presence is the extent

to which machines, such as robots, create the impression of being with another social entity [118]. This sense of social embodiment, where robots are designed to convey communication and emotional capabilities, could help foster perceptions of authenticity by contributing to perceived social competence. These insights suggest that factors such as social presence, anthropomorphism, and likeability are essential for creating an experience of human-likeness, which has significant implications for the design [88] of artificial agents to support user engagement and foster authentic interactions.

Hypothesis 2 (H2), which anticipated significant multicollinearity among SP, ANT, and LIK, was not confirmed, suggesting that these facets are distinct yet complementary in shaping human-likeness. This outcome underscores the importance of considering each of these facets separately, as their independent contributions to human-likeness reveal the limitations of unidimensional anthropomorphism measures. A unidimensional measure risks oversimplifying these constructs, as it may not capture social presence or likeability, both of which significantly affect user experience. The findings thus suggest the need for multidimensional assessment tools [42] to measure human-likeness in artificial agents, allowing for an accurate understanding of the distinct roles played by social presence, anthropomorphism, and likeability.

The study's findings also necessitate a reconsideration of the uncanny valley concept. Traditionally, the uncanny valley effect is associated with agents whose near-human appearance can cause discomfort. However, acknowledging likeability as a distinct contributor to human-likeness offers a potential mitigation strategy, where likeable agents might avoid negative reactions even at high levels of anthropomorphism. This insight suggests that likeability may serve as a valuable design consideration, potentially counteracting user discomfort in instances of high, but unease-inducing human-likeness.

The findings on Hypothesis 3 (H3), which explored the reliability of a new semantic differential scale for authenticity, reveal challenges in measuring this complex construct with a single scale. The initially low reliability score of 0.376 improved to 0.614 following the exclusion of the "Authorized/Unauthorized" item, indicating that this item may not align well with other facets of authenticity, such as Truthfulness, Consistency, Coherency, and Authenticity. This points to the challenges of encompassing all aspects of authenticity within a single measure and emphasises the context-specificity of authenticity. As the Pragmatic Authenticity framework posits, authenticity may vary depending on the specific interaction, agent type, and user characteristics, especially as the term itself can be seen as abstract, personal, and very malleable [110]. These insights

raise questions about the applicability of a universal authenticity scale and highlight the potential need for context-specific adjustments, or a universal scale, linguistically broad enough to encompass the abstractness of authenticity.

The literature extensively discusses the impact of context on perceptions of anthropomorphism and trust, in human-agent interactions, highlighting factors such as task type and user characteristics [126, 55]. Anthropomorphism itself is influenced by the perceived fit between a robot's human-like qualities and the demands of a specific task. According to Task-Technology Fit theory [111], if a robot's human-like features enhance its suitability for a given task, users are more likely to attribute human characteristics to it [126]. For instance, socially assistive robots benefit from anthropomorphic features that may enable emotional connection, fostering higher levels of perceived anthropomorphism in social interactions. Individual user characteristics, such as personality traits, past experiences with robots, and cultural backgrounds, also shape anthropomorphisation tendencies. For example, individuals with higher computer anxiety may anthropomorphise robots as a coping strategy to alleviate perceived uncertainty [126]. Such variability underscores the importance of a context-specific approach to AI design, suggesting that universal anthropomorphism designs may not be optimal for diverse user populations [96, 42].

Hypothesis 4 (H4), which hypothesised a positive correlation between an agent's human-likeness and perceived authenticity, was confirmed, supporting the notion that as agents become more human-like, their perceived authenticity may increase. However, because of the formative structure which was discovered within *human-likeness*, it seems important to individually assess the different aspects of said construct, in order to attain a balance between human-likeness and authenticity. As [55] suggests, excessive realism without contextual alignment could result in negative user reactions, and possibly matching facets of *human-likeness* with some other feature but *anthropomorphism* could be an interesting query.

Hypotheses 5a, 5b, 6, 7a, 7b, and 8 examined the roles of user expertise and attitudes towards AI in shaping perceptions of human-likeness and authenticity but did not find significant effects. These findings challenge the assumption that user characteristics like expertise or pre-existing attitudes directly influence perceptions of agents' human-likeness or authenticity. Rather, it appears that users' direct interactions with agents may play a more decisive role in shaping these perceptions, which supports the idea of a

focused interaction-centred design. However, these findings could be greatly affected by the limited sample size and choice of participants, as discussed in *Limitations*.

In summary, these findings confirm that social presence, anthropomorphism, and likeability independently contribute to human-likeness, supporting the need for multidimensional approaches in assessing and designing artificial agents. Additionally, while human-likeness contributes to perceived authenticity, the role of likeability as a possible moderator for user discomfort in cases of high anthropomorphism, challenges the traditional notion of the uncanny valley. Instead, likeability appears as a potentially critical design element that can help mitigate the uncanny valley effect when anthropomorphism reaches high levels. Lastly, the refutation of hypotheses related to expertise and attitudes towards AI (H5 through H8) underscores the complex nature of user-agent interactions, where users seem to place greater emphasis on their direct experience with the agent rather than on their preconceived expertise or attitudes. However, the lack of significant findings within the latter questions should be followed by the mention of a possible interference of a small sample size, as well as possibly smaller margins for grouping variables (based attitudes and expertise), than the ones that could be found in a *normal* population.

8.2 Implications

8.2.1 Individual Contribution of Human-Likeness Facets

By identifying that there is no significant multicollinearity among facets of social presence, anthropomorphism, and likeability, the findings underscore the need to evaluate each facet independently rather than relying on one to represent the total perception of human-likeness. In statistical terms, multicollinearity would imply that the independent variables of social presence, anthropomorphism, and likeability, are so highly correlated that it would be difficult to assess their unique effects on perceived human-likeness, which is the dependent variable in this context. The lack of significant multicollinearity, however, suggests that each facet offers unique information and contributes separately to the construct of human-likeness. This implies that a conversational agent's human-likeness cannot be achieved through a singular focus on one dimension, while neglecting others. This is supported by [83] who explored factors of "human nature" and ""uniquely human" facts, which both contained human-like facets, but have shown to elicit anthropomorphic perceptions to a different degree.

The implications of these findings are substantial for both the design and evaluation of conversational agents. Designers are urged to adopt a balanced approach that assesses each facet individually, ensuring a holistic perception of human-likeness that meets the desired interaction goals. Neglecting any of the facets, risks undermining the (assumed) goal of human-likeness. Current research supports the idea that focusing exclusively on anthropomorphic design does not suffice in achieving perceived authenticity in HAI. Rather, *(what we call)* authenticity arises from a blend of factors, embedded into the agent through design cues, met with the traits of the user, and situated in the interaction context [132]. Moreover, ethical implications arise from striving for overly human-like agents, where users may develop inappropriate attachments or misplaced trust if the agent's apparent human-likeness surpasses its actual capabilities [91, 18]. A nuanced approach should, therefore, be beneficial not only for developing likeable human-like agents, but also consider ethical dilemmas associated with creating convincing yet fundamentally artificial entities.

8.2.2 Designing for Likeability

In human-agent interaction, designing for likeability opens new possibilities for developing authentic relationships with agents by shifting away from an overemphasis on anthropomorphism, grounded in personalisation and context-specific design. Popular cases [92] have demonstrated the power of (and interest in) personalisation, such as customisation of ChatGPT to exhibit specific *boyfriend-like* traits. In this case, the user *customised* the agent's behaviour to reflect a "masculine personality", characterised by distinct voice features, proactive engagement, teasing and flirtatious humour, and use of nicknames. Such deliberate tailoring fostered a more enjoyable, personal connection between the user and the agent, showcasing how personalisation can drive engagement and likeability.

To scale this approach and make it more accessible to those less tech-savvy but still interested in interactions of this nature, designers could introduce *personas*, distinct agent archetypes with unique voices, traits, and personalities, something which user experience research has been familiar with for decades. Early iterations, such as OpenAI's introduction of different voices for ChatGPT, demonstrate how basic persona variations can already enhance user engagement. Building on this, future iterations could assign these personas more complex personalities grounded in psychological frameworks. By drawing from established personality models, such as the Big Five or similar typologies,

designers could create personas with traits calibrated to appeal to different user preferences. Developed from extensive usability testing, these archetypes would get refined, condensing them into a manageable set (five to ten personas) that are widely accepted and well-liked. Users would be able to choose their agent's persona based on the descriptions of their assigned traits.

The next step in this evolution could involve allowing users to mix and match traits to create semi-customised personas. For example, users could choose traits such as the agent's extroversion level, humour style, or preferred topics of interest. These traits could then be paired with specific attitudes, values, and beliefs designed to match the persona's personality. It is important to highlight that the users would be choosing from a list of pre-approved traits, such as *optimism* and *love for music*, without inputting their own values directly, ensuring the agent's adherence to strict ethical guidelines, and avoiding controversial or harmful configurations. That way, agents would also avoid engaging in explicit or inappropriate discourse, regardless of the selected persona.

An interesting idea regarding ethical concerns of deception would be introducing limitations to each persona; while *our* persona might have many things we like about it, even prefer, it could be limited to only a number of (positive) traits, meaning a switch to an uninteresting, task-specific agent would be the only way to get something done. Another way would be highlighting the limitations of a persona; if they are a good *mathematician*, they might not be able to help you with your interest in *agriculture*. This way, humanness is normalised within artificial agents, making them appear transparent about their capabilities and not all-knowing, insinuating artificial agents do not have all the answers.

This approach to designing for likeability can also be considered as designing for authenticity. Authenticity of human interactions is based on personal and situational cues, which can change based on numerous factors. By choosing the agent's persona which is likeable to us, then tailoring it further to our specific preferences, all the while introducing transparency and limited capabilities as *positive* traits, could lead to higher acceptance, engagement, and normalisation of agents and companions, developing authentic human-agent relationships.

8.2.3 Reframing Empathy

Staying on the topic of relationships, how we perceive and evaluate others, particularly our development of affection toward them, is heavily influenced by our interactions. This

applies to human-agent interactions, as much as human-human ones. According to [113], our judgements about another's empathy are based on what we can observe in their behaviour, rather than having direct access to their genuine internal state. For example, we might infer someone is *empathetic* if they express their concern when we are upset or offer helpful advice in case of a challenge. [113] further highlights this reliance on perceived empathy in discussing the development of the Empathy Quotient, a self-report questionnaire designed to measure an individual's perception of their own empathetic traits. This further supports the notion that even our own (human) understanding of empathy is shaped by observations and reflections on our behaviour.

[43] extend this concept to human-robot interaction, arguing that the success of robot companions relies on their ability to exhibit relevant behaviour, regardless of the presence, or absence, of genuine emotions. They illustrate this point with a thought experiment about caregiving, ultimately suggesting that the care receiver's perception of the care is crucially determined by the caregiver's actions, rather than an internal emotional state.

With the understanding that human interactions are based in the evaluation of behavioural cues of empathy, it seems quite unnecessary to be striving for the development of agents with embedded ontological empathy. Focusing on enabling agents to demonstrate appropriate empathetic behaviours, however, regardless of what exists within their black box, seems to be necessary for successful human-agent interaction.

Feeling, or *exhibiting* appropriate emotional reactions is one of the definitions [113] has offered for empathy, highlighting here the importance of the appropriateness of a reaction. This definition highlights that empathy is not limited to expressing only positive emotions, such as compassion or joy, rather it necessitates reacting in a way that aligns with the emotional context of the interaction. For instance, if someone expresses anger or frustration, an empathetic response would acknowledge and validate their anger, rather than offering generic words of comfort. This failure to match the emotional tone of the interaction can lead to the perception of insensitivity or a lack of understanding, which is a concept also relevant in human-agent interactions. To achieve truly human-like interactions, agents need to portray empathetic behaviour across a diverse range of contexts, including learning how to manage those involving negative emotions. The ability to tailor these responses might be crucial for fostering user comfort, trust, and a perception of authenticity. This dynamic, personalised design could be explored through the interplay of empathy components. In his work, [113] defines three concepts: cognitive

empathy, affective empathy, and sympathy, all of which will be relevant for this brief proposal. According to [113], affective empathy involves experiencing an emotional response that mirrors or is appropriate to the emotional state of another person, while cognitive empathy refers to the ability to understand another person's perspective and mental state, without necessarily feeling the emotions. In other words, [113] discusses an aspect of empathy which is affective and deeply embodied, and another which has also been compared to the Theory of Mind, also known to be interestingly embodied and uniquely human. In addition to these, [113] introduces the concept of sympathy, which he defines as a subset of affective empathy, occurring when our emotional response to another's distress motivates us to take action.

To create truly compelling agents, we do not need to replicate, however we must equip agents with the ability to exhibit a balance of these types of empathy. The operationalisation of these steps might include recognising and responding to the user's intentions and emotional cues, accurately interpreting the user's intentions and perspective, and taking appropriate actions to address the user's needs or concerns. Some of these concepts are already a design practice, implemented in current LLMs, for example ChatGPT's ability to verbalise (respond to) user's textual and verbal emotional cues. It was also discovered by [54] that ChatGPT, while far from possessing it, does portray behaviours resembling that of the Theory of Mind, which might have emerged as a result of iterations and interactions. However, it is essential to recognise that the ideal balance of these empathetic behaviours, design concepts, will vary depending on the individual user and the specific context of the interaction. What one user finds comforting, another might perceive and intrusive. This is because people have different expectations and preferences when it comes to interpersonal relationships, particularly those involving emotional support and understanding. Therefore, accessible personalisation of empathy will play a crucial role in tailoring the agent's empathetic behaviour to meet the unique needs of each user. This aligns with the initial emphasis on the importance of perceived empathy, and to the core idea that the design of empathetic agents should prioritise creating agents that behave empathetically, and are treated as dynamic systems, made to adapt to the environment in which they are operating. This approach goes beyond the *one-size-fits-all* or the *designing-for-all*, and rather emphasises truly personalised aspects of those traits which deem it necessary.

9. Conclusion

9.1 Limitations

This study's exploration of human-likeness and authenticity in AI, although valid in its contributions, encounters a series of limitations that highlight areas for methodological refinement and potential adaptations. These limitations, assessed through the criteria of rigour proposed by [40] and expanded by [1], offer a lens into the study's strengths and constraints. Within this framework, the limitations encountered in the study particularly underscore constraints in the criteria regarding *data abundance, richness, and resonance*. The exclusive use of English as the study language, despite the participants' diverse linguistic backgrounds, significantly impacts the abundance of data. With many participants being native German and Croatian speakers, using semantic differential scales exclusively in English might have influenced the accuracy of their responses. Semantic differential scales require a truly nuanced comprehension, and responses in a non-native language risk the loss of subtle distinctions in understanding and meaning, potentially limiting the collected data, thereby reducing the representational validity of the findings.

A further limitation relates to the study's small sample size (n=23), which has also constrained data abundance and led to the use of non-parametric analyses. Although the sample size sufficed for a preliminary pilot study, it limited the study's capacity to generalise findings and the statistical power necessary to detect more nuanced effects. Larger sample sizes are integral to conducting parametric analyses, which could reveal interaction effects among facets of human-likeness that are undetectable in smaller samples.

Additionally, the targeted recruitment strategy focused on students and experts within technical and technology-oriented interdisciplinary fields, which further constrained the abundance and representativeness of the data. Although intended to differentiate between expertise levels, both participant groups had familiarity with digital technologies, potentially blurring distinctions between expert and non-expert perceptions. This limits the study's generalisability to broader populations, where attitudes toward AI may vary significantly from those of technology-savvy individuals. The limited diversity in the participant sample may thus reduce the resonance of findings, as perceptions of AI tend to be influenced by varying levels of exposure and familiarity with technology.

The absence of familiarity with technology as a measured variable further limits the study's ability to capture the interplay of non-interaction-based factors influencing perceptions of authenticity. Technological familiarity, especially as it varies through generations, is likely to influence how expertise shapes authenticity perceptions in AI. While such an omission is understandable in a pilot context, it restricts the study's potential to provide an informed perspective on the interplay of expertise, attitudes, and familiarity with technology.

Moreover, the study's reliance on self-reported data introduces limitations in plausibility, as self-reports are susceptible to various biases, including social desirability bias and recall bias. While self-reporting provides valuable insights into subjective user perceptions, it inherently involves participants' self-assessment, which may not always align with actual behaviour or perceptions. Social desirability bias, for instance, might prompt respondents to answer in ways they perceive as more favourable rather than truthful. Similarly, recall bias may impact the accuracy of participants' recollections, particularly in assessing prior experiences with AI. While objective measures such as physiological responses could be incorporated, I believe the value of it would be insignificant and future research should rather focus on expanding the scope of subjective measures through incorporating qualitative data. Considering that, a final limitation regards the limited granularity of data which is derived from the sole use of quantitative scales, without incorporating open-ended or interview-based questions. While structured scales offer a systematic approach to quantifying user responses, they may not capture the richness of individual experiences that could emerge through qualitative briefs. This restriction affects the resonance of the study's findings, as it limits the contextual understanding.

In conclusion, while this study provides valuable initial insights into the complex constructs of human-likeness and authenticity in AI, these limitations underscore the importance of methodological rigour and the need for careful consideration of language, sample diversity, variable inclusion, data collection methods, and subjective biases. Each limitation highlights an area for refinement, suggesting that future research on authenticity in human-agent interaction would benefit from larger, more diverse samples, and the integration of qualitative inquiries.

9.2 Assessment

[40] and [1] offer criteria for assessing the rigour of research, particularly within design-oriented fields like human-computer interaction and visualisation research. While their frameworks were developed in different contexts, both allow for the application of their criteria to evaluating both the Pragmatic Authenticity (PA) framework and the empirical study on human-likeness in AI interactions. [40] criteria, rooted in an interpretivist perspective, emphasise the subjective and socially constructed nature of knowledge gained through design studies. [1] expands on by [40] criteria by incorporating ethical considerations and highlighting the relevance of rigour in the evolving field of human-robot interaction. However, applying these general criteria to disembodied systems such as conversational agents requires certain adaptations. Disembodied agents lack physical presence, which is often a focal point in evaluations of human-likeness and authenticity within human-robot interaction. Therefore, embodiment-specific criteria traditionally used to assess physical appearance, or gestures must be reinterpreted to address elements that are to disembodied agents, such as language use, and the conveyed personality through text or voice alone. This will be apparent through assessment of rigour by emphasising the quality and authenticity of language and interaction style (i.e. social embodiment), as these are replacements for physical embodiment. With these adaptations in mind, the PA framework and empirical study are evaluated according to by [1] adapted version of the criteria, focusing on six main areas: *Informed, Reflexive, Abundant, Plausible, Resonant, and Transparent*.

Both the PA framework and the empirical study demonstrate a strong foundation in existing knowledge, backing up the criteria for informed research and design. The framework draws upon established concepts of authenticity from fields including philosophy, psychology, and human-computer interaction, providing a comprehensive theoretical foundation. For example, the framework's structure, informed by [110] on subordinate concepts of authenticity, provides a strong basis for its components. In addition, the study's reliance on established measurement tools, such as the Godspeed and Social Presence Questionnaires, bolsters the reliability and validity of the data collected. Nonetheless, a limiting quality of such a framework is tied to the open-ended nature of developing it; of failing to include all potential factors which could affect human perception of authenticity. Proposing such a framework is always open to critique, and therefore, further development, since the introduction of new factors could completely dismantle the current structure of the framework and the intercorrelations of its facets.

Reflexivity is addressed by explicitly acknowledging the researcher's positionality in the discussion section, which outlines how their background and values may have influenced the research process. This transparency in positionality enhances the overall reflexivity and invites readers to consider the researcher's perspective when interpreting the findings. Furthermore, the study acknowledges the limitations of using likeability as a proxy for empathy, identifying the need for future research to explore this relationship in greater depth.

Regarding abundance, although the study incorporates various measures and data sources, concerns regarding the limited sample size are valid. The limitations subsection recognises that the study did not meet its initial target of 30 participants, which would have allowed for parametric statistical analysis. Due to the smaller sample size, non-parametric tests with potentially reduced statistical power were used. Despite this, the employment of multiple measurement tools does contribute to the abundance of data collected, offering a multifaceted perspective on the constructs under study.

The findings of the study are plausible, supported by robust data analysis. The confirmation of Hypothesis 1, which demonstrates a strong correlation between social presence, anthropomorphism, and likeability, aligns with existing literature on the CASA paradigm [36]. Furthermore, the reliability analysis of the authenticity scale is detailed, justifying the decision to exclude the "Authorized/Unauthorized" item, resulting in a refined four-item scale with improved consistency. However, a dedicated study for developing the authenticity scale should be conducted, to explore various aspects of authenticity which might not have been considered in the current scale. At the least, the study should work toward expanding on the introduced aspects, and confirming its internal consistency and validity for general, rather than just a specific use case.

Targeting resonance as a criterion, addressing questions about authenticity in human-agent interaction, the PA framework serves as a practical foundation for designers and engineers aiming to create agents that foster authentic relationships with human users. Notably, the study's findings, particularly regarding the role of likeability in potentially mitigating the uncanny valley effect, suggest that achieving likeability may be more impactful than attaining perfect human-likeness. The emphasis on interaction-centred design aligns with contemporary human-robot interaction trends that prioritise user experience and recognise the evolving, dynamic nature of human-agent interactions.

Transparency is a hallmark of this study, which offers a clear and thorough account of the study's design, data collection procedures, and analytical approach. The positionality

statement in the discussion section enhances this transparency by highlighting the researcher's perspective and potential biases. To facilitate research scrutiny, replication, and further exploration, the study's unique materials, such as novel questionnaires, and social vignettes, and codes for the alternative data visualisation generation, are provided in the appendices.

9.3 Future Work

9.3.1 Developing a Validated Scale for Authenticity

The future directions for advancing the authenticity scale, as indicated by findings in Hypothesis 3, reveal significant challenges and opportunities for improving the measurement of authenticity in human-agent interaction. Specifically, the study highlights the complexity of capturing social and relational dimensions such as trust and authenticity, areas that are foundational to HAI and HRI but inherently challenging to quantify. These findings underscore the pressing need for the development of standardised, reliable, and valid instruments that facilitate meaningful comparisons across studies, thereby advancing the field. Although the study achieved moderate reliability with the revised scale by removing problematic items, further refinement and validation are crucial to establish a robust measure for authenticity in HAI. Future work will need to address nuanced aspects of social connection, which are currently challenging to measure but essential for deepening our understanding of user-agent interactions. This could be explored through a study focusing solely on establishing and validating the semantic differential scale of authenticity for HAI research. Using knowledge from sociolinguistics, cognitive psychology, and HAI, research could aim to include an extensive list of potential relevant and valuable descriptors, narrowing the scale down to its most reliable and usable components. This would create a first step in introducing pragmatic authenticity in HAI, HRI, and potentially HCI, within research and the academic community, before it can become a useful tool in the design process and usability testing in more broad, industry contexts.

9.3.2 Interaction-Centred Design

Interaction-centred design, shifting away from merely static, embedded traits of individual interlocutors (human and artificial), emerges as another critical area for future research. As the empirical findings suggest, neither expertise nor pre-existing attitudes toward AI can definitively, on their own, define users' perceptions of human-likeness or

authenticity, indicating that the perception of agents and authentic interactions emerges through interaction. Therefore, interaction-centred design intends to focus on the interaction between the agent and the human, all within a specific context, because what may appear as friendly in one setting, could be perceived as inappropriate in a more formal domain. This context-dependence of authenticity reflects the need for a dynamic approach to AI design, where the parameters of interaction (agent, human, context), rather than agent's traits alone, are being calibrated to effectively meet user expectations.

9.3.3 Navigating Anthropocentrism

Closely aligned with aforementioned ideas, the sometimes-interchangeable use and role of humanness, human-likeness, and anthropomorphism in design, is being revisited. Research suggests that an overemphasis on anthropomorphism may not always serve as the optimal design pathway, potentially allowing for the idea of authenticity built on *likeability,* emerge as a new design ideal. This concept involves deliberately highlighting the agent's limitations as a machine, rather than attempting to create a perfect replication of human behaviour. When users attribute human-like qualities such as genuine empathy or consciousness to AI, this can foster expectations that, when unmet, result in user dissatisfaction, frustration, or distrust. This effect is particularly pronounced when agents are used as companions for emotional support or personal advice, in which cases user perception could end up exceeding the system's intended functionality, as well as ethical boundaries. By embracing *authenticity* as a goal, rather than *human-likeness*, designers can set more realistic expectations and aim to avoid the "uncanny valley" effect, which often results from agents that are nearly, but not entirely, human-like. User testing becomes an invaluable tool in this context, enabling designers to optimise agents for experiences that are both engaging and appropriate for their intended applications. Within this context, employing *alternative, domain-specific* design options was suggested [55, 50] as a way of inducing trust in users (aiding authenticity), while not having to employ human-likeness, which could be done by using objects (*metaphors*) to serve as agent personas. Examples given in [55] were various; that of a zoomorphic avatar, an encyclopaedia for a health-oriented agent, or a calculator for a financial advisor. All examples were curated to highlight aspects of their *personas* that are task-and-context-relevant, rather than portraying the idea of an AGI-like agent. In the case of an animal, the benefits were supposed in clearly showcasing the distinction of an artificial agent not being a human, indirectly limiting human perception of its capabilities.

This approach interestingly affects the PA framework, as I fully support the idea of pursuing context-specific alternative design, all the while wanting to highlight the importance of *human-likeness* or *likeability* within my framework. What I believe needs clarification, and possible linguistic and philosophical framing, is the operationalisation, and intended use of *human-likeness* within PA. The idea of human-likeness was initially incorporated to be a useful factor supporting concepts such as familiarity, trust, as well as serving as an *ideal*. However, the more in-depth and context-specific one goes into design, its ethical aspects, and nuances for what is necessary, expected and effective, it becomes obvious that human-likeness is *not* the objective. Human-likeness is a proxy for a parameter, to be switched out for any other *persona*, depending on the agent, human, and context. Exploring how this unravels in a practical sense, and how it will be adopted into the PA framework is something to also be considered as *Future Work*.

In conclusion, a call for a shift in AI design paradigms could mean moving from *human-likeness* as the ideal towards which engineers, designers, researchers, and developers should be striving for, and moving towards more nuanced and context-specific *metaphors*. This could be explored through further, rigorous implementations of transparency and explainability, highlighting AI's capabilities. These may be matched with both standardised and personalised design cues enabling likeability in interactions, grounded in concepts such as anthropomorphism and perceived empathy. By broadening the design perspective beyond strict anthropomorphic virtues, both physical and social, designers can work on minimising the risks associated with unrealistic user expectations, ethical dilemmas, and the potentially negative impacts of the uncanny valley.

9.4 Summary

This thesis has introduced a novel theoretical framework for defining, assessing, and designing pragmatic authenticity, putting emphasis on aspects of personalised, subjective experience and ethical development. The goal of the framework is to unify valuable insight from relevant fields, contributing to emergence of trust, responsibility, and comfort in human-agent relationships. An empirical study followed, exploring the initial validity of the proposed framework, specifically its factor of human-likeness. The study focused on the relationship between human-likeness and authenticity in artificial conversational agents, hoping to showcase that achieving authentic interactions requires

designing agents whose perceived human-likeness goes beyond anthropomorphism and mimicry. While the study confirms that social presence, anthropomorphism, and likeability contribute to perceptions of human-likeness, it also underscores the importance of considering these facets independently, highlighting the balance needed to strike when managing various facets of human-likeness and the potential for the uncanny valley effect. Future research should explore alternative design methods beyond mere anthropomorphism when designing for authentic social interactions with agents, as well as look further into establishing a culturally representative semantic differential scale of authenticity.

Acknowledgments

This thesis utilised ChatGPT-4 to generate the social vignette content, a Python code for generating partial dependence plots, and it was used as an interactive agent to conduct the empirical part of the study.
This thesis utilised Grammarly [62] to review grammar and spelling throughout the writing of this thesis.

References

[1] A. Weiss, ‘`One cannot know everything’ — On the Need of Epistemological Diversity in Human-centered HRI Research’, Habilitation treatise, Paris Lodron University of Salzburg, 2021.

[2] J. Rose and M. Jones, ‘2004) "The Double Dance of Agency: A Socio-Theoretic Account of How Machines and Humans Interact’, Jan. 2004.

[3] Özge N. Yalcin and S. DiPaola, ‘A computational model of empathy for interactive agents’, *Biologically Inspired Cognitive Architectures*, vol. 26, pp. 20–25, Oct. 2018, doi: 10.1016/j.bica.2018.07.010.

[4] E. L. Thorndike, ‘A constant error in psychological ratings.’, *Journal of Applied Psychology*, vol. 4, no. 1, pp. 25–29, Mar. 1920, doi: 10.1037/h0071663.

[5] A. Sciutti, ‘A Developmental Perspective on Cognitive Robotics for Interaction’, presented at the MEi:CogSci Conference, Bratislava, Slovakia, Jun. 13, 2024.

[6] I. Bečková, Š. Pócoš, G. Belgiovine, M. Matarese, A. Sciutti, and C. Mazzola, ‘A Multi-Modal Explainability Approach for Human-Aware Robots in Multi-Party Conversation’, May 20, 2024, *arXiv*: arXiv:2407.03340. Accessed: Nov. 01, 2024. [Online]. Available: http://arxiv.org/abs/2407.03340

[7] T. D. Do, R. P. McMahan, and P. J. Wisniewski, ‘A New Uncanny Valley? The Effects of Speech Fidelity and Human Listener Gender on Social Perceptions of a Virtual-Human Speaker’, in *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, Apr. 2022, pp. 1–11. doi: 10.1145/3491102.3517564.

[8] T. Franke, C. Attig, and D. Wessel, ‘A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale’, *International Journal of Human–Computer Interaction*, vol. 35, no. 6, pp. 456–467, Apr. 2019, doi: 10.1080/10447318.2018.1456150.

[9] R. R. Hoffman, P. J. Feltovich, K. M. Ford, and D. D. Woods, ‘A rose by any other name...would probably be given an acronym [cognitive systems engineering]’, *IEEE Intell. Syst.*, vol. 17, no. 4, pp. 72–80, Jul. 2002, doi: 10.1109/MIS.2002.1024755.

[10] T. Fong, I. Nourbakhsh, and K. Dautenhahn, ‘A survey of socially interactive robots’, *Robotics and Autonomous Systems*, vol. 42, no. 3–4, pp. 143–166, Mar. 2003, doi: 10.1016/S0921-8890(02)00372-X.

[11] J. Feine, U. Gnewuch, S. Morana, and A. Maedche, ‘A Taxonomy of Social Cues for Conversational Agents’, *International Journal of Human-Computer Studies*, vol. 132, pp. 138–161, Dec. 2019, doi: 10.1016/j.ijhcs.2019.07.009.

[12] L. Festinger, *A Theory of Cognitive Dissonance*. Redwood City: Stanford University Press, 1957. doi: doi:10.1515/9781503620766.
[13] L. B. McInroy and O. W. J. Beer, 'Adapting vignettes for internet-based research: eliciting realistic responses to the digital milieu', *International Journal of Social Research Methodology*, vol. 25, no. 3, pp. 335–347, May 2022, doi: 10.1080/13645579.2021.1901440.
[14] M. R. Leary, 'Affiliation, Acceptance, and Belonging: The Pursuit of Interpersonal Connection', in *Handbook of Social Psychology*, 1st ed., S. T. Fiske, D. T. Gilbert, and G. Lindzey, Eds., Wiley, 2010. doi: 10.1002/9780470561119.socpsy002024.
[15] V. Kaptelinin and B. Nardi, 'Affordances in HCI: toward a mediated action perspective', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Austin Texas USA: ACM, May 2012, pp. 967–976. doi: 10.1145/2207676.2208541.
[16] K. Yeung, A. Howes, and G. Pogrebna, 'AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing', *SSRN Journal*, 2019, doi: 10.2139/ssrn.3435011.
[17] E. J. De Visser *et al.*, 'Almost human: Anthropomorphism increases trust resilience in cognitive agents.', *Journal of Experimental Psychology: Applied*, vol. 22, no. 3, pp. 331–349, Sep. 2016, doi: 10.1037/xap0000092.
[18] M. Coeckelbergh, 'Are Emotional Robots Deceptive?', *IEEE Trans. Affective Comput.*, vol. 3, no. 4, pp. 388–393, 2012, doi: 10.1109/T-AFFC.2011.29.
[19] G. Papagni, J. De Pagter, S. Zafari, M. Filzmoser, and S. T. Koeszegi, 'Artificial agents' explainability to support trust: considerations on timing and context', *AI & Soc*, vol. 38, no. 2, pp. 947–960, Apr. 2023, doi: 10.1007/s00146-022-01462-7.
[20] D. Vernon, *Artificial cognitive systems: a primer*. Cambridge, Massachusetts: MIT Press, 2014.
[21] S. Russell and P. Norvig, *Artificial Intelligence, Global Edition : A Modern Approach*. Pearson Deutschland, 2021. [Online]. Available: https://elibrary.pearson.de/book/99.150005/9781292401171
[22] M. Heerink, B. Kröse, V. Evers, and B. Wielinga, 'Assessing Acceptance of Assistive Social Agent Technology by Older Adults: the Almere Model', *Int J of Soc Robotics*, vol. 2, no. 4, pp. 361–375, Dec. 2010, doi: 10.1007/s12369-010-0068-5.
[23] S. Turkle, 'Authenticity in the age of digital companions', Interaction Studies - Social Behaviour and Communication in Biological and Artificial Systems, vol. 8, pp. 501-517, Nov. 2007, doi: 10.1075/is.8.3.11tur
[24] C. Heeter, '*Being* There: The Subjective Experience of Presence', *Presence: Teleoperators & Virtual Environments*, vol. 1, no. 2, pp. 262–271, Jan. 1992, doi: 10.1162/pres.1992.1.2.262.
[25] N. Banovic, Z. Yang, A. Ramesh, and A. Liu, 'Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust', *Proc. ACM Hum.-Comput. Interact.*, vol. 7, no. CSCW1, pp. 1–17, Apr. 2023, doi: 10.1145/3579460.
[26] K. Dautenhahn and A. Billard, 'Bringing up robots or—the psychology of socially intelligent robots: from theory to implementation', in *Proceedings of the Third Annual Conference on Autonomous Agents*, in AGENTS '99.

New York, NY, USA: Association for Computing Machinery, 1999, pp. 366–367. doi: 10.1145/301136.301237.
[27] K. M. Lee, N. Park, and H. Song, 'Can a Robot Be Perceived as a Developing Creature?.: Effects of a Robot's Long-Term Cognitive Developments on Its Social Presence and People's Social Responses Toward It', *Human Comm Res*, vol. 31, no. 4, pp. 538–563, Oct. 2005, doi: 10.1111/j.1468-2958.2005.tb00882.x.
[28] K. M. Lee, W. Peng, S.-A. Jin, and C. Yan, 'Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction', *Journal of Communication*, vol. 56, no. 4, pp. 754–772, Dec. 2006, doi: 10.1111/j.1460-2466.2006.00318.x.
[29] K. J. Kim, E. Park, and S. Shyam Sundar, 'Caregiving role in human–robot interaction: A study of the mediating effects of perceived benefit and social presence', *Computers in Human Behavior*, vol. 29, no. 4, pp. 1799–1806, Jul. 2013, doi: 10.1016/j.chb.2013.02.009.
[30] K. S. Haring, K. Watanabe, D. Silvera-Tawil, M. Velonaki, and T. Takahashi, 'Changes in perception of a small humanoid robot', in *2015 6th International Conference on Automation, Robotics and Applications (ICARA)*, Queenstown, New Zealand: IEEE, Feb. 2015, pp. 83–89. doi: 10.1109/ICARA.2015.7081129.
[31] OpenAI, *ChatGPT*. (2024). OpenAI. [Online]. Available: https://chat.openai.com/
[32] L. N. Girouard-Hallam, H. M. Streble, and J. H. Danovitch, 'Children's mental, social, and moral attributions toward a familiar digital voice assistant', *Human Behav and Emerg Tech*, vol. 3, no. 5, pp. 1118–1131, Dec. 2021, doi: 10.1002/hbe2.321.
[33] K. Corti and A. Gillespie, 'Co-constructing intersubjectivity with artificial conversational agents: People are more likely to initiate repairs of misunderstandings with agents represented as human', *Computers in Human Behavior*, vol. 58, pp. 431–442, May 2016, doi: 10.1016/j.chb.2015.12.039.
[34] C. S. Peirce, C. Hartshorne, P. Weiss, and A. W. Burks, *Collected papers of Charles Sanders Peirce. Vol. 5 : Pragmatism and pragmaticism*, [New ed]. Cambridge, Mass.: Belknap Press of Harvard University Press, 1974.
[35] S. Marsella, J. Gratch, and P. Petta, 'Computational Models of Emotion', A Blueprint for Affective Computing-A Sourcebook and Manual, pp. 21-46, Jan. 2010.
[36] C. Nass, J. Steuer, and E. Siminoff, 'Computer are social actors', in *Conference on Human Factors in Computing Systems - Proceedings*, Apr. 1994, p. 204. doi: 10.1145/259963.260288.
[37] R. E. Mayer, W. L. Johnson, E. Shaw, and S. Sandhu, 'Constructing computer-based tutors that are socially sensitive: Politeness in educational software', *International Journal of Human-Computer Studies*, vol. 64, no. 1, pp. 36–42, Jan. 2006, doi: 10.1016/j.ijhcs.2005.07.001.
[38] D. Gefen and D. W. Straub, 'Consumer trust in B2C e-Commerce and the importance of social presence: experiments in e-Products and e-Services', *Omega*, vol. 32, no. 6, pp. 407–424, Dec. 2004, doi: 10.1016/j.omega.2004.01.006.
[39] M. Dubiel, S. Daronnat, and L. A. Leiva, 'Conversational Agents Trust Calibration: A User-Centred Perspective to Design', in *Proceedings of the*

4th Conference on Conversational User Interfaces, Glasgow United Kingdom: ACM, Jul. 2022, pp. 1–6. doi: 10.1145/3543829.3544518.
[40] M. Meyer and J. Dykes, 'Criteria for Rigor in Visualization Design Study', *IEEE Trans. Visual. Comput. Graphics*, pp. 1–1, 2019, doi: 10.1109/TVCG.2019.2934539.
[41] M. Lewis, 'Designing for Human-Agent Interaction', AI Magazine, vol. 19, pp. 67-78, Jun. 1998, doi: https://doi.org/10.1609/aimag.v19i2.1369
[42] J. Złotowski, E. Strasser, and C. Bartneck, 'Dimensions of anthropomorphism: from humanness to humanlikeness', in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, Bielefeld Germany: ACM, Mar. 2014, pp. 66–73. doi: 10.1145/2559636.2559679.
[43] B. Baumgaertner and A. Weiss, 'Do Emotions Matter in the Ethics of Human-Robot Interaction? - Artificial Empathy and Companion Robots'.
[44] Z. Rezaei Khavas, M. R. Kotturu, S. R. Ahmadzadeh, and P. Robinette, 'Do Humans Trust Robots that Violate Moral Trust?', *J. Hum.-Robot Interact.*, vol. 13, no. 2, pp. 1–30, Jun. 2024, doi: 10.1145/3651992.
[45] S. R. Cox and W. T. Ooi, 'Does Chatbot Language Formality Affect Users' Self-Disclosure?', in *Proceedings of the 4th Conference on Conversational User Interfaces*, Glasgow United Kingdom: ACM, Jul. 2022, pp. 1–13. doi: 10.1145/3543829.3543831.
[46] C. D. Kidd and C. Breazeal, 'Effect of a robot on user perceptions', in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, Sendai, Japan: IEEE, 2004, pp. 3559–3564. doi: 10.1109/IROS.2004.1389967.
[47] M. Chita-Tegmark, J. M. Ackerman, and M. Scheutz, 'Effects of Assistive Robot Behavior on Impressions of Patient Psychological Attributes: Vignette-Based Human-Robot Interaction Study', *J Med Internet Res*, vol. 21, no. 6, p. e13729, May 2019, doi: 10.2196/13729.
[48] M. Desai *et al.*, 'Effects of changing reliability on trust of robot systems', in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, Boston Massachusetts USA: ACM, Mar. 2012, pp. 73–80. doi: 10.1145/2157689.2157702.
[49] S. C. Marsella and J. Gratch, 'EMA: A process model of appraisal dynamics', *Cognitive Systems Research*, vol. 10, no. 1, pp. 70–90, Mar. 2009, doi: 10.1016/j.cogsys.2008.03.005.
[50] E. Deng, B. Mutlu, and M. Mataric, 'Embodiment in Socially Interactive Robots', *FNT in Robotics*, vol. 7, no. 4, pp. 251–356, 2019, doi: 10.1561/2300000056.
[51] C. R. Rogers, 'Empathic: An Unappreciated Way of Being', *The Counseling Psychologist*, vol. 5, no. 2, pp. 2–10, 1975, doi: 10.1177/001100007500500202.
[52] C. Frauenberger, 'Entanglement HCI The Next Wave?', *ACM Trans. Comput.-Hum. Interact.*, vol. 27, no. 1, pp. 1–27, Feb. 2020, doi: 10.1145/3364998.
[53] M. Ragni, A. Rudenko, B. Kuhnert, and K. O. Arras, 'Errare humanum est: Erroneous robots in human-robot interaction', in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, New York, NY, USA: IEEE, Aug. 2016, pp. 501–506. doi: 10.1109/ROMAN.2016.7745164.

[54] M. Kosinski, ‘Evaluating large language models in theory of mind tasks’, *Proceedings of the National Academy of Sciences*, vol. 121, Oct. 2024, doi: 10.1073/pnas.2405460121.
[55] S. Desai, M. Dubiel, and L. A. Leiva, ‘Examining Humanness as a Metaphor to Design Voice User Interfaces’, in *ACM Conversational User Interfaces 2024*, Luxembourg Luxembourg: ACM, Jul. 2024, pp. 1–15. doi: 10.1145/3640794.3665535.
[56] R. Setchi, M. B. Dehkordi, and J. S. Khan, ‘Explainable Robotics in Human-Robot Interactions’, *Procedia Computer Science*, vol. 176, pp. 3057–3066, 2020, doi: 10.1016/j.procs.2020.09.198.
[57] A. Barredo Arrieta *et al.*, ‘Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI’, *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.
[58] J. Zhang, J. Conway, and C. A. Hidalgo, ‘Why people judge humans differently from machines: The role of perceived agency and experience’, Sep. 19, 2023, *arXiv*: arXiv:2210.10081. Accessed: Dec. 05, 2024. [Online]. Available: http://arxiv.org/abs/2210.10081
[59] N. Inie, S. Druga, P. Zukerman, and E. M. Bender, ‘From “AI” to Probabilistic Automation: How Does Anthropomorphization of Technical Systems Descriptions Influence Trust?’, in *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, Rio de Janeiro Brazil: ACM, Jun. 2024, pp. 2322–2347. doi: 10.1145/3630106.3659040.
[60] K. Dautenhahn, B. Ogden, and T. Quick, ‘From embodied to socially embedded agents – Implications for interaction-aware robots’, *Cognitive Systems Research*, vol. 3, no. 3, pp. 397–428, Sep. 2002, doi: 10.1016/S1389-0417(02)00050-5.
[61] G. Hannibal and A. Weiss, ‘Fundamental and Methodological Perspectives on the Interrelation of Authenticity and Trustworthiness in Human-Agent Interaction’.
[62] Grammarly, Inc., *Grammarly for Desktop*. (2024). Grammarly, Inc. [Online]. Available: https://www.grammarly.com/.
[63] S. Amershi *et al.*, ‘Guidelines for Human-AI Interaction’, in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk: ACM, May 2019, pp. 1–13. doi: 10.1145/3290605.3300233.
[64] S. R. Fussell, S. Kiesler, L. D. Setlock, and V. Yew, ‘How people anthropomorphize robots’, in *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, Amsterdam The Netherlands: ACM, Mar. 2008, pp. 145–152. doi: 10.1145/1349822.1349842.
[65] A. Weiss, ‘Human Interactions With (Embodied) AI: The Future of Authenticity in Human–AI Relation(ship)s’, *The De Gruyter Handbook of Robots in Society and Culture*, vol. 3, p. 221, Sep. 2024.
[66] N. C. Krämer, A. Von Der Pütten, and S. Eimler, ‘Human-Agent and Human-Robot Interaction Theory: Similarities to and Differences from Human-Human Interaction’, in *Human-Computer Interaction: The Agency Perspective*, vol. 396, M. Zacarias and J. V. De Oliveira, Eds., in Studies in Computational Intelligence, vol. 396. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 215–240. doi: 10.1007/978-3-642-25691-2_9.

[67] M. M. E. Van Pinxteren, M. Pluymaekers, and J. G. A. M. Lemmink, 'Human-like communication in conversational agents: a literature review and research agenda', *JOSM*, vol. 31, no. 2, pp. 203–225, Jun. 2020, doi: 10.1108/JOSM-06-2019-0175.

[68] M. Mara, M. Appel, and T. Gnambs, 'Human-Like Robots and the Uncanny Valley: A Meta-Analysis of User Responses Based on the Godspeed Scales', *Zeitschrift für Psychologie*, vol. 230, no. 1, pp. 33–46, Jan. 2022, doi: 10.1027/2151-2604/a000486.

[69] M. A. Goodrich and A. C. Schultz, 'Human-Robot Interaction: A Survey', *FNT in Human-Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2007, doi: 10.1561/1100000005.

[70] J. M. Bradshaw, P. J. Feltovich, and M. Johnson, 'Human–Agent Interaction', in *The Handbook of Human-Machine Interaction*, 1st ed., G. A. Boy, Ed., CRC Press, 2017, pp. 283–300. doi: 10.1201/9781315557380-14.

[71] H. Osawa, 'Human–agent interaction as augmentation of social intelligence', *Artif Life Robotics*, vol. 28, no. 2, pp. 273–281, May 2023, doi: 10.1007/s10015-023-00874-y.

[72] S. Nyholm, *Humans and robots: ethics, agency, and anthropomorphism*. in Philosophy, technology and society. London ; New York: Rowman & Littlefield Publishing Group, 2020.

[73] D. Dellermann, P. Ebel, M. Söllner, and J. M. Leimeister, 'Hybrid Intelligence', *Bus Inf Syst Eng*, vol. 61, no. 5, pp. 637–643, Oct. 2019, doi: 10.1007/s12599-019-00595-2.

[74] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, 'Impact of robot failures and feedback on real-time trust', in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Tokyo, Japan: IEEE, Mar. 2013, pp. 251–258. doi: 10.1109/HRI.2013.6483596.

[75] A. Pereira, R. Prada, and A. Paiva, 'Improving social presence in human-agent interaction', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Toronto Ontario Canada: ACM, Apr. 2014, pp. 1449–1458. doi: 10.1145/2556288.2557180.

[76] A. Schepman and P. Rodway, 'Initial validation of the general attitudes towards Artificial Intelligence Scale', *Computers in Human Behavior Reports*, vol. 1, p. 100014, Jan. 2020, doi: 10.1016/j.chbr.2020.100014.

[77] F. Chen and J. Terken, 'Interaction Design Theory', in *Automotive Interaction Design*, in Springer Tracts in Mechanical Engineering. , Singapore: Springer Nature Singapore, 2023, pp. 129–144. doi: 10.1007/978-981-19-3448-3_8.

[78] C. Harms and F. Biocca, 'Internal Consistency and Reliability of the Networked Minds Measure of Social Presence', 2004.

[79] D. Parisi, 'Internal robotics', *Connection Science*, vol. 16, no. 4, pp. 325–338, Dec. 2004, doi: 10.1080/09540090412331314768.

[80] M. Hudson and P. Cairns, 'Interrogating social presence in games with experiential vignettes', *Entertainment Computing*, vol. 5, no. 2, pp. 101–114, May 2014, doi: 10.1016/j.entcom.2014.01.001.

[81] J. Cassell and A. Tartaro, 'Intersubjectivity in human–agent interaction', *IS*, vol. 8, no. 3, pp. 391–410, Nov. 2007, doi: 10.1075/is.8.3.05cas.

[82] R. Baeza-Yates, 'Introduction to Responsible AI', in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*,

Merida Mexico: ACM, Mar. 2024, pp. 1114–1117. doi: 10.1145/3616855.3636455.
[83] N. Mariacher, S. Schlögl, and A. Monz, 'Investigating Perceptions of Social Intelligence in Simulated Human-Chatbot Interactions', vol. 184, 2021, pp. 513–529. doi: 10.1007/978-981-15-5093-5_44.
[84] R. Heimgärtner, 'ISO 9241-210 and Culture? – The Impact of Culture on the Standard Usability Engineering Process'.
[85] A. Chandrasekaran, D. Yadav, P. Chattopadhyay, V. Prabhu, and D. Parikh, 'It Takes Two to Tango: Towards Theory of AI's Mind', Oct. 02, 2017, *arXiv*: arXiv:1704.00717. Accessed: Nov. 10, 2024. [Online]. Available: http://arxiv.org/abs/1704.00717
[86] G. Haas, M. Rietzler, M. Jones, and E. Rukzio, 'Keep it Short: A Comparison of Voice Assistants' Response Behavior', in *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, Apr. 2022, pp. 1–12. doi: 10.1145/3491102.3517684.
[87] G. E. Newman and R. K. Smith, 'Kinds of Authenticity', *Philosophy Compass*, vol. 11, no. 10, pp. 609–618, Oct. 2016, doi: 10.1111/phc3.12343.
[88] S. W. T. Chan, T. S. Gunasekaran, Y. S. Pai, H. Zhang, and S. Nanayakkara, 'KinVoices: Using Voices of Friends and Family in Voice Interfaces', *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, pp. 1–25, Oct. 2021, doi: 10.1145/3479590.
[89] L. M. Ahearn, 'Language and Agency', *Annu. Rev. Anthropol.*, vol. 30, no. 1, pp. 109–137, Oct. 2001, doi: 10.1146/annurev.anthro.30.1.109.
[90] J. C. R. Licklider, 'Man-Computer Symbiosis', *IRE Trans. Hum. Factors Electron.*, vol. HFE-1, no. 1, pp. 4–11, Mar. 1960, doi: 10.1109/THFE2.1960.4503259.
[91] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, 'Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots', *Int J of Soc Robotics*, vol. 1, no. 1, pp. 71–81, Jan. 2009, doi: 10.1007/s12369-008-0001-3.
[92] MidnightHowlingHuskyDog, 'midnighthowlinghuskydog', Instagram. Accessed: Dec. 08, 2024. [Online]. Available: https://www.instagram.com/midnighthowlinghuskydog/
[93] G. Abercrombie, A. C. Curry, T. Dinkar, V. Rieser, and Z. Talat, 'Mirages: On Anthropomorphism in Dialogue Systems', Oct. 23, 2023, *arXiv*: arXiv:2305.09800. Accessed: Nov. 01, 2024. [Online]. Available: http://arxiv.org/abs/2305.09800
[94] S. Beerends and C. Aydin, 'Negotiating the authenticity of AI: how the discourse on AI rejects human indeterminacy', *AI & Soc*, Feb. 2024, doi: 10.1007/s00146-024-01884-5.
[95] N. Epley, A. Waytz, and J. T. Cacioppo, 'On seeing human: A three-factor theory of anthropomorphism.', *Psychological Review*, vol. 114, no. 4, pp. 864–886, 2007, doi: 10.1037/0033-295X.114.4.864.
[96] M. Appel, S. Weber, S. Krause, and M. Mara, 'On the eeriness of service robots with emotional capabilities', in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Christchurch, New Zealand: IEEE, Mar. 2016, pp. 411–412. doi: 10.1109/HRI.2016.7451781.
[97] L. Pessoa, 'On the relationship between emotion and cognition', *Nat Rev Neurosci*, vol. 9, no. 2, pp. 148–158, Feb. 2008, doi: 10.1038/nrn2317.

[98] T. Ziemke and R. Lowe, ‘On the Role of Emotion in Embodied Cognitive Architectures: From Organisms to Robots’, *Cogn Comput*, vol. 1, no. 1, pp. 104–117, Mar. 2009, doi: 10.1007/s12559-009-9012-0.
[99] M. Neururer, S. Schlögl, L. Brinkschulte, and A. Groth, ‘Perceptions on Authenticity in Chat Bots’, *MTI*, vol. 2, no. 3, p. 60, Sep. 2018, doi: 10.3390/mti2030060.
[100] G. Luca Liehner, A. Hick, H. Biermann, P. Brauner, and M. Ziefle, ‘Perceptions, attitudes and trust toward artificial intelligence — An assessment of the public opinion’, presented at the 14th International Conference on Applied Human Factors and Ergonomics (AHFE 2023), 2023. doi: 10.54941/ahfe1003271.
[101] J. C. Nunnally and I. H. Bernstein, *Psychometric Theory*. in McGraw-Hill series in psychology, no. br. 972. McGraw-Hill Companies,Incorporated, 1994. [Online]. Available: https://books.google.at/books?id=r0fuAAAAMAAJ
[102] M. Luria, S. Reig, X. Z. Tan, A. Steinfeld, J. Forlizzi, and J. Zimmerman, ‘Re-Embodiment and Co-Embodiment: Exploration of social presence for robots and conversational agents’, in *Proceedings of the 2019 on Designing Interactive Systems Conference*, San Diego CA USA: ACM, Jun. 2019, pp. 633–644. doi: 10.1145/3322276.3322340.
[103] S. Goellner, M. Tropmann-Frick, and B. Brumen, ‘Responsible Artificial Intelligence: A Structured Literature Review’, Mar. 11, 2024, *arXiv*: arXiv:2403.06910. Accessed: Oct. 19, 2024. [Online]. Available: http://arxiv.org/abs/2403.06910
[104] J. Danaher, ‘Robot Betrayal: a guide to the ethics of robotic deception’, *Ethics Inf Technol*, vol. 22, no. 2, pp. 117–128, Jun. 2020, doi: 10.1007/s10676-019-09520-3.
[105] A. A. Hamed, M. Zachara-Szymanska, and X. Wu, ‘Safeguarding authenticity for mitigating the harms of generative AI: Issues, research agenda, and policies for detection, fact-checking, and ethical AI’, *iScience*, vol. 27, no. 2, p. 108782, Feb. 2024, doi: 10.1016/j.isci.2024.108782.
[106] M. Mende, M. L. Scott, J. Van Doorn, D. Grewal, and I. Shanks, ‘Service Robots Rising: How Humanoid Robots Influence Service Experiences and Elicit Compensatory Consumer Responses’, *Journal of Marketing Research*, vol. 56, no. 4, pp. 535–556, Aug. 2019, doi: 10.1177/0022243718822827.
[107] L. W. Barsalou, P. M. Niedenthal, A. K. Barbey, and J. A. Ruppert, ‘Social Embodiment’, in *Psychology of Learning and Motivation*, vol. 43, Elsevier, 2003, pp. 43–92. doi: 10.1016/S0079-7421(03)01011-9.
[108] H. S. Sætra, ‘Social robot deception and the culture of trust’, *Paladyn, Journal of Behavioral Robotics*, vol. 12, no. 1, pp. 276–286, Apr. 2021, doi: 10.1515/pjbr-2021-0021.
[109] K. Dautenhahn, ‘Socially intelligent robots: dimensions of human–robot interaction’, *Phil. Trans. R. Soc. B*, vol. 362, no. 1480, pp. 679–704, Apr. 2007, doi: 10.1098/rstb.2006.2004.
[110] N. Coupland, ‘Sociolinguistic authenticities’, *Journal of Sociolinguistics*, vol. 7, no. 3, pp. 417–431, Aug. 2003, doi: 10.1111/1467-9481.00233.
[111] D. Marikyan and S. Papagiannidis, ‘Task-Technology Fit: A review’, in *TheoryHub Book*, 2023. [Online]. Available: https://open.ncl.ac.uk
[112] D. Erwin. Byrne, *The attraction paradigm [by] Donn Byrne.* in Personality and psychopathology, 11. New York: Academic Press, 1971.

[113] S. Baron-Cohen and S. Wheelwright, 'The Empathy Quotient: An Investigation of Adults with Asperger Syndrome or High Functioning Autism, and Normal Sex Differences', *J Autism Dev Disord*, vol. 34, no. 2, pp. 163–175, Apr. 2004, doi: 10.1023/B:JADD.0000022607.19833.00.
[114] J. P. Cabral, B. R. Cowan, K. Zibrek, and R. McDonnell, 'The Influence of Synthetic Voice on the Evaluation of a Virtual Character', in *Interspeech 2017*, ISCA, Aug. 2017, pp. 229–233. doi: 10.21437/Interspeech.2017-325.
[115] V. Thomas, C. Remy, and O. Bates, 'The Limits of HCD: Reimagining the Anthropocentricity of ISO 9241-210', in *Proceedings of the 2017 Workshop on Computing Within Limits*, Santa Barbara California USA: ACM, Jun. 2017, pp. 85–92. doi: 10.1145/3080556.3080561.
[116] J. Bowlby, 'The nature of the child's tie to his mother 1', in *Influential Papers from the 1950s*, 1st ed., A. C. Furman and S. T. Levy, Eds., Routledge, 2018, pp. 222–273. doi: 10.4324/9780429475931-15.
[117] J. Wainer, D. Feil-seifer, D. Shell, and M. Mataric, 'The role of physical embodiment in human-robot interaction', in *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, Univ. of Hertfordshire, Hatfield, UK: IEEE, Sep. 2006, pp. 117–122. doi: 10.1109/ROMAN.2006.314404.
[118] J. Short, E. Williams, and B. Christie, *The social psychology of telecommunications*. London: Wiley, 1976.
[119] T. Jokela, N. Iivari, J. Matero, and M. Karukka, 'The Standard of User-Centered Design and the Standard Definition of Usability: Analyzing ISO 13407 against ISO'.
[120] H. Muller, M. T. Mayrhofer, E.-B. Van Veen, and A. Holzinger, 'The Ten Commandments of Ethical Medical AI', *Computer*, vol. 54, no. 7, pp. 119–123, Jul. 2021, doi: 10.1109/MC.2021.3074263.
[121] M. Mori, K. MacDorman, and N. Kageki, 'The Uncanny Valley [From the Field]', *IEEE Robot. Automat. Mag.*, vol. 19, no. 2, pp. 98–100, Jun. 2012, doi: 10.1109/MRA.2012.2192811.
[122] J. Finch, 'The Vignette Technique in Survey Research', *Sociology*, vol. 21, no. 1, pp. 105–114, Feb. 1987, doi: 10.1177/0038038587021001008.
[123] J. Klein, Y. Moon, and R. W. Picard, 'This computer responds to user frustration: Theory, design, and results', *Interacting with Computers*, 2002.
[124] C. Breazeal, 'Toward sociable robots', *Robotics and Autonomous Systems*, vol. 42, no. 3–4, pp. 167–175, Mar. 2003, doi: 10.1016/S0921-8890(02)00373-1.
[125] K. Reinhardt, 'Trust and trustworthiness in AI ethics', 2023.
[126] M. Blut, C. Wang, N. V. Wünderlich, and C. Brock, 'Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other AI', *J. of the Acad. Mark. Sci.*, vol. 49, no. 4, pp. 632–658, Jul. 2021, doi: 10.1007/s11747-020-00762-y.
[127] A. Kandemir and R. Budd, 'Using Vignettes to Explore Reality and Values With Young People'.
[128] J. Danaher, 'Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism', *Sci Eng Ethics*, vol. 26, no. 4, pp. 2023–2049, Aug. 2020, doi: 10.1007/s11948-019-00119-x.
[129] Y. Xu and M. Warschauer, 'What Are You Talking To?: Understanding Children's Perceptions of Conversational Agents', in *Proceedings of the*

2020 CHI Conference on Human Factors in Computing Systems, Honolulu HI USA: ACM, Apr. 2020, pp. 1–13. doi: 10.1145/3313831.3376416.
[130] L. Clark *et al.*, 'What Makes a Good Conversation?: Challenges in Designing Truly Conversational Agents', in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk: ACM, May 2019, pp. 1–12. doi: 10.1145/3290605.3300705.
[131] A.-M. Seeger, J. Pfeiffer, and A. Heinzl, 'When Do We Need a Human? Anthropomorphic Design and Trustworthiness of Conversational Agents', in *Proceedings of the Sixteenth Annual Pre-ICIS Workshop on HCI Research in MIS*, Seoul Korea, 2017.
[132] E. Elshan, N. Zierau, C. Engel, A. Janson, and J. M. Leimeister, 'Understanding the Design Elements Affecting User Acceptance of Intelligent Agents: Past, Present and Future', *Inf Syst Front*, vol. 24, no. 3, pp. 699–730, Jun. 2022, doi: 10.1007/s10796-021-10230-9.

Appendices

Appendix A: Authenticity Scale

Instructions:
Please rate your impression of the interaction on these scales:

Semantic differential scales (1-5):
Truthful-Deceptive
Consistent-Unreliable
Coherent-Confused
Authentic-Inauthentic
Authorized-Unauthorized

Appendix B: Social Vignettes

Career Vignette:
You've been offered a promotion at work that comes with increased responsibilities and a significant pay raise. However, the new role requires you to relocate to a different city, which would mean leaving behind your current social circle and adapting to a new environment. Additionally, you're unsure about the long-term stability of the new position. You're conflicted about whether to accept the offer or stay in your current role.

Health Vignette:
You have been feeling increasingly fatigued and stressed over the past few months. Despite getting enough sleep, you wake up tired and struggle to get through your daily tasks. You've also noticed frequent headaches and occasional dizziness. You've tried to manage your stress with exercise and relaxation techniques, but nothing seems to work. You're worried that something more serious might be going on and are unsure about what to do next.

Social Relationships Vignette:
You've recently had a falling out with a close friend due to a misunderstanding. This friend has been an important part of your social life, and the conflict has left you feeling hurt and isolated. Despite multiple attempts to reach out and clarify the situation, your friend has not responded positively. You're unsure how to mend the relationship and whether it's worth continuing to try.

Appendix C: Code for Generating PDPs

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib.cm as cm
from sklearn.ensemble import RandomForestRegressor
from sklearn.inspection import partial_dependence
from mpl_toolkits.mplot3d import Axes3D
import numpy as np

# Load the data
data_path = "/Users/barbarageld/Desktop/Thesis/Data/spss-working-
data.xlsx"
data = pd.read_excel(data_path)

# Rename columns for readability
data = data.rename(columns={
    'SPQ_Aggregated': 'Social Presence',
    'ANT_Aggregated': 'Anthropomorphism',
    'LIK_Aggregated': 'Likeability',
    'HL_Aggregated': 'Human-Likeness',
    'AUT_Aggregated': 'Authenticity'
})

# Define features and target variable
X = data[['Social Presence', 'Anthropomorphism', 'Likeability']]
y = data['Human-Likeness']

# Train the model
model = RandomForestRegressor(n_estimators=100, random_state=0)
model.fit(X, y)

# Define a consistent green color for the trend line
trendline_color = cm.get_cmap('viridis')(0.4)  # Consistent bluish-
green shade

# ========== Generate and Save Scatter Plots Independently
==========
scatter_specs = [
    ('Social Presence', 'Anthropomorphism', 'Social Presence vs
Anthropomorphism',
"/Users/barbarageld/Desktop/scatter_social_vs_anthropomorphism.png")
,
    ('Social Presence', 'Likeability', 'Social Presence vs
Likeability',
"/Users/barbarageld/Desktop/scatter_social_vs_likeability.png"),
    ('Anthropomorphism', 'Likeability', 'Anthropomorphism vs
Likeability',
"/Users/barbarageld/Desktop/scatter_anthropomorphism_vs_likeability.
png"),
    ('Human-Likeness', 'Authenticity', 'Human-Likeness vs
Authenticity',
"/Users/barbarageld/Desktop/scatter_human_likeness_vs_authenticity.p
ng")
]

# Generate and save each scatter plot independently
for x_var, y_var, title, filename in scatter_specs:
    plt.figure(figsize=(6, 6))
    sns.regplot(data=data, x=x_var, y=y_var, scatter_kws={'color':
'black'}, line_kws={"color": trendline_color, "linewidth": 1})
    plt.title(title)
    plt.xlabel(x_var)
    plt.ylabel(y_var)
```

```
    plt.savefig(filename, bbox_inches="tight", dpi=300)
    plt.close()
    print(f"Scatter plot saved as {filename}")



# ========== Generate 3D PDPs with Each Pair of Predictors
==========
# Function to plot and save 3D PDPs
def plot_3d_pdp(feature_indices, xlabel, ylabel, filename):
    pdp_result = partial_dependence(model, X=X,
features=feature_indices, grid_resolution=50)
    xx, yy = np.meshgrid(pdp_result['grid_values'][0],
pdp_result['grid_values'][1])
    zz = pdp_result['average'].reshape(xx.shape)

    fig = plt.figure(figsize=(8, 8))  # Adjusted size for balanced
view
    ax = fig.add_subplot(111, projection='3d')
    surf = ax.plot_surface(xx, yy, zz, cmap="viridis",
edgecolor='k', alpha=0.7)

    ax.set_xlabel(xlabel)
    ax.set_ylabel(ylabel)
    ax.set_zlabel("Partial Dependence on Human-Likeness")
    plt.title(f"3D Partial Dependence Plot: {xlabel} and {ylabel}")

    fig.colorbar(surf, shrink=0.5, aspect=5)
    plt.savefig(filename, bbox_inches="tight", dpi=300)
    plt.show()

# Generate and save 3D PDPs for each pair of predictors
plot_3d_pdp([0, 1], "Social Presence", "Anthropomorphism",
"3D_pdp_social_presence_anthropomorphism.png")
plot_3d_pdp([0, 2], "Social Presence", "Likeability",
"3D_pdp_social_presence_likeability.png")
plot_3d_pdp([1, 2], "Anthropomorphism", "Likeability",
"3D_pdp_anthropomorphism_likeability.png")
```